KALMAN FILTER ALGORITHM FOR RATING AND PREDICTION IN BASKETBALL

JIRKA POROPUDAS

University of Helsinki
Faculty of Social Sciences
Statistics
Master's thesis
May 2011

Acknowledgements

Helsinki, May 2011 Jirka Poropudas

I would like to thank following people for making everything possible. First of all, my gratitude goes to my parents and family, Eeva, and all my friends. I would also like to thank university teacher Pekka Pere, Jykke "Lammas" Venhola, and Jukka Myllyniemi for their contribution to this Thesis. Additionally, I am forever indebted to my teachers for their encouragement in my studies of science and, in particular, statistics (in somewhat chronological order): Seppo Hyyti, Anja Peitola, Yngve Lehtosaari, Pekka Norlamo, Elina Näsäkkälä, Juhani Kaila, Kai Virtanen, Ilkka Mellin, Pertti Laininen, and Juha Koljonen. Furthermore, this Thesis would not exist without the inspiration of John Stockton, Karl Malone, Mark Eaton, Manute Bol, Dikembe Mutombo, Shawn Bradley, Steve Nash, Dirk Nowitzki, and DeSagana Diop. I would also like to thank all my co-workers at the Systems Analysis Laboratory at the Aalto University School of Science for an enthusiastic and open working environment as well as for their invaluable advice and our innumerable pseudo-scientific discussions. Finally, my thanks go to all my teammates and coaches throughout the years – basketball truly is a beautiful sport.

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Faculty of Social Sciences	Statistics

Jirka Poropudas

Kalman Filter Algorithm for Rating and Prediction in Basketball

Time series analysis and econometrics

Master's thesis	May 3rd 2011	86 pages

The Thesis presents a state-space model for a basketball league and a Kalman filter algorithm for the estimation of the state of the league. In the state-space model, each of the basketball teams is associated with a rating that represents its strength compared to the other teams. The ratings are assumed to evolve in time following a stochastic process with independent Gaussian increments. The estimation of the team ratings is based on the observed game scores that are assumed to depend linearly on the true strengths of the teams and independent Gaussian noise. The team ratings are estimated using a recursive Kalman filter algorithm that produces least squares optimal estimates for the team strengths and predictions for the scores of the future games. Additionally, if the Gaussianity assumption holds, the predictions given by the Kalman filter maximize the likelihood of the observed scores.

The team ratings allow probabilistic inference about the ranking of the teams and their relative strengths as well as about the teams' winning probabilities in future games. The predictions about the winners of the games are correct 65-70% of the time. The team ratings explain 16% of the random variation observed in the game scores. Furthermore, the winning probabilities given by the model are concurrent with the observed scores.

The state-space model includes four independent parameters that involve the variances of noise terms and the home court advantage observed in the scores. The Thesis presents the estimation of these parameters using the maximum likelihood method as well as using other techniques. The Thesis also gives various example analyses related to the American professional basketball league, i.e., National Basketball Association (NBA), and regular seasons played in year 2005 through 2010. Additionally, the season 2009-2010 is discussed in full detail, including the playoffs.

Kalman filter algorithm, state-space model, ranking, rating, prediction, basketball, National Basketball Association

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Valtiotieteellinen tiedekunta	Tilastotieteen laitos

Jirka Poropudas

Kalman Filter Algorithm for Rating and Prediction in Basketball

Aikasarja-analyysin ja ekonometrian linja

Pro gradu –tutkielma	3.5. 2011	86 sivua

Työssä esitellään koripalloliigaa kuvaava diskreettiaikainen tila-avaruusmalli sekä Kalman-suotimeen perustuva algoritmi koripalloliigan tilan estimointiin. Tila-avaruusmallissa jokaisella liigan joukkueella on voimaluku ("rating"), joka kuvastaa joukkueen hyvyyttä suhteessa muihin liigan joukkueisiin. Joukkueiden voimasuhteiden oletetaan kehittyvän stokastisesti ajassa ja niiden muutokset oletetaan normaalijakautuneiksi riippumattomiksi satunnaismuuttujiksi. Joukkueiden voimalukujen estimointi perustuu havaittuihin ottelutuloksiin, joiden oletetaan riippuvan lineaarisesti joukkueiden todellisista voimaluvuista sekä normaalijakautuneesta kohinasta. Joukkueiden voimaluvut estimoidaan rekursiivisella Kalman-suodin algoritmilla, joka tuottaa pienimmän neliösumman mielessä optimaalisia estimaatteja joukkueiden voimaluvuille sekä tuleville ottelutuloksille. Lisäksi normaalijakauma-oletusten pätiessä Kalman-suotimen tuottamat ennusteet maksimoivat havaittujen tulosten uskottavuuden.

Voimalukuestimaattien avulla voidaan tehdä päätelmiä joukkueiden paremmuusjärjestyksestä ja niiden välisistä tasoeroista sekä ennustaa voitontodennäköisyyksiä tulevissa otteluissa. Mallin ennusteet otteluiden voittajista osuvat oikein 65-70%:sti. Mallin avulla pystytään selittämään noin 16% ottelutulosten satunnaisesta vaihtelusta. Lisäksi mallin tuottamat arviot joukkueiden voitontodennäköisyyksistä ovat sopusoinnussa havaittujen ottelutulosten kanssa.

Tila-avaruusmallissa on neljä riippumatonta parametria, jotka kuvaavat erilaisten kohinatermien varianssia sekä ottelutuloksien sisältämää kotietua. Työssä esitellään näiden parametrien estimointi sekä suurimman uskottavuuden menetelmällä että hyödyntäen muita keinoja. Työ tarjoaa myös lukuisia esimerkkianalyyseja, joissa tarkastellaan amerikkalaisen ammattilaiskoripallosarjan NBA:n (National Basketball Association) runkosarjan ottelutuloksia vuosina 2005-2010. Lisäksi esitetään kauden 2009-2010 yksityiskohtaisempi tarkastelu, mukaan lukien pudotuspelit.

Kalman-suodin algoritmi, tila-avaruusmalli, ranking, rating, ennustaminen, koripallo, National Basketball Association

Contents

List of Figures

List of Tables

1. Introduction

Ranking and rating of basketball teams (see, e.g., Govan 2009, Govan et al. 2009) as well as of individual players (see, e.g., Berri 1999, Oliver 2004, Berri et al. 2007, Doolittle and Pelton 2010) is an emerging topic of statistical analysis that captures the interest of both fans and professionals. There are also situations where the ranking of teams holds actual relevance in, e.g., determining the seeding of teams in tournament brackets (Shapiro et al. 2009). The common perception is that statistical analysis of basketball yields additional insight and added value in a manner similar to the analysis of other sports such as baseball (Lewis 2003) and soccer (Kuper and Szymanski 2009). In such analyses, observations of the performance levels of teams and individual players are aggregated into statistical evidence that is then exposed to comprehensive testing and study. The thinking is that given a large enough data set, the statistical analysis unveils trends and phenomena that are hard to discover by naked eye, i.e., by simply watching live basketball games.

This Thesis presents a state-space model for a basketball league that is based on observed game results and the Kalman filter algorithm (Kalman 1960). In the model, the state of the system represents the strengths of the teams and game point differentials are taken as observations of the relative differences between the teams' level of performance. New observations are combined with the current estimate of the system state recursively using the Kalman filter algorithm in order to produce updated estimates of the system state that minimize the mean square error. The estimates of the system state are used to rank and rate the basketball teams according to their level of performance as well as to predict the future outcomes of games.

1.1. **Background Information about Basketball.** Basketball is a team sport where the teams score points by shooting the ball to the basket of the opponent (FIBA 2010). Successful scoring attempt is called a basket. The scoring team is awarded one, two, or three points, depending on the position of the shooting player at the time of the shot attempt. The team with the most points at the end of the regulation time wins the game. A single game lasts 40 minutes when international rules are used (FIBA 2010). In American professional leagues, the regulation time lasts 48 minutes. A basketball game cannot end in a tie. If the teams have equal number of points at the end of the regulation, additional five minute periods are played until the winner is found. The teams' active rosters consist of maximum of 12 players out of whom five are on the court at one time.

Basketball competition is organized in leagues, where the teams first complete a regular season schedule and then the league championship is determined by playoffs. In general, the teams with best records receive the most advantageous playoff schedule and the worst teams are left out of playoffs entirely. The international competition in basketball is governed by International Basketball Association (Fédération Internationale de Basketball, FIBA). This Thesis concentrates on the National Basketball Association (NBA). It is the pre-eminent men's professional basketball league in North America which composes thirty franchised member clubs, of which twenty-nine are located in the United States and one in Canada. A more comprehensive presentation of the history of NBA is given, e.g., in Yilmaz and Chatterjee (2000).

During a single game, NBA teams make between 75 and 90 field goal attempts on average. There exists some controversy about the statistical dependence between the successfulness of consecutive shot attempts. Gilovich et al. claim that shot attempts can be considered as independent and that no so called "hot hand" phenomenon exists (Gilovich et al. 1985). In general, this view is not shared by all basketball analysts and it has been challenged in more recent analyses (Bar-Eli et al. 2006, Arkes 2010). Nevertheless, in practice the dependence between shot attempts, i.e., their correlation, is not dominant and it is reasonable to assume that the central limit theorem (see, e.g., Davidson and MacKinnon 1993, pp. 136) holds for total number of points scored by the two teams, and each of the individual teams. In other words, the joint probability distribution of the points scored by the home team and the visiting team should approximately follow a multinormal distribution with two variables. Due to the rules of basketball, the game cannot end in a tie, i.e., the two variables cannot have identical values. This results in ”a gap” in the middle of the joint probability distribution of the teams' points. Nevertheless, the joint probability distribution can be approximated with multinormal distribution (as is seen in Chapter 3).

The aim of this Thesis is to measure the level of performance of the basketball teams based on the observed outcomes of games. The team performance is a collection of multiple factors pertaining to different aspects of the game: offence, i.e., scoring points; defence, i.e., preventing the opponent from scoring; and winning the game, i.e., scoring more points than the opponent. In previous literature, various measures of team performance have been suggested such as points per game, points allowed per game, and winning percentage (Yilmaz and Chatterjee 2000, NBA 2011). Usually, it is assumed that the level of team performance changes in time both during the season and between consecutive seasons (Yilmaz and Chatterjee 2000, Chatterjee and Yilmaz 1999). The Thesis is based on the same starting points, but now the focus is on the evaluation of teams during a single basketball season.

1.2. **Previous Work in Ranking and Rating of Basketball Teams.** The objective of ranking and rating of teams is to order the teams and show which teams are the best. In practice, the game results are observed and the teams are ranked or rated according to their perceived level of performance. In this Thesis, the *result* or the *score* of a game is represented by either its outcome or its point differential. The *outcome* of a game refers to the information about the team that won the game, i.e., the team that had the most points at the end of the game. On the other hand, the final winning margin, i.e., the points scored by the home team less the points scored by the visiting team, is referred to as the *point differential* of the game. In *ranking*, the teams are placed on an ordinal scale, i.e., the separation between successive teams is not evaluated. In *rating*, the ordering takes place on an interval scale so that differences between teams are measurable and have an meaningful interpretation.

In basketball leagues, the final standings are generally determined based on the number of wins by each team (FIBA 2010) or by the teams' winning percentages (NBA 2011). Thus, the ranking of teams based on their cumulative number of wins or winning percentage offers the simplest and most straightforward approach to measuring the teams' level of play. This is particularly practical at the end of season when all teams have completed their schedules. During the season when the league schedule has not been completed, the effect of schedule is ignored as these metrics do not include the strength of the teams' opponents. Furthermore, the larger basketball leagues may not even use a balanced schedule and the teams may not face the same opponents during the season.

For additional accuracy, the method used to rank or rate the teams should also include the strength of the teams' opponents in the assessment (Fearnhead and Taylor 2009). Thus, when assessing a given teams' strength, it is not enough to just study the number of wins and losses by the team. On the other hand, the wins and winning percentages measure the teams' performance based on the current season in its entirety. This may not give a completely accurate picture of the teams' current level of play as such evaluation assumes implicitly that the teams level of play remains constant during the season and the teams' current level of performance is reflected by the point differentials of all the games played during the season.

The objective of the ranking or rating of teams is to answer the question: "which of the two teams is better?" When the performances of two teams are to be compared, it should be kept in mind that the number of games played between two teams is small and the results include random variation. For example in the NBA, the teams meet at most four times during the season. Thus, available ”direct” information is limited and the comparison of two teams needs to include also the ”indirect” information, i.e., the games played against other teams (Hu and Zidek 2004, Park and Newman 2005). In some leagues, the teams do not play a balanced schedule, i.e., the teams meet only a fraction of the opposing teams during the season. Thus, the teams overall performance is measured based on the games played against all other teams participating in the league.

In extreme cases, the number of teams in the league may be so large that most of the teams do not face each other at all during the season. The most important example of such league is the Men's basketball Division I of the National Collegiate Athletics Association (NCAA) where there are 346 teams in 34 conferences. During the regular season, the teams play between 26 and 36 games each (NCAA 2011). The NCAA basketball champion is determined through the NCAA Men's Division I Basketball Tournament where the top 68 teams compete for the championship (NCAA 2011, Shapiro et al. 2009)[1]. The participating teams are selected according to a ranking based on expert analysis, i.e., the decision of the 10 member NCAA

[1]This tournament is commonly known as the ”March Madness”.

tournament selection committee, and seeded to the tournament bracket. In practice, the winners of 31 conferences automatically qualify for the final tournament and the other 37 teams are selected by the committee. The committee uses various statistical guidelines in selecting and seeding basketball teams for the NCAA tournament, one of which is the Ratings Percentage Index, or RPI (NCAA 2011). For recent and more comprehensive presentations of the selection and seeding process (see, e.g., Shapiro et al. 2009, Coleman and Lynch 2009).

The RPI was created by the NCAA in 1981 to objectively assess the college basketball teams and it is calculated as a weighted sum of the team's winning percentage, its opponents' winning percentage, and its opponents' opponents' winning percentage. The teams' RPI are used to rank the college basketball teams in descending order from one to 346. Other guidelines used by the selection committee include, among others, the teams' win-loss record, non-conference record, non-conference RPI, conference record, conference RPI, etc. One should note that, the weights used to calculate the index were modified in 2005, but they still appear rather arbitrary. Nevertheless, the RPI continues to receive a lot of weight from the NCAA selection committee when decisions are being made about selecting and seeding the teams, especially after the 2005 adjustment (West 2006). Other rankings of NCAA Division I Men's basketball teams include, among others, the Sagarin rating (Sagarin 1985), the Massey rating (Massey 1995), and the Pomeroy rating Pomeroy (2003).

In the NBA, the schedule is a mixture of a balanced and unbalanced schedule – the teams face all the other teams during the season, but the number of meetings varies. Despite the fact the standings in the NBA are determined directly based on the teams' winning percentages (and, in case of a tie, various tiebreakers), the NBA teams are actively rated by the experts of the field. In actuality, all major sports web sites publish their own expert rankings for the NBA (see, e.g., Schuhmann 2010, Stein 2010, Hollinger 2011, Galinsky 2010). The published rankings are usually of ordinal but, e.g., the method used by ESPN's John Hollinger is based on the so called Hollinger Rating (Hollinger 2011). Overall, even though the rankings and ratings have been criticized as subjective, inconsistent, and even biased (Sanders 2007, Zimmer and Kuethe 2008), the ranking and rating of basketball teams is an active area of research and discussion.

Overall, the ranking and rating of teams can be performed ex ante, i.e., the estimates can be recursively updated in real time during the season as soon as new results are observed. This kind of models are needed for following the progress of the season or predicting game outcomes. On the other hand, the ranking and rating may be carried out ex post, i.e., after the season has finished and all the games have been played. For official standings, the ex post approach is used in order to make all the regular season games meaningful.

The simplest models for ranking of teams are calculated directly from the game results. For example, the official league standings are determined based on the teams' number of wins (FIBA 2010) or their winning percentage (NBA 2011), i.e., the number of wins divided by the total number of games played. More detailed information about the teams' performance is included when the ranking is done according to the teams' point differential, i.e., the difference between points scored by the team and its opponents. In such ranking, the teams are awarded for beating their opponents by wider margins which is seen as a quality of a better team. For a more current perspective on the teams level of play, it is possible to calculate the teams' point differential during, e.g., the last 60 days. This kind of approach concentrates of the teams' latest accomplishments and ignores the results of its older games.

In the following, a literary survey of the approaches to ranking and rating of sports teams is given. All the presented models are statistical or stochastic models where the randomness inherent in game results is included but, unless otherwise mentioned, no exact probability distributions are assessed for the ratings. All models listed do not necessarily involve basketball, but all of them could also be applied to ranking or rating of basketball teams. Also in general, the methods used for ranking basketball teams could be applied to other team sports.

Harville (Harville 1977) presents a linear additive model for rating of high school and college football teams where the expectation of the point differential of a game is the difference between teams' ratings with home court advantage added. In the model, the teams' observed performances include deviations from their true level that are generated by a first-order autoregressive stochastic process. Harville (Harville 1980) applies a similar model to National Football League (NFL) data over several years time. Stefani (Stefani 1980) improves the existing models by adding a heuristic reduction factor to the model in order to correct the bias found in the previously predicted margins of victory.

Harville (Harville 2003) discusses the desirable attributes of a fair and balanced ranking method used in the selection of NCAA basketball teams to the final tournament. The article presents a model having all but one of these attributes by applying least squares to a linear additive model in which the expected point differential in each game is modeled as a difference in team ratings plus the home court advantage. Additionally, the benefit of the teams "running up the score" is avoided by weighing down the effects of large victories. With this modification, the teams cannot affect their own rating by intentionally increasing the scoring margin after the outcome of the game has already been decided.[2] Similar robustness factors are discussed in (Bassett 1997) where linear additive model is used in ranking but now, instead of the minimization of least squares errors, the least absolute value of the errors is minimized. Thus, the unwanted effect of large victory margins is diminished.

Massey (Massey 1997) discusses three approaches to rating of teams based on the game results. The first approach is based on a linear additive model where the expected value of the victory margin is the difference between the teams' ratings with the home court advantage added. Then, the teams' ratings are solved by minimizing the sum of the squared errors when compared to the scores of the complete season. In so called maximum likelihood rating, the probability of the first team winning a game is relative to the ratio of the first team's rating and the sum of the two teams ratings. The ratings of the teams are solved by searching for the values that maximize the total likelihood of the observed game outcomes. The third discussed approach is the Elecs rating method that models win probabilities by applying continuous time Markov chains. In Elecs rating, the ratio of two teams' ratings approximates the ratio of their number of wins in a conceptually infinite series of games played between the teams.

Bassett (Bassett 2007) proposes an alternative for the linear additive model by introducing quantile regression to rating of sports teams. In quantile regression, the point differentials are not used to estimate the team ratings as the expected values of team strengths but instead the random variables representing the teams strengths are modeled as a collection of quantiles of their probability distribution. The quantile functions include, as a special case, the median equivalent to the expected victory margin of a game, i.e., the predictions are based on the median of the probability distribution of the point differential. The quantile regression

[2]The meaningless playing time that takes place after the winner of the game has been determined is often referred to as "garbage time".

approach reproduces the entire probability distribution of the team strengths which can be used in predicting game outcomes, e.g. for betting purposes. The article illustrates the approach with an application to NFL teams. Koenker and Bassett (Koenker and Bassett Jr. 2010) apply quantile regression to predicting the NCAA tournament results and calculating different betting opportunities against the point spreads offered by the bookmakers.[3]

There are several optimization formulations for the ranking and rating of teams. In general, the optimization approaches aim to find a ranking or rating of teams that minimizes/maximizes an objective function that measures the correspondence between the ordering of teams and the point differentials observed in the past. Such objectives include, e.g., the maximization of the number of games won by the higher ranked team and the minimization of the distance between the point differentials predicted by the rating and the observed point differentials. Coleman (Coleman 2005) proposes a ranking that is based on minimizing game outcome violations, i.e., the number of times a past game's winner is ranked behind its loser. This results in an optimization problem that is solved using mixed-integer programming. Govan and Langville (Govan et al. 2009) present an iterative approach to rating of teams that produces an offensive and a defensive rating for each team based on the observed point differentials. The final ranking of the teams is then calculated as the ratio of these two ratings. Cassady et al. (Cassady et al. 2005) formulate a quadratic assignment problem where the teams are rated so that their relative distances match with the observed point differentials as well as possible. In the article, a heuristic solution for the resulting optimization problem is obtained by employing a genetic algorithm.

In addition to the linear models and optimization approaches, various completely different methodologies have also been used in ranking and rating of teams based on observed game outcomes. Park and Newman (Park and Newman 2005) present the collegiate football system as a directed network where the arcs of the network represent one teams victory over another. Then, the structure of the network and the resulting "indirect victories" are used to calculate the teams' rankings. On the other hand, Rotshtein et al. (Rotshtein et al. 2005) propose a fuzzy model with

[3]The point spread is essentially a handicap towards the underdog. In so called spread betting, the point spread is determined by the bookmaker as the difference between the levels of performance of the two teams playing. Then, the customer bets her money (on equal odds) on either "the favorite will win by at least x points" or "the favorite will lose or win by less than x points." See, e.g., Bassett (1981).

genetic and neural tuning for predicting the outcome of a soccer match from the previous results of both teams.

In the previous models, the results of a season are considered as a whole. In order to follow the teams' level of play in real time, dynamic models have been presented. Glickman and Stern (Glickman and Stern 1998) present a dynamic model for NFL using a state-space representation. In the model, the observations of the system state are acquired a linear additive model including the home field advantage. The actual team rating are solved using Markov Chain Monte Carlo method and Bayesian statistics. Knorr-Held (Knorr-Held 2000) considers the problem of dynamically rating sports teams on the basis of results from the German soccer league Bundesliga as well as from the NBA. In the article, the leagues are presented as state-space models where the teams level of performance follows a specific constrained random walk. The presented method uses a cumulative link model for the observed outcomes of the games and takes a Bayesian approach to the estimation of the system state with an extended Kalman filter and a smoothing algorithm.

There exists many models based on the available expert rankings of basketball teams, viz., the seeding order of the NCAA tournament. Schwertman et al. (Schwertman et al. 1991) introduce three ad hoc probability models to predict the four finalists in the NCAA basketball tournament based on the seeding of the teams. Schwertman et al. (Schwertman et al. 1996) then consider the use of eight additional logistic and ordinary least squares regression models to further develop the probability models aimed for the same purpose. Smith and Schwertman (Smith and Schwertman 1999) study the predictive power of expert rankings using regression analysis and it is discovered that the margin of victory in NCAA basketball tournament games can be predicted fairly reliably based on the teams' seeding. Caudill and Godwin (Caudill and Godwin 2002) apply a skewed logit model with heterogenous skewness to predicting the NCAA basketball tournament based on the teams' seeding. Coleman and Lynch (Coleman and Lynch 2009) analyzed the statistical information used by the NCAA tournament selection committee in the seeding of basketball teams and found eight relevant predictors for team performance in the large data set.

On the other hand, the performances of mathematical models have also been compared to the expert rankings in the literature. Boulier and Stekler (Boulier and Stekler 1999) use probit regression with difference of rankings as the predictor to model the outcomes of NCAA basketball games as well as tennis matches. It is concluded that the rankings by are useful predictors by themselves and their predictive performance can be improved using probit regression. In Ref. (Boulier and Stekler 2003), the previously mentioned analysis is continued by predicting the point differentials of NFL games using the NFL power scores published by *The New York Times*, the betting market, and the opinions of the sports editor of *The New York Times*. The article concludes that the best predictions are given by the betting market and followed by the probit predictions based on the power scores. Kvam and Sokol (Kvam and Sokol 2006) combine logistic regression and Markov chain modeling in order to predict the outcomes of NCAA tournament games given the scores of the regular season games. In a comparative analysis, the presented method outperforms other common methods, including the NCAA tournament seedings as well as the RPI, Sagarin, and Massey ratings.

In addition to previously mentioned approaches, there are several articles where the inputs of the models are something other than simple game scores. Stern (Stern 1991) studies the probability of team winning in football game using a linear additive model based on the point spreads offered by the betting bookmakers and a normal distribution. Carlin (Carlin 1996) uses a similar Gaussian model to predict the teams' probability of winning the NCAA tournament using the information given by the point spreads. West (West 2006) proposes a rating method based on ordinal logistic regression that was designed to predict the teams' success in the NCAA tournament. The proposed method involves the teams' winning percentages, point differentials, Sagarin ratings as well as their performance against the top rated teams. Parker (Parker 2010) applies multinomial logistic regression to basketball teams' points per possession data[4]. The teams' productivity is then used to rank NCAA basketball teams and predict probabilities of winning future games.

1.3. **Objectives of the Presented Work.** This Thesis presents a mathematical algorithm that is aimed for ranking and rating of basketball teams according to their level of performance, i.e., the observed point differentials up to the current date. The algorithm is based on the state-space model for a system representing a basketball league (see, e.g., Glickman and Stern 1998, Knorr-Held 2000). The estimation of the system state is done in accordance to a linear additive model

[4]Points per possession is an advanced basketball statistic that measures the productivity of the teams' offence using the average number of points the team has scored per the number of ball possessions. Note that the number of possessions has to be calculated from the more elementary statistics using an approximative formula.

for the point differentials including the home court advantage. This approach is very common in literature (see,e.g., Bassett 1997, Carlin 1996, Harville 1977, 1980, 2003, Massey 1997, Schwertman et al. 1996, Stefani 1980, Stern 1991, Parker 2010). The current estimate of the system state is updated using a recursive Kalman filter algorithm (Kalman 1960, Hamilton 1994) when new scores are observed. The Kalman filter algorithm produces estimates of the system state that minimize the mean square error. Therefore, the estimates produced by the Kalman filter algorithm are analogous with those produced by the optimization approaches involving the minimization least squares errors (Stefani 1980, Bassett 1997, Massey 1997, Harville 2003). The state estimates also produce a method for rating the teams, i.e., measurable information about the relative differences between the strengths of the teams. The updated state estimates are applicable for prediction of scores of future games. This is used for validating and confirming the accuracy of the produced ratings as a valid rating should also have predictive power, i.e., ”the better teams” should be able to beat ”the weaker teams”. One should note that the aim of this Thesis is not to predict the results of entire playoff series (Niemi 2005) as the dynamics of a payoff series include various complications that are beyond the scope of this study (see, e.g., Bassett and Hurley 1998). On the other hand, this Thesis omits entirely the possibility of generating optimal or profitable strategies for betting in general (see, e.g., Marttinen 2001, Spann and Skiera 2008) or for betting pools (see, e.g., Hu and Zidek 2004).

1.4. **Structure of the Thesis.** Chapter 2 presents the state-space model for a system representing a basketball league as well the estimation of the system state using the Kalman filter algorithm. In Section 2.1, the state-space model and a general Kalman filter algorithm are introduced. Section 2.2 presents a state-space model for a basketball league and the application of the Kalman filter algorithm to the estimation of team ratings from observed point differentials. In Section 2.5, it shown how the parameters of the model are estimated from data using maximum likelihood. The calibration of the model using expert assessment in cases where history data is unavailable is discussed is also Section 2.5. Sections 2.3 and 2.4 present some mathematical properties of the ratings produced by the Kalman filter algorithm as well as some qualitative characteristics of the introduced rating approach. The utilization and validation of the algorithm are briefly discussed in Sections 2.6 and 2.7. In Chapter 3, the rating of basketball teams using a Kalman filter is illustrated by analyzing data from NBA Season 2009-2010. Some alternative models and ideas for further development are discussed Section 4. Finally, the conclusions are give in Chapter 5.

2. Kalman Filter Algorithm for Analysis of a Basketball League

2.1. **General Kalman Filter.** Kalman filter is recursive algorithm for estimating the unobservable true state of a system from observations that include random noise (Kalman 1960, Harvey 1989, Hamilton 1994, Bertsekas 2005). The Kalman filter is based on a state-space model for the system under consideration. The Kalman filter combines the possibly contradictory observations giving both the optimal estimates for the current state of the system as well as optimal predictions that minimize the expected quadratic prediction error. The results of the algorithm can be used for predicting the future state of the system and the future observations. If the noise included in the observations is Gaussian white noise, the Kalman filter gives maximum likelihood estimates and predictions for the system state (Hamilton 1994).

2.1.1. *State-Space Model.* Let $\mathbf{x_t} \in \mathbb{R}^n$ be the state of the system at time t. The state-space representation for the system is the following system of equations:

$$\mathbf{x_{t+1}} = \mathbf{A_t x_t} + \mathbf{B_t u_t} + \mathbf{v_{t+1}} \tag{1}$$

$$\mathbf{y_t} = \mathbf{C_t x_t} + \mathbf{D_t u_t} + \mathbf{w_t} \tag{2}$$

where equation (1) is known as the state equation and equation (2) as the observation equation (e.g., Bertsekas 2005, Hamilton 1994). In the equations, $\mathbf{u_t} \in \mathbb{R}^l$ is the control applied to the system at time t, and $\mathbf{y_t} \in \mathbb{R}^m$ is the vector of observations made at time t. The matrix $\mathbf{A_t} \in \mathbb{R}^{n\times n}$ describes the evolution of the system state if no control is applied and $\mathbf{C_t} \in \mathbb{R}^{m\times n}$ is the observation matrix. On the other hand, the matrix $\mathbf{B_t} \in \mathbb{R}^{n\times l}$ models the effect of applied control on the system state and $\mathbf{D_t} \in \mathbb{R}^{m\times l}$ the effect of the applied control on the made observation.

The vectors $\mathbf{v_{t+1}} \in \mathbb{R}^n$ and $\mathbf{w_t} \in \mathbb{R}^m$ are random white noise affecting evolution of the system state and observation vector, respectively:

$$E(\mathbf{v_t v_\tau^T}) = \begin{cases} \mathbf{M_t} & \text{if} \quad t = \tau \\ \mathbf{0} & \text{if} \quad t \neq \tau \end{cases} \tag{3}$$

$$E(\mathbf{w_t w_\tau^T}) = \begin{cases} \mathbf{N_t} & \text{if} \quad t = \tau \\ \mathbf{0} & \text{if} \quad t \neq \tau \end{cases}, \tag{4}$$

where $\mathbf{M_t} \in \mathbb{R}^{n\times n}$ and $\mathbf{N_t} \in \mathbb{R}^{m\times m}$ are positive definite covariance matrixes. The noise vectors are assumed to be uncorrelated for all lags, i.e., $E(\mathbf{v_t w_\tau^T}) = 0$ for all values of t and τ. The random vectors $\mathbf{v_t}$ and $\mathbf{w_t}$ are interpreted as follows. Vector $\mathbf{v_t}$ represents the uncertainty about the time evolution of the system, i.e., the random variations in the state of the system, while vector $\mathbf{w_t}$ is the random noise included in the observations made about the system state.

2.1.2. *Kalman Filter Algorithm.* Kalman filter algorithm is used for calculating linear least squares estimates of the state vector on the basis of data observed up to time instant t (e.g., Bertsekas 2005, Hamilton 1994). The estimation is done recursively, i.e., the estimate is updated whenever new observations are made. Now, it is assumed that there is no control vector $\mathbf{u_t} \equiv \mathbf{0}$. Then, the Kalman filter algorithm is as follows:

(i) Initialization: The initial estimate of the system state is set as $\mathbf{\hat{x}_{0|0}} = E(\mathbf{x_0})$ and a positive definite covariance matrix $\mathbf{\hat{\Sigma}_{0|0}}$ is defined to represent the uncertainty related to the initial estimate of the system state.

(ii) Prediction of system state: The state equation (1) is used to predict the system state for the following time instant giving

$$\mathbf{\hat{x}_{t|t-1}} = \mathbf{A_{t-1}}\mathbf{\hat{x}_{t-1|t-1}}.$$

(iii) Predictive covariance matrix: The uncertainty related to the prediction is given by the predictive covariance matrix

$$\mathbf{\hat{\Sigma}_{t|t-1}} = \mathbf{A_{t-1}}\mathbf{\hat{\Sigma}_{t-1|t-1}}\mathbf{A_{t-1}}' + \mathbf{M_{t-1}}.$$

(iv) Posterior covariance matrix: The covariance matrix is updated to include the effect of the made observations

$$\mathbf{\hat{\Sigma}_{t|t}} = \mathbf{\hat{\Sigma}_{t|t-1}} - \mathbf{\hat{\Sigma}_{t|t-1}}\mathbf{C_t}'(\mathbf{C_t}\mathbf{\hat{\Sigma}_{t|t-1}}\mathbf{C_t}' + \mathbf{N_t})^{-1}\mathbf{C_t}\mathbf{\hat{\Sigma}_{t|t-1}}.$$

(v) Updated estimate of the system state: The estimate of the system state is updated based on the previous estimate and the made observations $\mathbf{y_t}$

$$\mathbf{\hat{x}_{t|t}} = \mathbf{\hat{x}_{t-1|t-1}} + \mathbf{\hat{\Sigma}_{t|t}}\mathbf{C_t}'\mathbf{N_t}^{-1}(\mathbf{y_t} - \mathbf{C_t}\mathbf{\hat{x}_{t-1|t-1}}).$$

The algorithm produces linear least squares estimates for the system state, i.e., $\mathbf{\hat{x}_{1|1}}, \mathbf{\hat{x}_{2|2}}, \ldots, \mathbf{\hat{x}_{t|t}}$, in accordance to the made observations. The covariance matrixes $\mathbf{\hat{\Sigma}_{t|t}}$ represent the accuracy of the current estimate of the system state. The main attractive feature of the Kalman filter algorithm is that the estimate $\mathbf{\hat{x}_{t+1|t}}$ is obtained by means of a simple equation that involves only the previous estimate $\mathbf{\hat{x}_{t|t-1}}$ and the new measurement $\mathbf{y_t}$, but none of the past measurements $\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_{t-1}}$.

2.2. **Model for Basketball League.**

2.2.1. *System State.* Now, in a basketball setting, the system under consideration is a basketball league that consists of n teams. The state of the system is a vector describing the relative strengths of the teams at time t, denoted by $\mathbf{x_t} \in \mathbb{R}^n$. In this presentation, the ith element of the state vector, $\mathbf{x_t}(i)$ where $i \in \{1, \ldots, n\}$, gives the strength of the ith team in relation to the "average team". With a suitable selection of the model parameters, the "average team" is represented by the value 0. This is further discussed in Section 2.3.

The strengths of the teams cannot be directly observed. Instead, the teams strengths are measured in relation to other teams by observing the scores of games played between the teams. In the presented model, it is assumed that the point differential of a given game depends linearly on the strength of the teams, home court advantage, and random variation, viz., Gaussian white noise, resulting from the non-deterministic nature of the game of basketball. For now, all model parameters, including the home court advantage, are taken as time invariant constants which are equal for all teams. Other modeling possibilities such as team dependent home advantage as well as the inclusion of the effect of rest in the model are discussed in Chapter 4.

2.2.2. *System Dynamics.* The strength of a team depends on numerous factors such as coaching (Colquitt et al. 2007); development and decline of individual players, injuries, and player transactions between teams. The team performance depends also on many intangible factors such as "team spirit" that are hard to model mathematically (despite some recent advances, see, e.g., Kraus et al. 2010). For modeling purposes, most of these factors can be assumed to be more or less independent and the evolution of the state of the system is presented as a (discrete time variant of) Brownian motion. In other words, the changes in the team strengths from the time instant t to the time instant $t+1$ are modeled as a random vector $\mathbf{v_{t+1}}$ following the multivariate normal distribution, $\mathbf{v_{t+1}} \sim N(\mathbf{0}, \mathbf{M})$. The changes in strengths of individual teams are assumed to be independent and identically distributed so that the covariance matrix is diagonal matrix $\mathbf{M} = \sigma_v^2 \cdot \mathbf{I}$, where σ_v^2 is the variance of the change in the strength of a single team. In theory, player transactions between teams should benefit some teams on at other teams' expense which could raise negative correlation between the changes in team strengths. Now, such additions are considered beyond the scope of this Thesis and left as a topic for future research (see, Chapter 4).

There are no deterministic trends in the model and it is defined that $\mathbf{A_t} = \mathbf{I}$ in equation (1). In other words, the expected value of the system state remains constant in one time step. No controls are applied on the system.[5] Therefore, the respective matrixes in equations (1) and (2) can be defined as zero, $\mathbf{B_t} = \mathbf{0}$ and $\mathbf{D_t} = \mathbf{0}$. Based on these assumptions, the state equation 1 takes the form

$$\mathbf{x_{t+1}} = \mathbf{x_t} + \mathbf{v_{t+1}}. \tag{5}$$

Note that the resulting system is in fact a discrete-time Brownian motion (Hamilton 1994) where the time instant t represents the tth game day.

2.2.3. *Observations of System State.* The state of the system is observed indirectly by following the scores of the games played between teams. The league schedule, i.e., the m games played on the tth day, is represented by the matrix $\mathbf{C_t} \in \mathbb{R}^{m\times n}$. If the team i is the home team in the kth game played on the tth day, $k \in \{1, \ldots, m\}$, $\mathbf{C_t}(i, k) = 1$. Similarly, if the team j is the visiting team in the kth game, $\mathbf{C_t}(j, k) = -1$. The rest of the matrix is set equal to zero, i.e., the outcome of the kth game is independent of the other teams' strengths. For example, in a league with $n = 4$ teams, where on the tth day the team 1 hosts team 2 and team 3 hosts team 4, the schedule matrix is

$$\mathbf{C_t} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

The point differentials of the games played on the tth day, represented by the vector $\mathbf{y_t} \in \mathbb{R}^m$, are assumed to depend additively from the difference of the teams' strengths resulting in the model

$$\mathbf{y_t} = \mathbf{C_t}\mathbf{x_t} + \mu \cdot \boldsymbol{\iota_t} + \mathbf{w_t} \tag{6}$$

where $\mu \in \mathbb{R}$ is a constant representing the home court advantage, $\boldsymbol{\iota_t} = (1\,1\,1\,\ldots\,1)' \in \mathbb{R}^m$, and $\mathbf{w_t} \sim N(\mathbf{0}, \mathbf{N_t})$ is the random variation in the point differentials of respective games. The random factors affecting individual games are assumed independent and identically distributed so that $\mathbf{N_t} = \sigma_w^2 \cdot \mathbf{I}$.

---

[5]Certain unsportsmanlike and even illegal actions might be seen as controls applied to the system. Such activities might include, e.g., biased (or bribed) referees, ”tanking” where teams lose intentionally to improve their position in draft of future players (e.g., Soebbing and Humphreys 2011), or players involved in a some sort of point shaving scheme where the exact point differential of a game is intentionally manipulated for betting purposes (e.g., Gibbs 2007).

For the kth game on tth day where team j visits team i, equation (6) simplifies to

$$\mathbf{y_t}(k) = \mathbf{x_t}(i) - \mathbf{x_t}(j) + \mu + \mathbf{w_t}(k). \tag{7}$$

If it is observed that $\mathbf{y_t}(k) > 0$, the home team i wins the kth game. Otherwise, the game results in a win for the visiting team. Note that due to the home court advantage and the random variation included in the model, the observation $\mathbf{y_t}(k) > 0$ does not necessarily mean that team i is better than team j. In other words, the observation $\mathbf{y_t}(k) > 0$ does not necessarily mean that $\mathbf{x_t}(i) > \mathbf{x_t}(j)$. On the contrary, a winning margin for the home team that is smaller than the home court advantage would actually imply that $\mathbf{x_t}(i) < \mathbf{x_t}(j)$.

2.2.4. *Kalman Filter Algorithm.* The Kalman filter algorithm is used in the updating of the estimate of the system state at time t, $\hat{\mathbf{x}}_{\mathbf{t|t}} \in \mathbb{R}^n$, and the respective posterior covariance matrix of the estimate $\hat{\mathbf{\Sigma}}_{\mathbf{t|t}} \in \mathbb{R}^{n\times n}$. The algorithm is initialized with the prior estimate of the system state, $\hat{\mathbf{x}}_{\mathbf{0|0}}$, and the related covariance matrix, $\hat{\mathbf{\Sigma}}_{\mathbf{0|0}}$. The estimates and covariance matrixes are updated recursively according to the observed game outcomes $\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_t}$.

(i) Initialization: The recursive Kalman filter algorithm is initialized by giving the initial estimate for system state, i.e., $\hat{\mathbf{x}}_{\mathbf{0|0}} = E(\mathbf{x_0})$. The uncertainty related to the initial estimate is represented by the covariance matrix,

$$\hat{\mathbf{\Sigma}}_{\mathbf{0|0}} = \sigma_0^2 \cdot \mathbf{I}. \tag{8}$$

The matrix $\hat{\mathbf{\Sigma}}_{\mathbf{0|0}}$ is diagonal making the initial estimates independent. The variance parameter σ_0^2 measures the confidence in the initial estimates of the system state.

(ii) Prediction of system state: The predicted system state at time t as given at time $t-1$ is calculated using equation (5) which gives

$$\hat{\mathbf{x}}_{\mathbf{t|t-1}} = \hat{\mathbf{x}}_{\mathbf{t-1|t-1}}.$$

Thus, according to the made assumptions, the prediction is equal to the latest estimate of the system state.

(iii) Predictive covariance matrix: The uncertainty related to the predicted system state is calculated as the covariance matrix of the predicted system state

$\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}} \in \mathbb{R}^{n \times n}$ at time t as given at time $t-1$

$$\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}} = \hat{\mathbf{\Sigma}}_{\mathbf{t-1|t-1}} + \mathbf{M}. \tag{9}$$

In practice, the effect of the multinormal noise with covariance matrix $\mathbf{M} = \sigma_v^2 \cdot \mathbf{I}$ is added to the covariance matrix to represent the random changes in the teams' strengths between time instants $t-1$ and t. The covariance matrix $\mathbf{M}$ is diagonal and the changes in team strengths are assumed independent. Variance parameter σ_v^2 measures the magnitude of variation in the team strengths.

(iv) Posterior covariance matrix: The covariance matrix is updated to include the effect of the observed scores

$$\hat{\mathbf{\Sigma}}_{\mathbf{t|t}} = \hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}} - \hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\mathbf{C_t}'(\mathbf{C_t}\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\mathbf{C_t}' + \mathbf{N_t})^{-1}\mathbf{C_t}\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}. \tag{10}$$

(v) Updated estimate of the system state: The observations made according to equation (6) are used to update the estimate of the system state at time t, $\hat{\mathbf{x}}_{\mathbf{t|t}}$.

$$\begin{aligned} \hat{\mathbf{x}}_{\mathbf{t|t}} &= \hat{\mathbf{x}}_{\mathbf{t-1|t-1}} + \underbrace{\hat{\mathbf{\Sigma}}_{\mathbf{t|t}}\mathbf{C_t}'\mathbf{N_t}^{-1}}_{=:\mathbf{K_t}} \underbrace{(\mathbf{y_t} - \mathbf{C_t}\hat{\mathbf{x}}_{\mathbf{t-1|t-1}} - \mu \cdot \boldsymbol{\iota}_t)}_{=:\mathbf{d_t}} \\ &= \hat{\mathbf{x}}_{\mathbf{t-1|t-1}} + \mathbf{K_t}\mathbf{d_t}, \end{aligned} \tag{11}$$

where $\mathbf{K_t} \in \mathbb{R}^{n \times m}$ is the so called the optimal Kalman gain matrix (see, e.g., Hamilton 1994, pp. 380) and $\mathbf{d_t} \in \mathbb{R}^{m \times 1}$ is the difference between the expected and observed point differentials. Note that the observed differences between the prediction given by the algorithm and the observations made, i.e., $\mathbf{d_t}$, are used as the innovations that dictate the evolution of the estimated system state.

2.2.5. *Initialization of the Algorithm.* The recursive Kalman filter algorithm is initialized by giving the initial estimate for system state. According to Ref. (Hamilton 1994, pp. 378), the proposed initial value is the expected value of system state which gives

$$\hat{\mathbf{x}}_{\mathbf{0|0}} = E(\mathbf{x_0}).$$

If no information about the teams' strengths is available, all teams are assumed equal and the initial state estimate is set as

$$\hat{\mathbf{x}}_{\mathbf{0|0}} = \mathbf{0}.$$

The uncertainty related to the initial estimate is represented by the covariance matrix,

$$\hat{\mathbf{\Sigma}}_{\mathbf{0|0}} = \sigma_0^2 \cdot \mathbf{I}.$$

Here, larger values of the variance parameter σ_0^2 imply greater uncertainty about the initial state estimate. In maximum likelihood estimation of the parameters of the model (see, Section 2.5), parameter σ_0^2 determines the amount of variation that takes place at the beginning of the season. If the parameters are determined based on expert analysis, the variance parameter can also be seen as the accuracy of the expert assessment determining $\hat{\mathbf{x}}_{\mathbf{0|0}}$.

If the analysis of the basketball league is continued from one season to another, the final state of the previous season can be used as the initial estimate of the following season. In such a case, the possible changes in team rosters and other unknown factors should be accounted for by adding considerable random noise to the state estimate. In practice, the uncertainty related to off-season activities could be modeled by increasing the variance of the state estimates during the off season.

2.3. Mathematical Properties of the Model.

2.3.1. *Conditional Probability Distributions.* If the assumptions presented in Section 2.2 hold, the conditional probability distribution of the system state, i.e., the true strengths of the teams, is

$$\mathbf{x_t} | \, \mathcal{Y}_t \sim N(\hat{\mathbf{x}}_{\mathbf{t|t}}, \hat{\mathbf{\Sigma}}_{\mathbf{t|t}}), \tag{12}$$

where $\mathcal{Y}_t := (\mathbf{y_1}', \mathbf{y_2}', \ldots, \mathbf{y_t}')'$ denotes the observations made up to time instant t. Because the initial state and the innovations $(\mathbf{v_1}', \mathbf{v_2}', \ldots, \mathbf{v_t}', \mathbf{w_1}', \mathbf{w_2}', \ldots, \mathbf{w_t}')'$ are multivariate Gaussian, the conditional probability distribution used for predicting future scores is

$$\mathbf{y_{t+1}} | \, \mathcal{Y}_t \sim N(\underbrace{\mathbf{C_{t+1}}\hat{\mathbf{x}}_{\mathbf{t+1|t}} + \mu \cdot \boldsymbol{\iota}_{\boldsymbol{t+1}}}_{=:\hat{\mathbf{y}}_{\mathbf{t+1}}}, \underbrace{\mathbf{C_{t+1}}\hat{\mathbf{\Sigma}}_{\mathbf{t+1|t}}\mathbf{C_{t+1}}' + \mathbf{N_{t+1}}}_{=:\hat{\mathbf{O}}_{\mathbf{t+1}}}).$$

2.3.2. *Observability.* A n-dimensional state vector $\mathbf{x_t}$ is called observable if it can be determined exactly given the observations $\mathbf{y_t}, \ldots, \mathbf{y_{t+n-1}}$ (Harvey 1989, pp. 115). This does not hold for the state of the basketball league, as defined in Section 2.2, which cannot be uniquely determined because only the differences of team strengths are observed. For example, two systems $\mathbf{x_t}$ and $\widetilde{\mathbf{x}}_{\mathbf{t}} = \mathbf{x_t} + c \cdot \boldsymbol{\iota}$, where $\mathbf{c} \in \mathbb{R}$ is a constant and $\boldsymbol{\iota} = (1 \; 1 \; 1 \; \ldots \; 1)' \in \mathbb{R}^n$, cannot be distinguished from each other by observing the differences between elements of the state vector as

$$\widetilde{\mathbf{x}}_{\mathbf{t}}(i) - \widetilde{\mathbf{x}}_{\mathbf{t}}(j) = \mathbf{x}_{\mathbf{t}}(i) + c - \mathbf{x}_{\mathbf{t}}(j) - c = \mathbf{x}_{\mathbf{t}}(i) - \mathbf{x}_{\mathbf{t}}(j).$$

In practice, this means that two teams playing in completely disjoint leagues cannot be compared based on intra-league games only. On the other hand, within an individual league the sum of the initial team ratings can be set as equal to zero, i.e., $\boldsymbol{\iota}'\mathbf{x_0} = \mathbf{0}$, resulting in unambiguous ratings for this league. Most importantly, this somewhat arbitrary constraint does not affect the ranking of the teams. Additionally, the non-uniqueness of the observations has no effect on the maximum likelihood estimation of the model parameters discussed in Section 2.5.

2.3.3. *Optimality.* The Kalman gain matrix $\mathbf{K_t}$ in equation (11) is determined in order to minimize the mean square error of the state estimates given by the Kalman filter algorithm (Hamilton 1994, pp. 377). That is, the state estimates produced by the Kalman filter are linear projections of the observed point differentials that minimize $E\left((\mathbf{x_t} - \hat{\mathbf{x}}_{\mathbf{t|t}})(\mathbf{x_t} - \hat{\mathbf{x}}_{\mathbf{t|t}})'\right)$ and $E\left((\mathbf{y_{t+1}} - \hat{\mathbf{y}}_{\mathbf{t+1}})(\mathbf{y_{t+1}} - \hat{\mathbf{y}}_{\mathbf{t+1}})'\right)$. This derivation does not involve the Gaussianity assumption. In fact, the state estimates $\hat{\mathbf{x}}_{\mathbf{t|t}}$ and predictions $\hat{\mathbf{y}}_{\mathbf{t+1}}$ are optimal within the set of linear unbiased estimators in $\mathcal{Y}_t$ (Harvey 1989, pp. 105). That is, for any other linear unbiased estimators $\tilde{\mathbf{x}}_{\mathbf{t|t}}$ and $\tilde{\mathbf{y}}_{\mathbf{t+1}}$ with corresponding covariance matrixes $\tilde{\mathbf{\Sigma}}_{\mathbf{t|t}}$ and $\tilde{\mathbf{O}}_{\mathbf{t+1}}$, respectively, it holds that $\tilde{\mathbf{\Sigma}}_{\mathbf{t|t}} - \hat{\mathbf{\Sigma}}_{\mathbf{t|t}}$ and $\tilde{\mathbf{O}}_{\mathbf{t+1}} - \hat{\mathbf{O}}_{\mathbf{t+1}}$ are positive semidefinite matrixes.

Furthermore, if the joint distribution of the initial state and the innovations $(\mathbf{v_1}', \mathbf{v_2}', \ldots, \mathbf{v_t}', \mathbf{w_1}', \mathbf{w_2}', \ldots, \mathbf{w_t}')'$ is multivariate Gaussian, the estimates given by the Kalman filter algorithm are optimal among any functions of $\mathcal{Y}_t$ (Hamilton 1994, pp. 385). It can also be shown that, in case of Gaussian initial state and innovations, the estimates produced by the Kalman filter are the maximum likelihood estimates for both the current state and the predicted future scores (Hamilton 1994, pp. 385).

2.3.4. *Effect of Parameter Values.* In the following, the effects of the parameter values are explored using a simple example where home team 1 hosts visiting team 2 on the tth day of the season. Assume that the current estimate of the system state is $\hat{\mathbf{x}}_{\mathbf{t-1|t-1}} = (x_1\ x_2)'$ and the predictive covariance matrix is

$$\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}.$$

The schedule of the tth day is represented by $\mathbf{C_t} = (1\ \ -1)$ and the variance related to its point differentials is $\mathbf{N_t} = \sigma_v^2$.

When the point differential of the game is observed to be $\mathbf{y_t} = y \in \mathbb{R}$, the estimate of the system state is updated using equation (11). Noting that $\mathbf{C_t}\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\mathbf{C_t}' = 2(1-\rho)\sigma^2$, the Kalman gain matrix is

$$\begin{aligned}
\mathbf{K_t} &= \hat{\mathbf{\Sigma}}_{\mathbf{t|t}}\mathbf{C_t}'\mathbf{N_t}^{-1} \\
&= \left(\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}} - \hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\mathbf{C_t}'(\mathbf{C_t}\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\mathbf{C_t}' + \mathbf{N_t})^{-1}\mathbf{C_t}\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\right)\mathbf{C_t}'\mathbf{N_t}^{-1} \\
&= \frac{1}{\sigma_v^2}\left(\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\mathbf{C_t}' - \frac{2(1-\rho)\sigma^2\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\mathbf{C_t}'}{2(1-\rho)\sigma^2+\sigma_v^2}\right) \\
&= \frac{1}{2(1-\rho)\sigma^2+\sigma_v^2}\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\mathbf{C_t}' = \frac{1}{2(1-\rho)\sigma^2+\sigma_v^2}\begin{pmatrix} (1-\rho)\sigma^2 \\ -(1-\rho)\sigma^2 \end{pmatrix} \\
&= \frac{1}{2+\sigma_v^2/((1-\rho)\sigma^2)}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{1}{2+r^2}\begin{pmatrix} 1 \\ -1 \end{pmatrix},
\end{aligned}$$

where $r^2 := \sigma_v^2/((1-\rho)\sigma^2) \geq 0$. It should be noted that r^2 can be interpreted as the ratio of the uncertainty related to the point differential of the game and the uncertainty related to the predicted state of the system. This can be seen as the amount of information included in the observation related to the existing state information. That is, if the state is known accurately, σ^2 is small, and if there is a lot of noise in the observation, the observing the value of the point differential is less informative and σ_v^2 is large.

The difference between the predicted and observed point differential of the game is $\mathbf{d_t} = y - x_1 + x_2 - \mu \in \mathbb{R}$. Together with $\mathbf{K_t}$, this is used to update the estimate of the state:

$$\begin{aligned}
\hat{\mathbf{x}}_{\mathbf{t|t}} &= \hat{\mathbf{x}}_{\mathbf{t-1|t-1}} + \mathbf{K_t}\mathbf{d_t} \\
&= \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{y - x_1 + x_2 - \mu}{2+r^2}\begin{pmatrix} 1 \\ -1 \end{pmatrix} \\
&= \frac{1}{2+r^2}\begin{pmatrix} y + (1+r^2)x_1 + x_2 - \mu \\ -y + x_1 + (1+r^2)x_2 + \mu \end{pmatrix}.
\end{aligned}$$

From this result, several observations can be made. First, it is seen that $\boldsymbol{\iota}'\hat{\mathbf{x}}_{\mathbf{t|t}} = x_1 + x_2 = \boldsymbol{\iota}'\hat{\mathbf{x}}_{\mathbf{t-1|t-1}}$, i.e., the sum of ratings remains constant regardless of the outcome of the game y or the values of the model parameters. Second, if the outcome of the game matches its predicted value, i.e., $y = x_1 - x_2 + \mu$, the estimate for the state of the system remains constant $\mathbf{\hat{x}_{t|t}} = \mathbf{\hat{x}_{t-1|t-1}}$.

Third, the correction made to the state estimate is inversely proportional to r^2. That is, the highly uncertain observations give little new information about the state of the system. If the variance σ_v^2 is large compared to $(1-\rho)\sigma^2$ and $r^2 \to \infty$, the estimate of the system state is not changed regardless of the point differential of the game, and $\mathbf{\hat{x}_{t|t}} = \mathbf{\hat{x}_{t-1|t-1}}$. On the other hand, if variance σ_v^2 is small in comparison with $(1-\rho)\sigma^2$ and $r^2 \to 0$,

$$\mathbf{\hat{x}_{t|t}} = \frac{1}{2}\begin{pmatrix} y + x_1 + x_2 - \mu \\ -y + x_1 + x_2 + \mu \end{pmatrix}.$$

In this case, the estimate for the state of the system is updated to match the outcome of the game exactly, so that $\mathbf{\hat{x}_{t|t}}(1) - \mathbf{\hat{x}_{t|t}}(2) = y$. In other words, if there is no uncertainty in the game results, the difference between teams' ratings is simply set equal to the outcome of their latest game.[6]

2.3.5. *Sum of Ratings.* In the presented state-space model, the uncertainties related to teams' initial state and changes in team strengths are represented by the covariance matrixes $\mathbf{\hat{\Sigma}_{0|0}}$ and $\mathbf{M}$ in equations (8) and (9), respectively. If the state of each team is known with equal accuracy, the variation of each team's performance is equal, and the changes in teams performance levels are independent, it holds that $\mathbf{\hat{\Sigma}_{0|0}} = \sigma_0^2 \cdot \mathbf{I}$ and $\mathbf{M} = \sigma_v^2 \cdot \mathbf{I}$. From this it follows that the sum of team ratings remain constant, i.e., $\boldsymbol{\iota}'\mathbf{\hat{x}_{t|t}} = \boldsymbol{\iota}'\mathbf{\hat{x}_{t-1|t-1}} = \ldots = \boldsymbol{\iota}'\mathbf{\hat{x}_{0|0}}$ where $\boldsymbol{\iota} = (1\ 1\ 1\ \ldots\ 1)' \in \mathbb{R}^n$. In other words, the corrections made to the team ratings sum to zero.

This is seen by first calculating the difference between consecutive team ratings from equation (11)

$$\mathbf{\hat{x}_{t|t}} - \mathbf{\hat{x}_{t-1|t-1}} = \mathbf{K_t d_t}.$$

and adding the changes in team strengths together resulting in

$$\boldsymbol{\iota}'(\mathbf{\hat{x}_{t|t}} - \mathbf{\hat{x}_{t-1|t-1}}) = \boldsymbol{\iota}'\mathbf{K_t d_t} = \boldsymbol{\iota}'\mathbf{\hat{\Sigma}_{t|t} C_t{}' N_t}^{-1} \cdot \mathbf{d_t}. \tag{13}$$

Now, it is shown by induction that $\boldsymbol{\iota}'\mathbf{\hat{\Sigma}_{t|t}} = \tilde{\alpha}\boldsymbol{\iota}'$, i.e., the sums of the columns of covariance matrix $\mathbf{\hat{\Sigma}_{t|t}}$ are equal at all time instants t.

[6]In Finnish basketball circles, this kind of naïve updating of implicit team ratings is playfully referred to as *jabastics* after Jari "Jaba" Suvanto.

For the initial covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{0|0}}$ in equation (8), it is seen that

$$\boldsymbol{\iota}'\hat{\mathbf{\Sigma}}_{\mathbf{0|0}} = \sigma_0^2\boldsymbol{\iota}'. \tag{14}$$

If it is assumed that $\boldsymbol{\iota}'\hat{\mathbf{\Sigma}}_{\mathbf{t-1|t-1}} = \alpha\boldsymbol{\iota}'$, i.e., the sum of every column of matrix $\hat{\mathbf{\Sigma}}_{\mathbf{t-1|t-1}}$ is equal to some $\alpha \in \mathbb{R}$. By equation (9), this holds also for the columns of the predictive covariance matrix

$$\boldsymbol{\iota}'\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}} = \boldsymbol{\iota}'(\hat{\mathbf{\Sigma}}_{\mathbf{t-1|t-1}} + \mathbf{M}) = \underbrace{(\alpha + \sigma_v^2)}_{=:\tilde{\alpha}}\boldsymbol{\iota}' = \tilde{\alpha}\boldsymbol{\iota}'.$$

When the covariance matrix is updated to include the new observations in equation (10), the sums of the columns of posterior covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{t|t}}$ are also equal

$$\begin{aligned} \boldsymbol{\iota}'\hat{\mathbf{\Sigma}}_{\mathbf{t|t}} &= \boldsymbol{\iota}'\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}} - \boldsymbol{\iota}'\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\mathbf{C_t}'(\mathbf{C_t}\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\mathbf{C_t}' + \mathbf{N_t})^{-1}\mathbf{C_t}\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}} \\ &= \tilde{\alpha}\boldsymbol{\iota}' - \tilde{\alpha}\underbrace{\boldsymbol{\iota}'\mathbf{C_t}'}_{=\mathbf{0}}(\mathbf{C_t}\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\mathbf{C_t}' + \mathbf{N_t})^{-1}\mathbf{C_t}\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}} = \tilde{\alpha}\boldsymbol{\iota}'. \end{aligned} \tag{15}$$

Together, equations (14) and(15) imply that $\boldsymbol{\iota}'\hat{\mathbf{\Sigma}}_{\mathbf{t|t}} = \tilde{\alpha}\boldsymbol{\iota}'$ for some $\tilde{\alpha}$ for all time instants t. Thus, it holds for all time instants t that

$$\boldsymbol{\iota}'\hat{\mathbf{\Sigma}}_{\mathbf{t|t}}\mathbf{C_t}' = \tilde{\alpha}\underbrace{\boldsymbol{\iota}'\mathbf{C_t}'}_{=\mathbf{0}} = \mathbf{0}.$$

The latter equality holds regardless of the value of $\tilde{\alpha}$ as the sum of each column of the matrix $\mathbf{C_t}'$ is zero.

Finally, from equation (13) it is seen that

$$\begin{aligned} \boldsymbol{\iota}'(\hat{\mathbf{x}}_{\mathbf{t|t}} - \hat{\mathbf{x}}_{\mathbf{t-1|t-1}}) &= \underbrace{\boldsymbol{\iota}'\hat{\mathbf{\Sigma}}_{\mathbf{t|t}}\mathbf{C_t}'}_{=\mathbf{0}}\mathbf{N_t}^{-1} \cdot \mathbf{d_t} = \mathbf{0} \\ \Rightarrow \boldsymbol{\iota}'\hat{\mathbf{x}}_{\mathbf{t|t}} &= \boldsymbol{\iota}'\hat{\mathbf{x}}_{\mathbf{t-1|t-1}}. \end{aligned} \tag{16}$$

Thus, when the uncertainty about all teams' level of play is the same, the sum of team ratings is constant. For example, if the sum of ratings is set equal to zero at the beginning of the season, it remains such for the duration of the season. On the other hand, if the variances on the diagonals of $\hat{\mathbf{\Sigma}}_{\mathbf{t|t}}$ or $\mathbf{M}$ are not constant, the ratings of teams with higher variances are corrected more drastically, and equation (16) does not hold.

2.3.6. *Dependence between Ratings.* The estimates of the teams' strengths are dependent as only the differences between team strengths are observed. In effect, the initially diagonal covariance matrix "fills out" accordingly to reflect the covariances between estimates. For example, if there are three teams in a basketball league and the predictive covariance matrix is given by

$$\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}} = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}.$$

Note that the covariance matrix is diagonal and the predicted state estimates of the individual teams are independent prior to the game played on the tth day. Assuming that team 1 hosts team 2, i.e., $\mathbf{C_t} = (1 \ \ -1 \ 0)$ and $\mathbf{N_t} = \sigma_v^2$. Then, noting that $\mathbf{C_t}\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\mathbf{C_t}' = 2\sigma^2$, the covariance matrix is updated according to (10) yielding

$$\begin{aligned} \hat{\mathbf{\Sigma}}_{\mathbf{t|t}} &= \hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}} - \hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\mathbf{C_t}'(\mathbf{C_t}\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\mathbf{C_t}' + \mathbf{N_t})^{-1}\mathbf{C_t}\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}} \\ &= \frac{1}{2\sigma^2 + \sigma_v^2}\begin{pmatrix} \sigma^4 + \sigma^2\sigma_v^2 & \sigma^4 & 0 \\ \sigma^4 & \sigma^4 + \sigma^2\sigma_v^2 & 0 \\ 0 & 0 & 2\sigma^4 + \sigma^2\sigma_v^2 \end{pmatrix} \\ &= \frac{1}{2 + \sigma_v^2/\sigma^2}\begin{pmatrix} \sigma^2 + \sigma_v^2 & \sigma^2 & 0 \\ \sigma^2 & \sigma^2 + \sigma_v^2 & 0 \\ 0 & 0 & 2\sigma^2 + \sigma_v^2 \end{pmatrix}. \end{aligned}$$

Clearly, it is seen that there now exists a positive correlation between teams 1 and 2 while team 3 remains independent of the other two.

The magnitude of the introduced correlation depends on the ratio of σ_v^2 and σ^2 (cf. Subsection 2.3.4). For example, if there is no noise in the observed point differential, it is seen that

$$\lim_{\sigma_v^2 \to 0} \hat{\mathbf{\Sigma}}_{\mathbf{t|t}} = \frac{1}{2}\begin{pmatrix} \sigma^2 & \sigma^2 & 0 \\ \sigma^2 & \sigma^2 & 0 \\ 0 & 0 & 2\sigma^2 \end{pmatrix},$$

resulting in coefficient of linear correlation $\rho = 1$ between ratings of the teams 1 and 2. On the other hand, if the observed point differential is completely uninformative, the covariance matrix remains unchanged by the observation

$$\lim_{\sigma_v^2 \to \infty} \hat{\mathbf{\Sigma}}_{\mathbf{t|t}} = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}.$$

2.3.7. *Geometric Interpretation for State-Space Model.* The presented state-space model for a basketball league has a clear geometric interpretation. The current state of the system, i.e., $\mathbf{x_t}$, is a point in $\mathbb{R}^n$ where each of the coordinates represents an individual team. When the teams' performance levels change, the point moves in $\mathbb{R}^n$ accordingly. Similarly, the current estimate of the system state $\hat{\mathbf{x}}_{\mathbf{t|t}} \in \mathbb{R}^n$. When new scores are obtained, the estimate is updated using the Kalman filter algorithm and, hopefully, the corrected estimate is closer to the true state of the system. The direction of the correction is innovated by the differences between the predictions given by the algorithm and the observed point differentials, viz., $\mathbf{d_t}$ in equation (11). The direction and magnitude of the correction is determined by the Kalman gain matrix $\mathbf{K_t}$ in equation (11) which combines the relative variances of the state estimate and observations in an optimal manner. The covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{t|t}}$ presents the uncertainty related to the location of the true system state. Therefore, the state estimate, the covariance matrix, and the assumed Gaussianity of the system can be used together to construct confidence regions for the true system state, i.e., regions where the true state lies with a given probability.

2.3.8. *Measures of Uncertainty.* The uncertainty about the state of the basketball league (Yilmaz and Chatterjee 2000) can be measured using various measures. The simplest approach is to study the difference between two consequent covariance matrixes representing the accuracy of the state estimate before and after an observations, i.e., $\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}} - \hat{\mathbf{\Sigma}}_{\mathbf{t|t}}$. For the uncertainty related to a multivariate normal distribution, there exists also more specific measures such as the total variation and generalized variance (see, e.g., Mardia et al. 1979).

The total variation is used as measure of uncertainty in principal component analysis and is defined as the sum of individual variances

$$tr\left(\hat{\mathbf{\Sigma}}_{\mathbf{t|t}}\right) = \lambda_1 + \lambda_2 + \ldots + \lambda_n \tag{17}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the covariance matrix.

The generalized variance is defined as the determinant of the covariance matrix

$$\left|\hat{\mathbf{\Sigma}}_{\mathbf{t|t}}\right| = \lambda_1 \cdot \lambda_2 \cdots \lambda_n \tag{18}$$

and it has an information theoretic background because the entropy of the multivariate normal distribution is

$$\frac{1}{2}\left(n \log(2\pi) + \log\left|\hat{\mathbf{\Sigma}}_{\mathbf{t|t}}\right| + n\right).$$

In the example discussed in Subsection 2.3.6, the effect of observations have on the uncertainty related to the state estimate is studied by comparing the covariance matrixes before and after the observation:

$$\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}} - \hat{\mathbf{\Sigma}}_{\mathbf{t|t}} = \frac{1}{2+\sigma_v^2/\sigma^2}\begin{pmatrix} \sigma^2 & -\sigma^2 & 0 \\ -\sigma^2 & \sigma^2 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{19}$$

This matrix is positive semidefinite which means that the observation does not increase the amount of uncertainty. Stronger results are obtained when measures specific to the multinormal distribution are used.

The total variation before the observation is

$$tr\left(\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\right) = 3\sigma^2.$$

For the updated covariance matrix, the total variation is

$$tr\left(\hat{\mathbf{\Sigma}}_{\mathbf{t|t}}\right) = \frac{4\sigma^4 + 3\sigma^2\sigma_v^2}{2\sigma^2+\sigma_v^2} = 3\sigma^2 - \frac{2\sigma^4}{2\sigma^2+\sigma_v^2} < 3\sigma^2 = tr\left(\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\right).$$

Thus, the made observation decreases the uncertainty about the state of the system when it is measured using total variation.[7]

Similarly in Subsection 2.3.6, the generalized variance before the observation is

$$\left|\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\right| = \sigma^6.$$

After the covariance matrix is updated, the generalized variance is

$$\left|\hat{\mathbf{\Sigma}}_{\mathbf{t|t}}\right| = \frac{(\sigma^4+\sigma_v^2\sigma^2)^2(2\sigma^4+\sigma_v^2\sigma^2)}{(2\sigma^2+\sigma_v^2)^3} < \frac{(2\sigma^4+\sigma_v^2\sigma^2)^3}{(2\sigma^2+\sigma_v^2)^3} = \sigma^6 = \left|\hat{\mathbf{\Sigma}}_{\mathbf{t|t-1}}\right|$$

and also the generalized variance is reduced.[7]

It should be noted that both the total variation and the generalized variance suffer from severe limitations (Mustonen 1997). Most notably, the total variation ignores the correlation between variables and the generalized variance collapses to zero if there exists linear dependence between them. There exists also extensions for the generalized variance that might perform better (Mustonen 1997), but their analysis is now omitted.

2.4. **Qualitative Characteristics of Model.** In addition to the mathematical properties of the discussed state-space model and the Kalman filter algorithm, the presented model includes several qualitative features that are beneficial for its use and analysis. These positive characteristics are listed and described in the following.

[7]Unless, of course, the observation is completely uninformative and $\sigma_v^2 \to \infty$.

Ranking and rating: The state estimates given by the Kalman filter algorithm can be used to order the teams based on their level of performance.

Simplicity: The state-space model for a basketball league is intuitive and comprehensible. The system dynamics as well as the modeling of game scores involve no complex assumptions or *ad hoc* modeling choices. Furthermore, the model includes only four independent parameters with clear interpretation which aids in their analysis as well as in their calibration if it is based on expert knowledge.

Interpretability: The system state has a clear interpretation, i.e., the difference $\hat{\mathbf{x}}_{\mathbf{t|t}}(i) - \hat{\mathbf{x}}_{\mathbf{t|t}}(j)$ represents the difference in the two teams' performance as measured in points. In addition, sum $\hat{\mathbf{x}}_{\mathbf{t|t}}(i) - \hat{\mathbf{x}}_{\mathbf{t|t}}(j) + \mu$ gives the expected point differential for the game between teams i and j at i's home court.

Transparency: The mathematics involved in the Kalman filter algorithm are simple and the effects of the model parameters are transparent as presented in Subsection 2.3.4.

Limited Input: After the parameters of the model have been estimated or calibrated, the model uses only scoreboard information that is readily available immediately after the end of the game.

Measurable Uncertainty: Covariance matrix $\hat{\mathbf{\Sigma}}_{\mathbf{t|t}}$ gives topical and probabilistic information about the uncertainty of system state. The covariance matrix and the normality assumption can be used to calculate confidence regions for the system state.

Time Evolution: The changes in the estimated system state tell the story behind the ongoing season and the observed scores. This is further discussed in Subsection 2.6.1.

2.5. Estimation and Calibration of Model.

2.5.1. *Parameters of the Model:* $\boldsymbol{\theta} := (\sigma_0^2, \sigma_w^2, \sigma_v^2, \mu)$. The state-space model and Kalman filter algorithm discussed in Section 2.2 include four parameters that determine the behavior of both the system representing a basketball league and the state estimates used for rating the teams. These parameters are σ_0^2, σ_w^2, σ_v^2, and μ in equations (5), (6), and (8). In the following, the parameters are collected into a parameter vector $\boldsymbol{\theta} := (\sigma_0^2, \sigma_w^2, \sigma_v^2, \mu)$ for convenience. The three variance parameters have been discussed earlier and they are as follows. Variance σ_0^2 reflects the uncertainty about the initial state of the system, i.e., the team strengths at the beginning of the season. As discussed in Subsection 2.2.2, the performance of teams is affected by many factors that include, e.g., evolution and devolution of individual players, injuries and player transactions between teams. Variance σ_v^2 is a measure of this variability in team strength during the season, i.e., the volatility in team performance from game to game. There are many random factors affecting the scores of individual games. The intensity of these aleatory elements in represented by variance σ_w^2, i.e., the amount of noise in the observed point differentials.

The observation equation (6) includes the parameter μ representing the advantage the teams have for playing in their home courts. The home court advantage has been studied in the literature (see, e.g., Courneya and Carron 1992, Jones 2007, Entine and Small 2008, Moskowitz and Wertheim 2011) and it has a significant effect on the game scores. Therefore, home court advantage is also included in many models discussed in Section 1.2. Courneya and Carron (Courneya and Carron 1992) propose four factors that could account for the home court advantage: crowd factors, familiarity with local conditions, travel factors, and effects related to rule differences for the home versus visiting team. Entine (Entine and Small 2008) notes that rest is also a significant factor in the home court advantage as the visiting team has often played a game also on the previous night because the NBA teams usually play their away-games on road trips consisting of several successive games. Jones (Jones 2007) presents the home advantage as a game-long process. Jones (Jones 2008), on the other hand, discusses the team dependence of the home court advantage.

In the following, maximum likelihood estimation (ML, see, e.g., Davidson and MacKinnon 1993, Casella and Berger 2001) and expert calibration of the parameters are presented. The notable difference between ML estimation and expert calibration is that the calibration can be done ex ante – even before any game scores have been observed. The ML estimation requires available data and, therefore, it is done ex post, i.e., after a suitable number of games has been played. Naturally, the parameters can also be estimated based on the previous seasons' scores. In theory, the ML estimation should produce most accurate results. On the other hand, it also involves an assumption about the normality of system state and observed point differentials that may not hold in practice.

2.5.2. *Maximum Likelihood Estimation:* $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$. The parameters presented in Subsections 2.2.2 and 2.2.3 can be estimated using the ML method (see, e.g., Davidson and MacKinnon 1993, Casella and Berger 2001) provided that the necessary data are available. In practice, the parameters can be estimated if the scores of one season's games are known up to the time instant T. Let us denote the observed results up to time instant $t-1 \leq T$ by $\mathcal{Y}_{t-1} := (\mathbf{y_1}', \mathbf{y_2}', \ldots, \mathbf{y_{t-1}}')'$. If the initial state and the innovations $(\mathbf{v_1}', \mathbf{v_2}', \ldots, \mathbf{v_t}', \mathbf{w_1}', \mathbf{w_2}', \ldots, \mathbf{w_t}')'$ are multivariate Gaussian (as was assumed in Section 2.2), it follows that

$$\mathbf{y_t} | \, \mathcal{Y}_{t-1} \sim N(\hat{\mathbf{y}}_\mathbf{t}, \hat{\mathbf{O}}_\mathbf{t}), \tag{20}$$

where $\hat{\mathbf{y}}_\mathbf{t}$ and $\hat{\mathbf{O}}_\mathbf{t}$ are as defined in equation (13). Thus, the conditional probability distribution function of $\mathbf{y_t} | \, \mathcal{Y}_{t-1}$ is

$$\begin{aligned} f_{\mathbf{y_t}| \, \mathcal{Y}_{t-1}}(\mathbf{y_t} | \, \mathcal{Y}_{t-1}) &= (2\pi)^{-m/2} |\hat{\mathbf{O}}_\mathbf{t}|^{-1/2} \\ &\quad \times \exp\left(-\frac{1}{2} (\mathbf{y_t} - \hat{\mathbf{y}}_\mathbf{t})' \hat{\mathbf{O}}_\mathbf{t}^{-1} (\mathbf{y_t} - \hat{\mathbf{y}}_\mathbf{t}) \right), \end{aligned} \tag{21}$$

where $|\hat{\mathbf{O}}_\mathbf{t}|$ denotes the determinant of the covariance matrix $\hat{\mathbf{O}}_\mathbf{t}$. The probability of making a sample of observations $\mathcal{Y}_T = (\mathbf{y_1}', \mathbf{y_2}', \ldots, \mathbf{y_T}')'$ is then proportional to the likelihood function

$$L(\boldsymbol{\theta}; \mathcal{Y}_T) = \prod_{t=1}^{T} f_{\mathbf{y_t}| \, \mathcal{Y}_{t-1}}(\mathbf{y_t} | \, \mathcal{Y}_{t-1}).$$

For computational purposes, the analysis is often based on the logarithm of the likelihood function, viz., the log-likelihood function

$$\ell(\boldsymbol{\theta}; \mathcal{Y}_T) = \sum_{t=1}^{T} \log f_{\mathbf{y_t}| \, \mathcal{Y}_{t-1}}(\mathbf{y_t} | \, \mathcal{Y}_{t-1}).$$

In order to find the maximum likelihood estimator for the model parameters, the log-likelihood function is maximized with respect to the parameter values (Harvey 1989, pp. 126) yielding

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \ \ell(\boldsymbol{\theta}; \mathcal{Y}_T). \tag{22}$$

where the parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^4$ is defined so that the variance parameters are non-negative (e.g., Davidson and MacKinnon 1993, pp. 248).

The ML estimator can also be defined as the solution of the likelihood equations or the following first-order conditions for the maximum of the likelihood function:

$$\mathbf{g}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}) = \mathbf{0}. \tag{23}$$

Here, the vector $\mathbf{g} \in \mathbb{R}^4$ is the gradient of the log-likelihood function with typical element

$$\mathbf{g}_i(\boldsymbol{\theta}) := \frac{\partial \ell(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}(i)}$$

where $\partial\boldsymbol{\theta}(i)$ refers to the individual model parameters. In practice, equation (23) is solved numerically using, e.g., the Newton's method or quasi-Newton methods (e.g., Davidson and MacKinnon 2004, pp. 228). In these methods, the log-likelihood function is evaluated by performing the Kalman filter algorithm presented in Subsection 2.2.4 repeatedly for given parameter values. This is straightforward as the Kalman filter algorithm readily produces all the state estimates and covariance matrixes needed in the calculations, i.e., $\hat{\mathbf{x}}_{\mathbf{t}|\mathbf{t-1}}$ and $\hat{\mathbf{O}}_{\mathbf{t}}$ in equation (21). The Newton method makes use of the Hessian matrix, which is now a 4×4 matrix $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta})$ consisting of the second derivatives of the log-likelihood function

$$\boldsymbol{\mathcal{H}}_{ij}(\boldsymbol{\theta}) := \frac{\partial^2 \ell(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}(i) \partial \boldsymbol{\theta}(j)}.$$

According to the information matrix equality (e.g., Davidson and MacKinnon 1993, pp. 263), the asymptotic covariance matrix of the ML estimates is given by

$$\mathbf{V}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}) = -\boldsymbol{\mathcal{H}}^{-1}(\boldsymbol{\theta}_{\mathbf{0}}),$$

where the Hessian matrix is evaluated in the numerical solution of the likelihood equations $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$. This matrix is also known as the observed information matrix. The ML estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ is a consistent estimator for the true values of the parameters (e.g., Davidson and MacKinnon 1993, pp. 145). When the number of observations increases, the asymptotic distribution of the ML estimator is

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}} - \boldsymbol{\theta}_{\mathbf{0}}\right) \underset{a}{\sim} N(\mathbf{0}, -\boldsymbol{\mathcal{H}}^{-1}(\boldsymbol{\theta}_{\mathbf{0}})), \tag{24}$$

given that certain regularity conditions hold. These conditions include that the true value $\boldsymbol{\theta}_{\mathbf{0}}$ is an interior point of the parameter space,[8] the derivatives of $\ell(\boldsymbol{\theta}; \mathcal{Y}_T)$ with respect to $\boldsymbol{\theta}$ exist and are continuous in the neighborhood of $\boldsymbol{\theta}_{\mathbf{0}}$ (up to order three), and the model is identifiable (Harvey 1989, pp. 128).

[8]Note that this condition does not hold for one of the examples in Section 3.1.

2.5.3. *Calibration of Using Other Techniques:* $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$. If there is no available data for the ML estimation of the parameter vector $\boldsymbol{\theta}$, the parameter values can also be calibrated based on expert knowledge, historical data involving the league in question or even some similar basketball league. In practice, a subject matter expert is asked for ballpark estimates for the parameter values. These values are then used for following the ongoing basketball season and more detailed estimation is carried out when the necessary data have become available.

Naturally, the expert assessment of the parameter values can depend on the data from previous seasons or similar leagues. For example, the home court advantage can be estimated from the previous season by calculating the average difference between the points scored by the home team and the visiting team. In the NBA 2004-05 and 2005-06 seasons, the home court advantage has been estimated as 3.24 points (Entine and Small 2008). Similar calculations can be performed for more recent NBA seasons resulting in a rough estimate $\hat{\mu} \approx 3$.

The calibration of the variance parameters, on the other hand, is not as straightforward. If the random variables in question are assumed Gaussian, rough estimates for the variance parameters can be obtained using a line of reasoning that is based on confidence regions, e.g., 95% confidence intervals. That is, it may be easier to evaluate intervals for likely variation of the modeled system, instead of the direct values of the variance parameters.

Variance parameter σ_0^2 represents the uncertainty related to the initial state estimates of the system. One might argue, e.g., the expected value of an individual team's state should lie with a 95% probability within ± 20 points from the league average. Given the assumption about normality, this would imply that in the expert's opinion $\hat{\sigma}_0^2 \approx 10^2 = 100$.

An interesting way for estimating variance parameter σ_v^2, i.e., the amount of random noise in the point differential of an individual game, can be constructed using the betting odds and point spreads offered by bookmakers.[9] The betting odds for the winner of a game can be used to calculate the implied probability of a home win. On the other, the point spread offered by the bookmaker represents the median of the probability distribution. Under the normality assumption, this equals the

[9]In actuality, the bookmakers continuously adjust the odds and point spreads according to the bets placed by the bettors in order to maintain a profitable position regardless of the realized result of the game. Thus, the betting odds and point spreads offered by the bookmakers in fact represent the entire betting market's view on the relative strength of the teams. See, e.g., Bassett (1981), Gandar et al. (1998).

mean of the probability distribution. Together, the home team's winning probability and the location of the mean are used to calculate the variance (or equivalently the standard deviation) of the underlying probability distribution. This calculation can be performed to almost any betting odds and respective point spreads resulting in the implied estimate $\hat{\sigma}_v^2 \approx 13.5^2 = 182.25$.

The variance of the team strengths can be estimated, e.g., by studying a typical difference in the level of performance for a team at the beginning of the season and the same team at the end of the season. One can try to imagine a virtual game between these teams and project a 95% confidence interval for the game outcome. This interval, together with the normality assumption, is then used to calculate an estimate for the variance of the team performance during a season. Furthermore, as the changes in team strength are assumed independent during non-overlapping time intervals, an approximation for the variance between games is calculated by dividing the variance over the entire season by the number of games played. This kind of reasoning leads to an estimate $\hat{\sigma}_w^2 \approx 0.25$.

To summarize, the estimates for the model parameters resulting from this analysis are $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}} = (100, 0.25, 182.25, 3)$. These estimates are compared with the ML estimates in Section 3.1.

2.6. **Utilization of Model.**

2.6.1. *Following the Current Season.* State estimates $\hat{\mathbf{x}}_{\mathbf{t|t}}$ represent the progress of the ongoing season as a n-dimensional time series. The perceived performance levels of the teams can be presented graphically which gives a good overall picture of the evolution of the entire league. Naturally, the analysis can also focused on individual teams, e.g., $\hat{\mathbf{x}}_{\mathbf{t|t}}(i)$, or comparison of two teams, e.g., $\hat{\mathbf{x}}_{\mathbf{t|t}}(i)$ and $\hat{\mathbf{x}}_{\mathbf{t|t}}(j)$. Examples of time series representing NBA Season 2009-2010 are given in Section 3.3.

2.6.2. *Ranking of Teams.* The presented state-space model and the Kalman filter algorithm are used to rank and rate the teams according to their perceived level of performance. If an ordinal listing of teams is needed, the teams are simply sorted in a descending order based on their ratings. Furthermore, if more detailed presentation of the teams' strength is desired, the teams are first sorted based on their ratings and the ranking is augmented by presenting the exact ratings of the teams.

In effect, the team ratings are used for statements such as "team i is better than team j because it has a higher rating". The ranking of two basketball teams can be presented as a hypoThesis test where the null and alternative hypotheses are as follows

$$H_0: \ \mathbf{x_t}(i) - \mathbf{x_t}(j) \leq 0 \text{ and } H_1: \ \mathbf{x_t}(i) - \mathbf{x_t}(j) > 0.$$

The P-value for the test is solved using the conditional distribution

$$\mathbf{x_t}(i) - \mathbf{x_t}(j) | \ \mathcal{Y}_t \sim N\left(\hat{\mathbf{x}}_{\mathbf{t|t}}(i) - \hat{\mathbf{x}}_{\mathbf{t|t}}(j), \underbrace{\hat{\mathbf{\Sigma}}_{\mathbf{t|t}}(i,i) + \hat{\mathbf{\Sigma}}_{\mathbf{t|t}}(j,j) - 2\hat{\mathbf{\Sigma}}_{\mathbf{t|t}}(i,j)}_{=:\sigma^2}\right)$$

which gives

$$P(\mathbf{x_t}(i) - \mathbf{x_t}(j) \leq 0 | \ \mathcal{Y}_t) = \Phi\left(\frac{-\hat{\mathbf{x}}_{\mathbf{t|t}}(i) + \hat{\mathbf{x}}_{\mathbf{t|t}}(j)}{\sigma}\right), \tag{25}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The P-value is used to analyze the differences between the teams' level of play and study the credibility of the presented ordering of teams. The teams could also be classified into statistically similar groups based on their performance. This classification is considered beyond the scope of this Thesis and omitted.

2.6.3. *Prediction of Game Outcomes and Odds.* The Kalman filter algorithm concentrates all available information about observed scores into the predicted state of the system and its covariance matrix in a form that can be used to predict future scores. That is

$$\mathbf{y_{t+1}} | \ \mathcal{Y}_t \sim N(\hat{\mathbf{y}}_{\mathbf{t+1}}, \hat{\mathbf{O}}_{\mathbf{t+1}}),$$

The state estimates given by the Kalman filter algorithm and equation (13) can be used to calculate odds for future game, i.e., the teams' winning probabilities for each game played at time instant $t+1$. In particular,

$$\mathbf{y_{t+1}}(k) | \ \mathcal{Y}_t \sim N(\hat{\mathbf{x}}_{\mathbf{t+1|t}}(i) - \hat{\mathbf{x}}_{\mathbf{t+1|t}}(j) + \mu, \sigma^2)$$

where team i is the home team and team j is the visiting team on the kth game played on day $t+1$, and $\sigma^2 := \hat{\mathbf{O}}_{\mathbf{t+1}}(i,j)$. This information can be used to predict the probability of the home team winning the game k on $t+1$th day:

$$P(\mathbf{y_{t+1}}(k) > 0 | \ \mathcal{Y}_t) = 1 - \Phi\left(\frac{-\hat{\mathbf{x}}_{\mathbf{t+1|t}}(i) + \hat{\mathbf{x}}_{\mathbf{t+1|t}}(j) - \mu}{\sigma}\right), \tag{26}$$

where $\Phi(\cdot)$ is again the cumulative distribution function of the standard normal distribution. This information can be used in two ways. First, the team with a higher probability of winning is seen as the predicted favorite to win the game under consideration. Second, the probability given by equation (26) is a measure of confidence on this prediction with a clear interpretation, i.e., the favorite's predicted winning probability ranges from 0.50 to 1.00.

2.7. **Validation of Team Ratings.** The Kalman filter algorithm and its predictions are compared with observed results in order to ascertain that the produced team ratings give a valid representation for the basketball league under consideration. Techniques applicable for the validation analysis include:

Graphical Analysis: The state estimates given by the Kalman filter algorithm depend on time and they are easily presented as time series representing the evolution of the teams' ratings. These time series give a narrative of the season and a subject matter expert can compare them with teams' actual performance in order to assess the credibility of the portrayed ratings. This kind of assessment is referred to as "passing the laugh test" in Oliver (2004).

Accuracy of Predictions Given by Algorithm: The Kalman filter algorithm predicts the favorite for each match and the accuracy of these predictions reflects the accuracy of the team ratings. The prediction accuracy is measured by comparing the predictions with the realized game outcomes. Then, the percentage of games predicted correctly can be calculated and compared with alternative models. Naturally, inaccurate predictions imply an inaccurate rating.

Accuracy of Odds Given by Algorithm: The Kalman filter algorithm gives an estimate for the teams' winning probabilities for each match, e.g., the probability of the home team winning. The accuracy of the odds given by the algorithm are compared with the realized game results using a logistic regression model with normal errors. If the predicted odds are accurate there should be an almost linear dependence between the predicted winning probability and the realized match results.

Sensitivity Analysis: The parameters of the state-space model affect the team ratings. In sensitivity analysis, the values of the parameters are varied in order to see the dependence between the team ratings and the parameter values. If the performance or the predictions of the algorithm are heavily sensitive to parameters, this source of uncertainty should be emphasized in the presentation of results.

The validation of the team ratings is discussed in further detail in Sections 3.2 and 3.3.

3. Example Analyses

In the following, the utilization of the Kalman filter algorithm illustrated by analyzing NBA data. First, in Section 3.1 the parameters of the model are for the NBA seasons 2005-2006 through 2009-2010 using ML estimation technique discussed in Subsection 2.5.2. The estimates are also compared to those obtained in Subsection 2.5.3. Second, in Section 3.2 the performance of the Kalman filter algorithm is validated by studying the predictions given by the algorithm for the NBA seasons 2005-2006 through 2009-2010. Third, in Section 3.3 the Kalman filter algorithm is used to study NBA season 2009-2010 in full detail. Initially, the assumptions of the state-space model are compared with actual game scores. Then, the Kalman filter algorithm is used to estimate team ratings for the entire season using both the ML estimates for the parameter vector and the parameter values suggested by an expert analyst. In Section 3.5, the final rankings and ratings for the NBA season 2009-2010 are compared. Finally, in Section 3.4 the scores of individual games of the NBA 2010 Finals are predicted and discussed.

3.1. **Model Parameters Estimated for the NBA Seasons 2005-2010.** The parameters of the state-space model are estimated using ML estimation presented in Section 2.5.2 for NBA regular seasons 2005-2006 through 2009-2010 where each of the seasons consists of 1230 games. The ML estimates are estimated independently for each of the seasons and the estimates are presented in Table 1.

$\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$	'05-'06	'06-'07	'07-'08	'08-'09	'09-'10	$\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$
$\hat{\sigma}_0^2$	11.48	6.88	25.53	19.48	16.76	100
$\hat{\sigma}_w^2$	0.0174	0.0917	0.0138	0.0000	0.0533	0.25
$\hat{\sigma}_v^2$	127.08	133.99	132.31	131.45	133.58	182.25
$\hat{\mu}$	3.37	3.01	3.26	3.28	2.72	3
no. of games	1230	1230	1230	1230	1230	-

Table 1. Maximum likelihood estimates for the model parameters for NBA regular seasons 2005-2006 through 2009-2010. The estimates based on expert knowledge are also presented for comparison.

The estimate $\hat{\sigma}_0^2$ varies heavily from season to season, ranging from 6.88 to 25.53. This may be the result of differences in teams' level of play or the variations in the teams's performance at the beginning of the seasons. The estimate $\hat{\sigma}_w^2$ ranges between 0.0000 and 0.0917. Interestingly the largest estimate coincides with the smallest value of $\hat{\sigma}_0^2$. That is in season 2006-2007, the initial uncertainty is quite small, but the changes in team ratings are the greatest. In season 2008-2009, the ML estimate is equal to zero, i.e., the probability of the data is highest if the model does not include any variation in team strengths.[10] Compared to the other variance estimates, the residual variation in point differentials is more consistent from season to season which seen in the estimates $\hat{\sigma}_v^2$ that range from 127.08 to 133.99. It would appear that the models explain some of the variation in observed scores as the estimates are smaller than the estimate given by the expert. This explanatory power is discussed also in Section 3.2. Overall, the variance estimates given by the expert knowledge are much larger than the ML estimates – reflecting the higher uncertainty. This is natural as the expert estimates are given a priori, i.e., before the game results have been estimated.

The estimate $\hat{\mu}$ varies slightly during the interval 2005-2010, but changes are moderate with range from 2.72 to 3.37. Notably, these estimates match to the average point differentials observed in the corresponding seasons. The ML estimates of home court advantage are congruent with the estimate given by the expert that falls in the middle of the range of the ML estimates.

	'05-'06	'06-'07	'07-'08	'08-'09	'09-'10
Kalman rating (ML)	**0.6577***	**0.6539***	**0.6894***	**0.6995***	0.6862*
Kalman rating (ML, prev.)	-	0.6426*	0.6829*	**0.6995***	0.6830*
Kalman rating (EK)	0.6561*	0.6491*	0.6854*	0.6922*	**0.6911***
home team	0.6033	0.5922	0.5951	0.6099	0.5941
wins	0.5724	0.5597	0.6211	0.6262	0.6031
winning percentage	0.5984	0.5865	0.6393	0.6393	0.6194
point differential	0.6341	0.6288	0.6772*	0.6938*	0.6626
point differential (60 days)	0.6472*	0.6426*	0.6764*	0.6922*	0.6650

TABLE 2. Probability of correctly predicting the game outcomes for the NBA regular seasons 2005-2006 through 2009-2010 using different predictors.

3.2. **Performance of Kalman Filter Algorithm during the NBA Seasons 2005-2010.** The performance of the Kalman filter algorithm is tested by using it to predict the outcomes of the games in the NBA during the seasons 2005-2006 through 2009-2010. The probabilities of correctly predicting the game outcomes

[10]As stated in Subsection 2.5.2, the asymptotic distribution of the estimator of the parameters is not normal in this case. This is an occurrence that may warrant further analysis.

using various performance measures are given in Table 2. Table 2 includes the predictions given by the Kalman filter with three alternative parameter estimates: ML estimates based on the season's results (ML), ML estimates based on the previous season's results (ML, prev.), and estimates given by the expert (EK). For comparison, the predictions given by five other performance measures are also presented. In these predictions, the favorite is either the home team, the team with more wins, the team with higher winning percentage, the team with higher point differential, or the team with the higher point differential during last 60 days. For each season, the highest probability is bolded denoting the best predictions. The predictive power of different methods is compared with the predictions given by the Kalman filter with ML based parameter estimates using the McNemar's test for marginal homogeneity with dependent observations (McNemar 1947). The probabilities marked with an asterisk (*) do not differ statistically significantly from those given by the Kalman filter on the 0.05 significance level.

Table 2 shows that, among the discussed predictors, the Kalman filter algorithm and the scoring differentials give the most accurate predictions. The predictions given by the Kalman filter are not sensitive to the values of the parameter estimates as there is no statistically significant differences between three parameters values. Notably, the predictions given by the two scoring differentials are almost as accurate as those given by the Kalman filter and the difference is not statistically significant. This is further discussed in Section 3.5.

	'05-'06	'06-'07	'07-'08	'08-'09	'09-'10
Kalman rating (ML)	0.6916*	**0.6623***	0.7256*	**0.7138***	0.7071
Kalman rating (ML, previous)	-	0.6510*	0.7208*	0.7106*	0.6909
Kalman rating (EK)	**0.6916***	0.6575*	**0.7289***	0.7122*	0.7039*
home team	0.6266	0.5828	0.6201	0.6276	0.5615
wins	0.6218	0.5958	0.6769	0.6488	0.6472
winning percentage	0.6396	0.6153	0.6883	0.6569	0.6505
scoring differential	0.6429	0.6299	0.6997	0.6699	0.6748
scoring differential (60 days)	0.6429	0.6136	0.6851	0.6650	0.6699

TABLE 3. Probability of correctly predicting the game outcomes for the second halves of NBA regular seasons 2005-2006 through 2009-2010 using different predictors.

Table 3 presents data similar to Table 2 but now the predictor performances are studied only for the second half of each of the seasons. In this table, the predictors have had time to ”warm up”, i.e., the first half of the season is only used for producing the estimates and their performance during the this period is ignored in the comparison. The best predictors are now those given by the Kalman filter algorithm. Interestingly for some of the seasons, the best predictions are given by the parameter estimates based on expert knowledge or the scoring differential. None of the differences among Kalman filters are statistically significant.

The predictions given by the Kalman filter algorithm are correct $65-70\%$ of the time when only the complete seasons are considered. This quite a good result and the algorithm outperforms the simplest predictors in Tables 2 and 3. On the other hand, the results do not differ statistically significantly from those given by the scoring differential during last 60 days. For the second half of the season, the Kalman filter algorithm predicts correctly $66-73\%$ of the games and outperforms the other predictors. There is no notable differences between the alternative values of the parameters.

Unfortunately, the comparison of the predictive accuracy of the Kalman filter algorithm with the existing literature is not straightforward. Out of the articles discussed in Section 1.2, only Knorr-Held discusses the ranking of NBA teams (Knorr-Held 2000) and even in this article the accuracy of the model predictions is not reported. In most of the articles mentioned in Section 1.2, statistical models are used to predict the results of NCAA Tournament games. Boulier and Stekler (Boulier and Stekler 1999) report that the seedings set by the NCAA tournament selection committee predict correctly 70% of the NCAA tournament games played during seasons $1986-1995$. The models presented in (Harville 2003) predict correctly $69-71\%$ of the NCAA 2001 Tournament games while the betting market is correct 66.5% of the time. Coleman and Lynch (Coleman and Lynch 2009) give a comprehensive presentation of the NCAA Tournaments for seasons $1999-2008$. According to this article, the accuracies of different predictors during this time are as follows: RPI 69.6% (NCAA 2011), NCAA seedings 70.0% (NCAA 2011), Sagarin rating 71.4% (Sagarin 1985), Kvam-Sokol 71.4% (Kvam and Sokol 2006), and betting market 74.6%. In the same time interval, the regression models presented by Coleman and Lynch are correct in $73.2-73.8\%$ of the tournament games (Coleman and Lynch 2009).

However, one should note that the prediction of NCAA Tournament games is not entirely comparable with the prediction of NBA regular season games as one could argue that the differences between team strengths are higher in college basketball

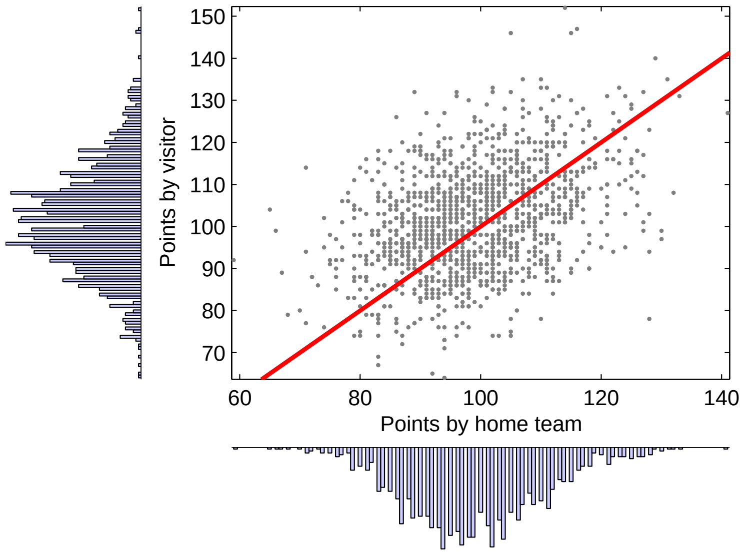

FIGURE 1. Scatterplot of the scores of the games as well as the histograms representing the marginal distributions of the points scored by the home and visiting team for the NBA season 2009-2010.

than in the NBA. Additionally, the teams participating in the final tournament are guaranteed to be playing at full force which does not necessarily hold for all the NBA regular season games. For example, NBA teams that have secured their playoff position have been known to rest their key players for the playoff run in the last games of the regular season. Therefore, one could expect a slightly lower accuracy when predicting the outcomes of NBA regular season games compared to the NCAA Tournament. Overall, it is safe to conclude that the accuracy of the predictions given by the Kalman filter algorithm is good and in line with the existing literature.

3.3. **NBA Season 2009-2010.**

3.3.1. *Observed Scores.* Scatterplot of the scores of the games played in the NBA season 2009-2010 including the playoffs is given in Figure 1. The figure also displays the marginal distributions of the points scored by the home team and visitor that appear approximately normally distributed. On average, the home team scores slightly more points, i.e., 2.86 points per game, and wins 60.0% of the games. The

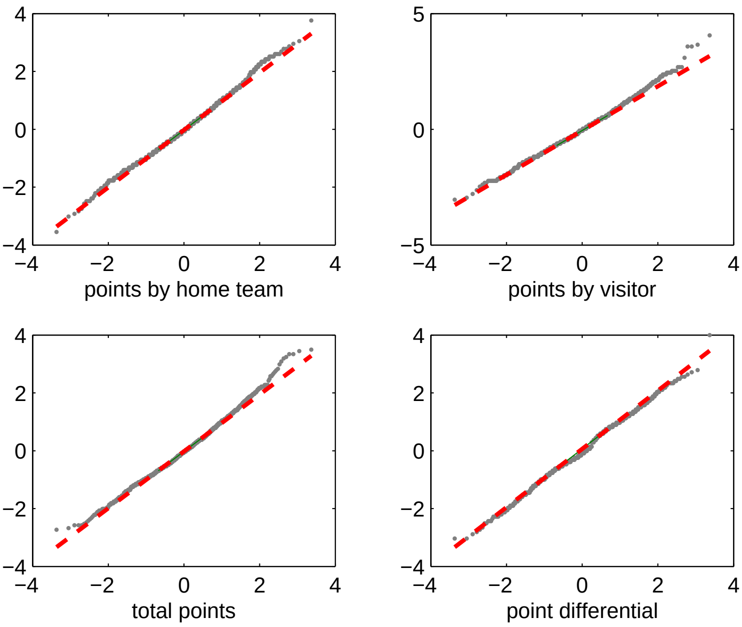

FIGURE 2. Quantile-quantile plot comparing the distributions of points scored by the home and visiting team as well as the point differential and the total points to respective normal distributions for the NBA season 2009-2010.

game cannot end in a tie and there is a gap in the joint probability distribution. This makes the actual distribution of point differentials bimodal.

variable	test			
	χ^2 $(dof = 5)$	Kolmogorov-Smirnov	Jarque-Bera	Lilliefors
home team	0.0001	0.0319	0.0070	0.0010
visiting team	0.0001	0.0320	0.0010	0.0010
total points	0.0031	0.0629	0.0010	0.0010
point differential	0.0000	0.0008	0.0010	0.5000

TABLE 4. P-values of statistical tests used in testing the normality of observed distributions of points scored by home team, points scored by visiting team, total points, and point differential. The number of observations is 1267.

The normality of the observed distributions is tested using quantile-quantile plots presented in Figure 2. The figure shows that the distributions appear approximately normal, despite slight curvature at the tails of the distributions. The normality of the different point distributions is tested using several statistical tests and rejected (Table 4). In other words, the observed distributions do not exactly follow a normal distribution. Nevertheless, the probability distribution of the point differential appears approximately normal and in accordance with the assumptions of the state-space model presented in Chapter 2. Furthermore, the derivation of the Kalman filter algorithm does not assume normality and it produces best linear estimates for the system state and observations even without the normality assumption (see, Subsection 2.3.3). Thus, it is not sensitive to slight deviations from normality.

3.3.2. *Model Parameters.* The data are used to calculate the ML estimates for the parameters of the model, i.e.,

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}} = (\hat{\sigma}_0^2, \hat{\sigma}_w^2, \hat{\sigma}_v^2, \hat{\mu}) = (17.27, 0.0390, 135.68, 2.85). \tag{27}$$

The covariance matrix of the parameter estimates is

$$\mathbf{V}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}) = -\boldsymbol{\mathcal{H}}^{-1}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}) = \begin{pmatrix} 0.0110 & 0.0010 & 0.0495 & -0.0011 \\ 0.0010 & 0.008 & -0.0514 & -0.0015 \\ 0.0495 & -0.0514 & 5.7885 & 0.1121 \\ -0.0011 & -0.0015 & 0.1121 & 0.0031 \end{pmatrix}. \tag{28}$$

parameter	lower bound	estimate	upper bound
$\hat{\sigma}_0^2$	17.07	17.28	17.48
$\hat{\sigma}_w^2$	-0.0172	0.0390	0.0951
$\hat{\sigma}_v^2$	130.96	135.68	140.39
$\hat{\mu}$	2.74	2.85	2.96

TABLE 5. Estimates and confidence intervals with confidence level 0.95 for the ML parameters in the NBA 2009-2010 season.

The covariance matrix and the normality assumption can be used to construct 95% confidence intervals for the parameters. The parameter estimates and corresponding confidence intervals are presented in Table 5. The confidence intervals and the covariance matrix imply that the parameters have been estimated reasonably well as the intervals are quite narrow relative to the absolute values of the parameters and the models's sensitivity to parameter values.[11]

[11]However, one should note that these estimates are not exactly ML estimates as the normality assumption does not hold for the observed point differentials. Therefore, it might be more accurate to refer to these estimates as pseudo-ML estimates (e.g., Gourieroux et al. 1984).

3.3.3. *Team Ratings for the NBA Season 2009-2010.* In Figure 3, the team ratings $\hat{\mathbf{x}}_{\mathbf{t+1|t}}$ calculated with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ are presented for individual teams for the duration of the NBA season 2009-2010. For clarity, the ratings for Western and Eastern Conferences are also presented separately in Figures 4 and 5.

The season, "first half" between days 1-108, "All-Star break" 109-112 no games are played, "second half" 113-170, playoffs from day 173 forward... Not all teams qualify for the playoffs which is seen in their ratings, viz., the ratings are stopped at the level of the last game played. Similarly, after teams are eliminated from playoffs, their ratings stop. Slight variations due to correlations between ratings, especially after team is eliminated from playoffs as its rating is tangled with the last opponent due to several games played between them.

In the Eastern Conference (Figure 4), the teams separated to better and worse teams gradually. The worst team's rating (New Jersey Nets, NJN) plummets at the beginning of the season. Seven teams separate themselves from the rest after All-star break and qualify for the playoffs. Note that the eight team entering the playoffs (Chicago Bulls, CHI) is ranked notably worse than the other playoff bound teams. Consequently, they are eliminated in five games by the Cleveland Cavaliers (CLE, ranked second in East by the Kalman filter). In advanced round of the playoffs, Boston Celtics (BOS) are able to upset both Orlando Magic (ORL) and CLE despite these teams' better performances in the regular season and the subsequent higher playoff seed and home court advantage. After three playoff rounds, BOS qualifies for the NBA Finals from the Eastern Conference.

Western Conference ratings (Figure 5) reveal the parity within the conference. The rating of the worst team (Minnesota Timberwolves, MIN) falls behind the pack at the beginning of the season. Other weaker teams such as Sacramento Kings (SAC) and Los Angeles Clippers (LAC) follow suit and their seasons are effectively over well before the All-star break. In the Western Conference, the teams qualifying to the playoffs are clearly separated from the other teams at the end of the regular season. Notably, the defending champion Los Angeles Lakers (LAL) entered the playoffs as the fifth ranked team in West. At the beginning of the playoffs, the best ranked teams appear to be Phoenix Suns (PHO), Utah Jazz (UTA), and San Antonio Spurs (SAS). The rating of PHO is further escalated by the 4-0 victory over SAS with very large point differentials in the second round of the playoffs. Nevertheless, LAL are able to sweep UTA 4-0 in the second round and eliminate PHO 4-2 in the Western Conference finals qualifying for the NBA Finals. The

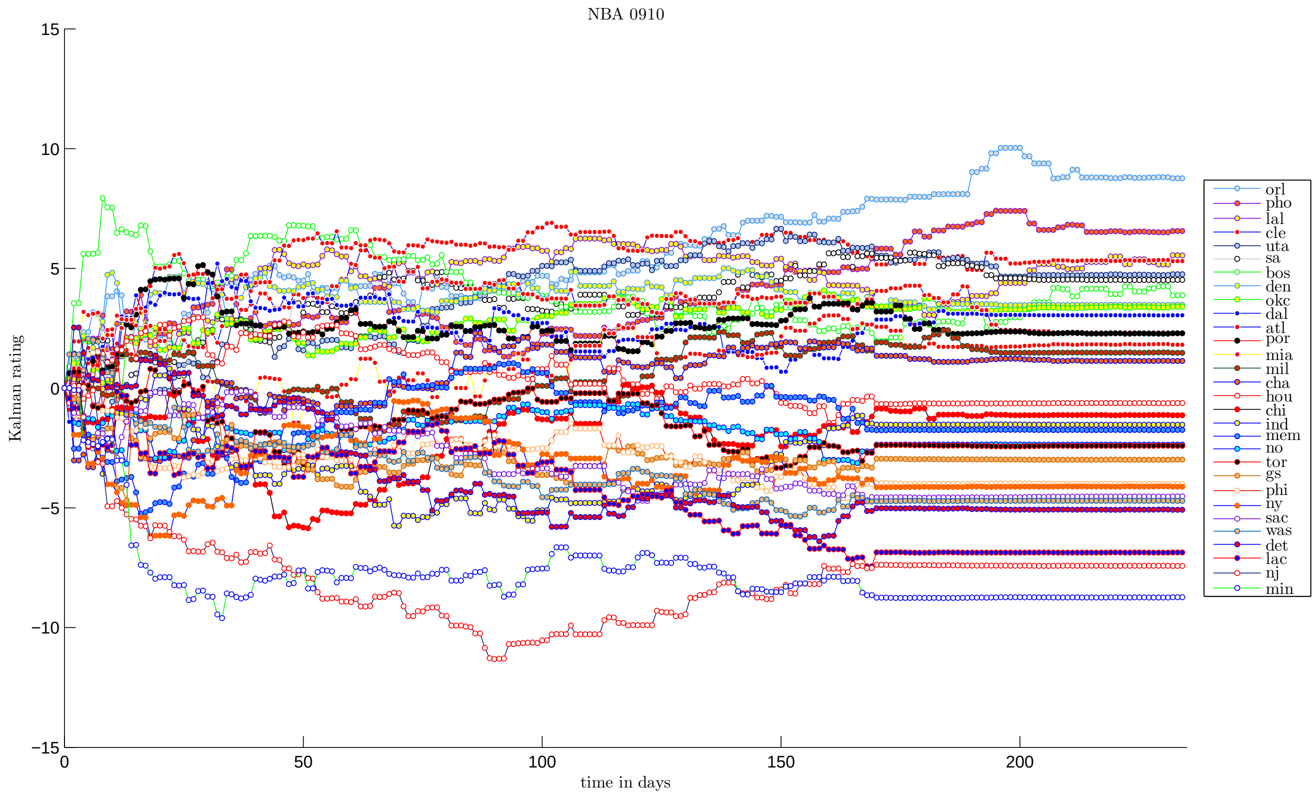

FIGURE 3. Team ratings for the NBA teams during season 2009-2010 with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$.

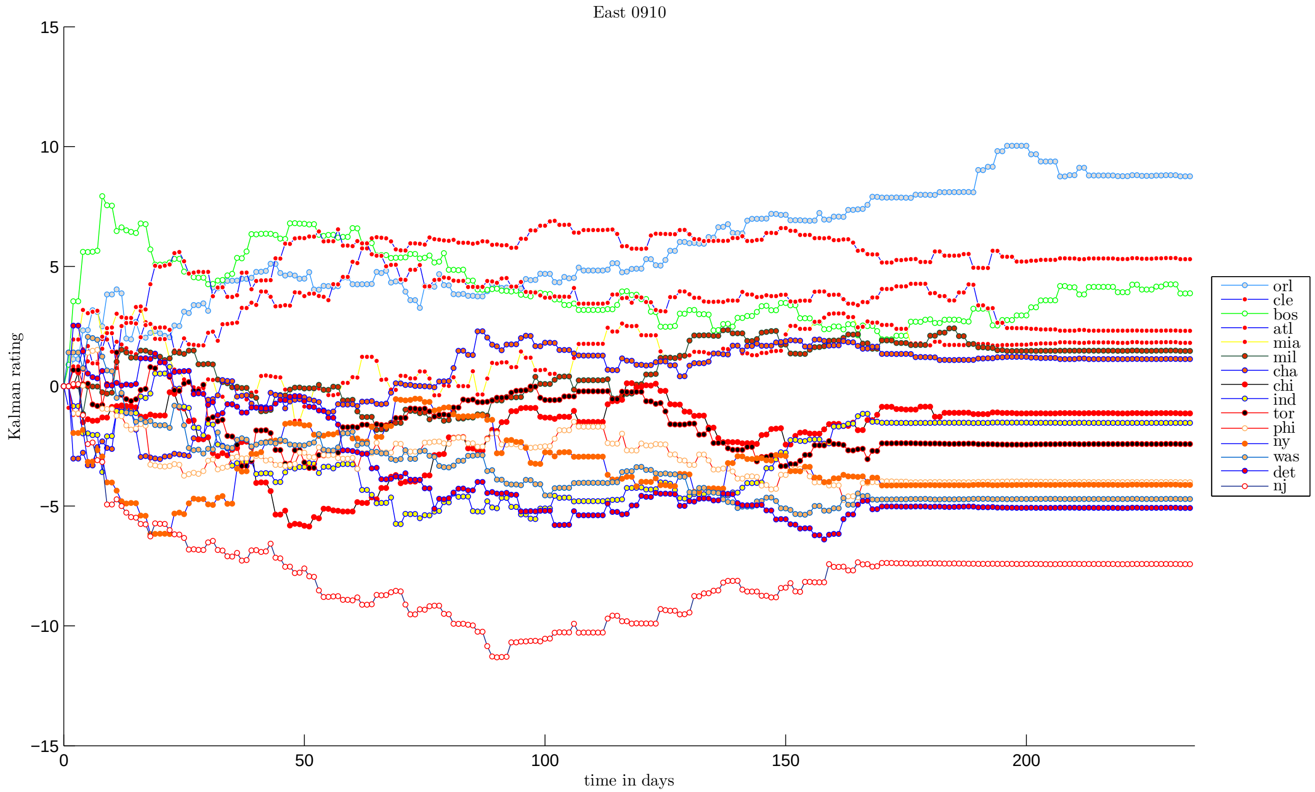

FIGURE 4. Team ratings for the Eastern Conference teams during season 2009-2010 with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$.

performances of LAL and BOS in the NBA 2010 Finals are further discussed in Section 3.4.

For the comparison of alternative parameter values and sensitivity analysis, the team ratings $\hat{\mathbf{x}}_{\mathbf{t+1|t}}$ obtained using parameter estimates $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ are presented in Figures 18, 19 and 20 in Appendix B. The notable difference between ratings given by parameter estimates $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ is in the variation of the ratings. The changes in the ratings produced by $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ are much more volatile and ranking of the teams shifts continuously throughout the regular season. This is a natural consequence of the larger variance parameters discussed in Section 3.1. These differences are discussed in further detail in Subsection 3.3.6. At the end of the season, though, the effect of the parameter values has diminished which is seen in the similarities of the teams' ratings and rankings presented in Section 3.5.

3.3.4. *Predicted Point Differentials.* The point differentials predicted by the Kalman filter algorithm with $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ and the actual observed point differentials are presented in Figure 6. From the ranges of variables, it is seen that the predictions of the Kalman filter algorithm match the expected value of observations and individual observations have a larger range due to random variation. The figure also includes red line corresponding to perfect and black dashed line representing the linear regression between predicted and observed point differentials. In the regression model, the constant term is 0.109 and the slope is 0.987. Therefore, the predicted point differentials give almost perfectly unbiased estimates for the actual point differentials. The P-value for the complete model is 0.000. The model included considerable variation, though, as the R^2 statistic is only 0.165 and therefore there exists plenty of variation unaccounted for by the model. Overall, the Kalman filter algorithm with $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ predicts correctly 69.2% of the games.

The predictions given by the Kalman filter algorithm with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ are presented in a similar scatterplot in Figure 7. The most notable difference between Figures 6 and 7 is in the regression model fitted to the predictions produced with $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$. The regression model in Figure 7 has a constant term 0.802 and slope 0.511 which is smaller than one leading to inconsistency in the expected values of the two variables. For example, if the model predicts a 10 point win for the home team, the expected point differential is in fact only 5.91 points. In other words, the model gives predictions whose expected values are exaggerated – even though the

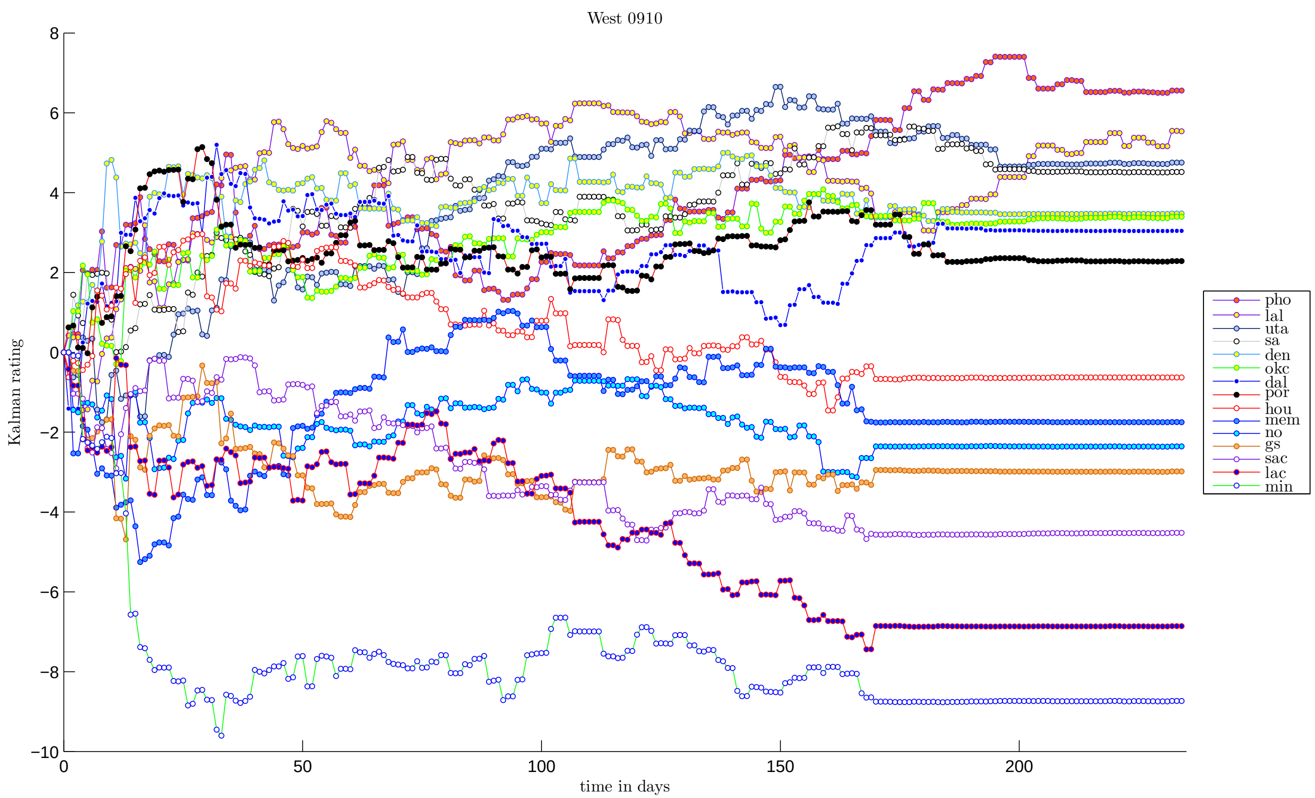

FIGURE 5. Team ratings for the Western Conference teams during season 2009-2010 with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$.

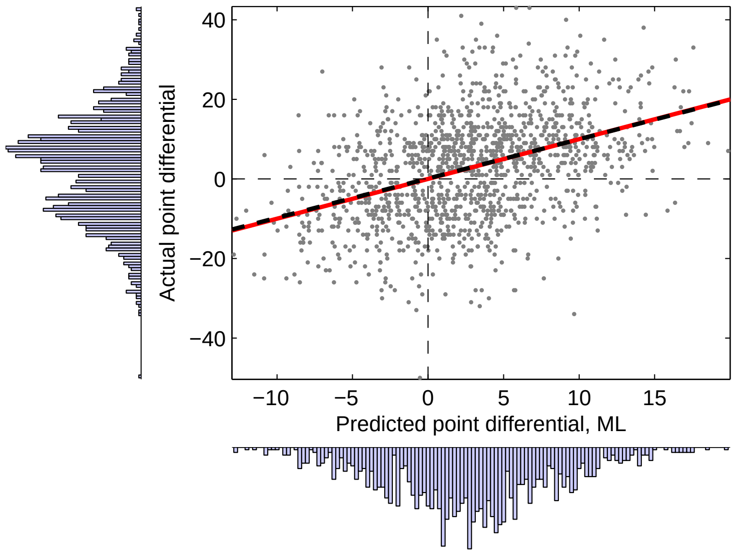

FIGURE 6. Scatterplots presenting the predicted and observed point differentials with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$.

range of predictions is much narrower than that of observed point differentials. The regressor has P-value of 0.000 and $R^2 = 0.161$.

Figure 8 shows that there is no non-linear dependence between predictions given by the Kalman filter algorithm with either of the two parameter vectors. The residuals appear homoscedastic and their scatterplots resemble normal probability distributions. All in all, the observed point differentials could not be predicted any better based on the team ratings.

3.3.5. *Predicted Winning Probabilities.* The Kalman filter algorithm and equation (26) in Subsection 2.6.3 are used to calculate the predicted probability of a home win for each of the games played in the NBA season 2009-2010. The predicted probabilities are presented in Figures 9 and 10 for the parameter vectors $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$, respectively. In the figures, the observed home win is denoted by 1 and a home loss by 0. The figures present also a logistic regression model where the observed wins are explained using the predicted winning probabilities. A dashed line representing perfect match between predicted winning probabilities and the observed outcomes is also presented for comparison.

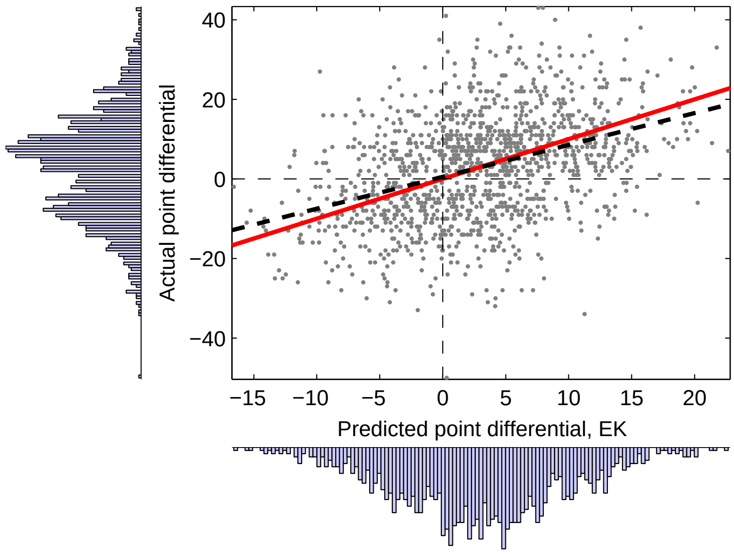

FIGURE 7. Scatterplot presenting the predicted and observed point differentials with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$.

In Figure 9, the logistic regression model has a constant term −1.681 and slope 3.371. The deviance of the model is 1554 which is small compared to the model's error degrees of freedom 1306, i.e., the model gives an adequate representation for the observed outcomes. The theoretical dispersion parameter of the model is 1 and estimated dispersion parameter is 1.001 In other words, the observed outcomes do not include significant variation excluding that included in the model. The predictions appear unbiased as the only deviation from the line representing perfect fit is due to the functional form of the logistic regression model. For example, if the predicted winning probability is 0.50 for the home team, the logistic regression fitted to the observed outcomes gives the value 0.501 for the probability.

In Figure 10, the logistic regression model has a constant term −1.578 and slope 3.250. The deviance of the model is 1559 which small compared to the model's error degrees of freedom 1306, i.e., the model gives an adequate representation for the observed outcomes. The theoretical dispersion parameter of the model is 1 and estimated dispersion parameter is 1.000 In other words, the observed outcomes do not include significant variation excluding that included in the model. The predictions appear unbiased as the only deviation from the line representing perfect

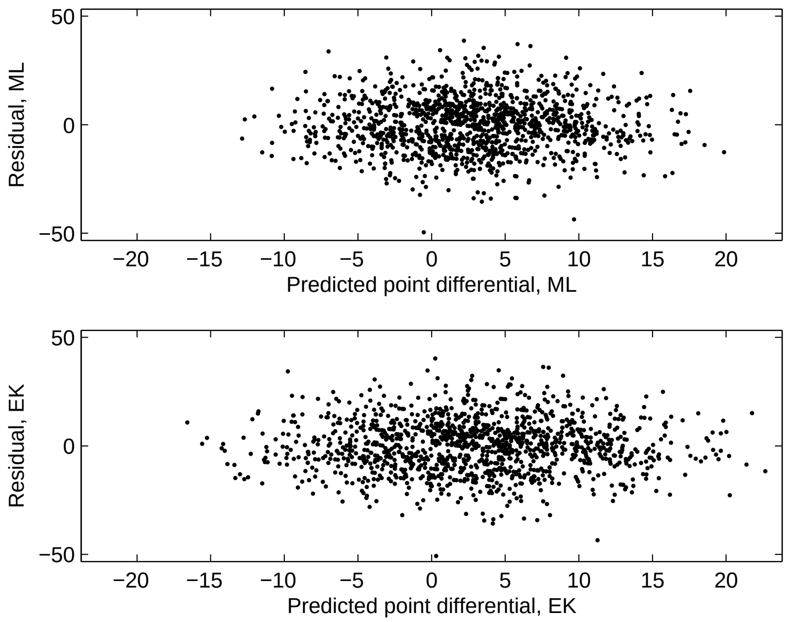

FIGURE 8. Scatterplot presenting the predicted point differentials and the residuals with parameter vectors $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$.

fit is due to the functional form of the logistic regression model. For example, if the predicted winning probability is 0.50 for the home team, the logistic regression fitted to the observed outcomes gives the value 0.511 for the probability.

3.3.6. *Comparison of Models with Alternative Parameter Estimates.* To evaluate the effect of alternative parameter estimates, the behavior of the ratings produced by parameter vectors $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$. In Figure 11, the predicted point differentials for models with parameter vectors $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ are plotted. The red line denotes the line where $x = y$ and the black dashed line is the linear regression model fitted to the data. In the regression model, constant term is 0.424, slope 0.807, $R^2 = 0.962$, and P-value for the complete model is 0.000. The fit between the two predictions is good. The predictions given by $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ are more conservative, i.e., smaller in absolute value. This is due to the greater variation in the ratings given by $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ and its larger variance estimates. There is also heteroscedasticity in the data. That is, the differences between the two predictions increase in magnitude towards the edges of the data set.

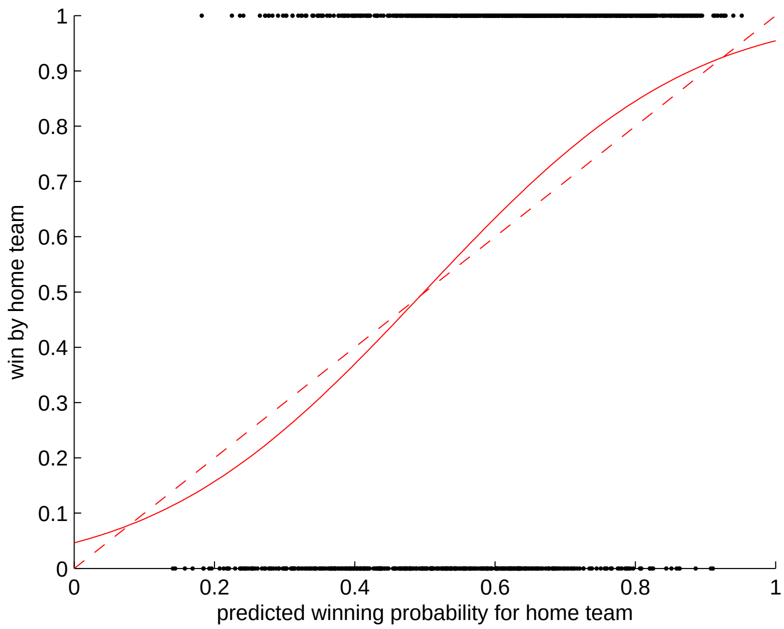

FIGURE 9. Scatterplot presenting the predicted winning probabilities for the home team and the wins by the home team with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$.

As discussed in Subsection 2.3.8, the total uncertainty related to the system state can be evaluated using various measures of uncertainty. In Figure, two measures of uncertainty about the system state with parameter vectors $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ are presented, viz., the total dispersion and generalized variance of the corresponding covariance matrixes. For parameter estimates $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ the uncertainty is always greater than for $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$. This is reasonable as the larger variance estimates in $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ represent the uncertainty related to the system before the season. In ML estimation of $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$, the scores of the games have already been observed and optimal parameter values can be calculated from the data.

The effect of the larger variance estimates is also presented in Figure 13. The figure shows the changes in the maximum of the diagonal elements of the Kalman gain matrix in equation (11). It is seen that the corrections made to the estimated system state are approximately twice as large when estimates $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ are used, in comparison to those resulting from $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$. In practice, this means that there exists larger fluctuations in the team ratings.

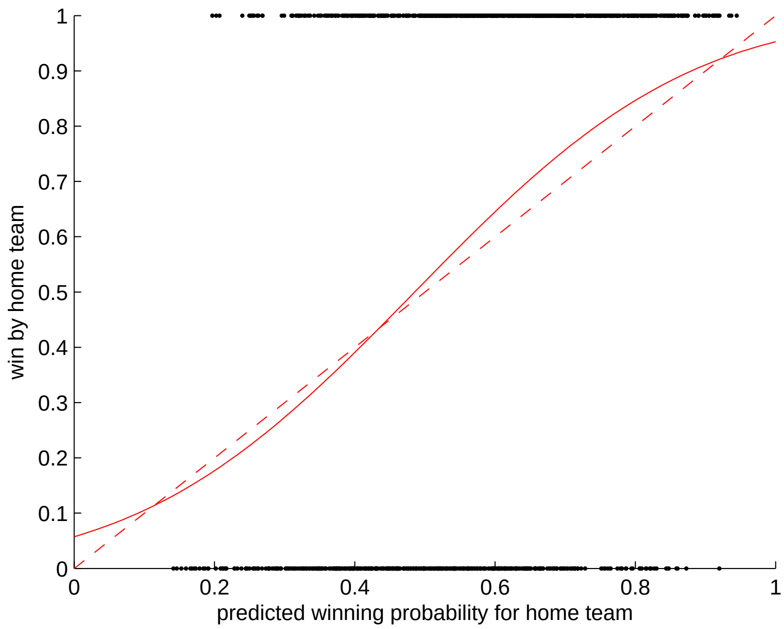

FIGURE 10. Scatterplot presenting the predicted winning probabilities for the home team and the wins by the home team with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$.

The final state estimate $\mathbf{x_{T|T}}$ given by $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ can be seen as the best available estimate for the system state when the scores of the entire season are utilized. Therefore, in Figure 14 the state estimates $\mathbf{x_{t|t}}$ given by the two parameter vectors are compared to the final state estimate by using linear coefficient of correlation. It is seen that the correlation between the state estimates $\mathbf{x_{t|t}}$ and $\mathbf{x_{T|T}}$ increases rapidly at the beginning of the season and by the 50th day a rough understanding of the teams' relative strengths has emerged. The state estimates given by parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ approach the final estimates more slowly, but in the end, they too approach the similar ordering of the teams (cf. Section 3.4). In other words, the "warm-up" time is smaller when the ML estimates are used for the parameters and true ranking of the teams is found more rapidly.

3.4. **NBA 2010 Finals.** Before the final series, LAL is rated at 5.27 and BOS at 4.14, meaning that LAL is appraised approximately one point better than BOS. During the series, there are only moderate changes in the ratings, only exception being game 6 which was a blowout for LAL. One should note that the sum of

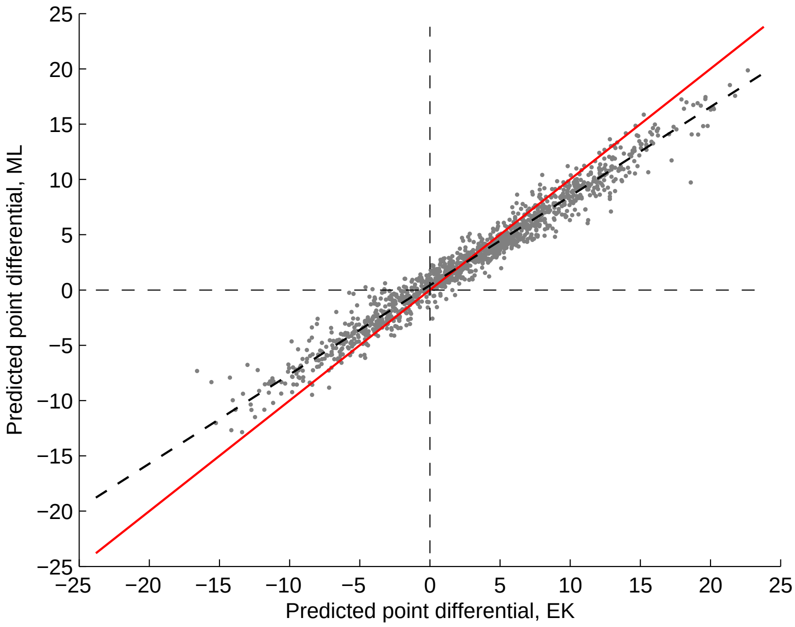

FIGURE 11. Scatterplot presenting the predicted point differentials for models with parameter vectors $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$.

Game	LAL	BOS	P-value	Home team	Pred. point diff.	Winning prob. (game)	Winning prob. (series)	Winner	Obs. point diff.
1	5.27	4.14	0.335	LAL	3.98	0.822	0.776	**LAL**	13
2	5.49	3.92	0.272	LAL	4.42	0.849	0.881	BOS	−9
3	5.17	4.24	0.357	BOS	−1.92	0.326	0.536	LAL	7
4	5.38	4.04	0.296	BOS	−1.51	0.361	0.867	**BOS**	−7
5	5.26	4.16	0.328	BOS	−1.75	0.338	0.666	**BOS**	−6
6	5.16	4.25	0.353	LAL	3.77	0.816	0.344	**LAL**	22
7	5.55	3.87	0.241	LAL	4.53	0.862	0.862	**LAL**	4
-	5.54	3.88	0.241	-	-	-	1.000	**LAL**	-

TABLE 6. Results of the NBA 2010 Finals. The columns LAL and BOS present the teams' ratings before the game. In the Winner column, correctly predicted winners are bolded.

ratings remains almost constant (up to the second decimal). Thus, the games played between LAL and BOS do not affect their average rating compared to other teams. In a sense, the playoff series are a ”zero sum game” where the increase in one team's rating causes a decrease in the other's.

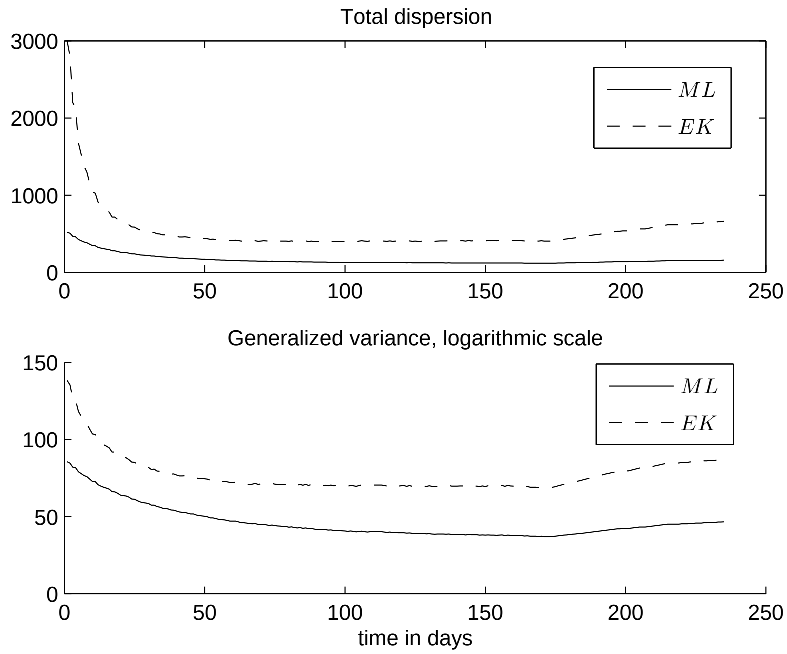

FIGURE 12. Two measures of uncertainty about the system state with parameter vectors $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ for the NBA season 2009-2010.

If the teams are compared with the test presented in Subsection 2.6.2, the P-value for the test is equal to 0.3352, i.e., the difference between teams' ratings is not statistically significant at 0.05 significance level. After seven game series, the P-value has decreased only little, but the direction of the change makes a stronger case for the superiority of LAL.

The teams are so evenly matched that the home is the favorite in every game. Additionally, the predicted winning probabilities and point differential remain almost constant during the series. At home, LAL is expected to win with a probability ranging from 0.816 to 0.862. Away, the predicted winning probabilities are between 0.326 and 0.361. Similarly, the predicted point differentials are from 3.77 to 4.53 in Los Angeles and from -1.92 to -1.51 in Boston. One should note that the observed point differentials are much larger in absolute value than the predictions – which is how it should be.

Before the series, the predictions given by the Kalman filter algorithm present LAL as the favorite to win the series as the probability of LAL first reaching four wins is 0.776. This probability changes with every observed score and the changes

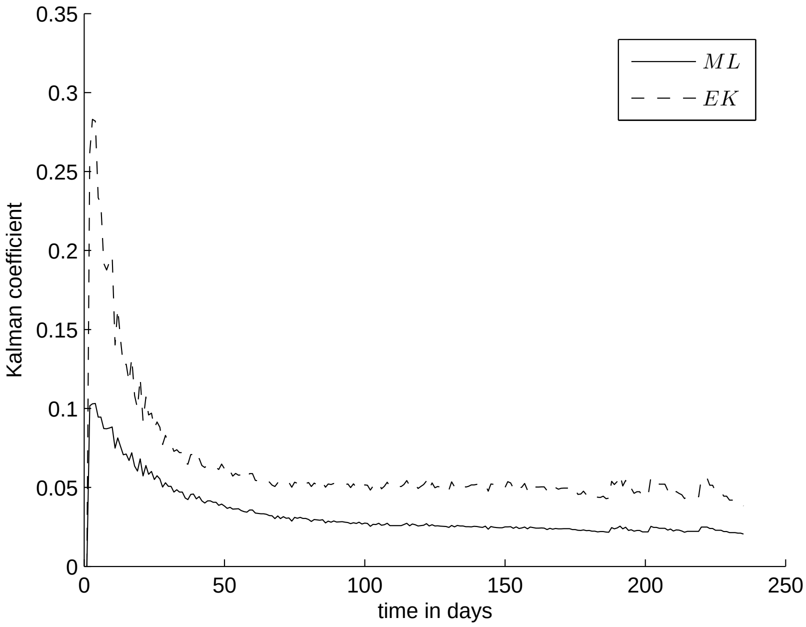

FIGURE 13. Maximum of the Kalman coefficients as a function of time with parameter vectors $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ for the NBA season 2009-2010.

are much more dramatic than the changes in team strengths.[12] The most influential games are those where the team is able to upset its opponent away from home. In the 2010 Finals, BOS was able "to steal the home court advantage" in game 2, tying the series at $1-1$ and dropping the LAL probability of winning the series to 0.536. Unfortunately for BOS, LAL reclaims the home court advantage in game 3 by winning in Boston and raising its probability of winning the series back to 0.867. After BOS wins games 4 and 5 in Boston and leads the series $3-2$, it is in fact the favorite to win the entire series as LAL championship would necessitate two consecutive LAL victories in game 6 and 7. For BOS on the other hand, winning either one of those game would clinch the series. Then, LAL wins game 6 by 22 points extending the finals to the seventh and decisive game that is played in Los Angeles. At this point, LAL probability of winning the game 7, and the entire

[12]This is completely natural as every game played means an additional win for one of the teams bringing the team closer to the championship.

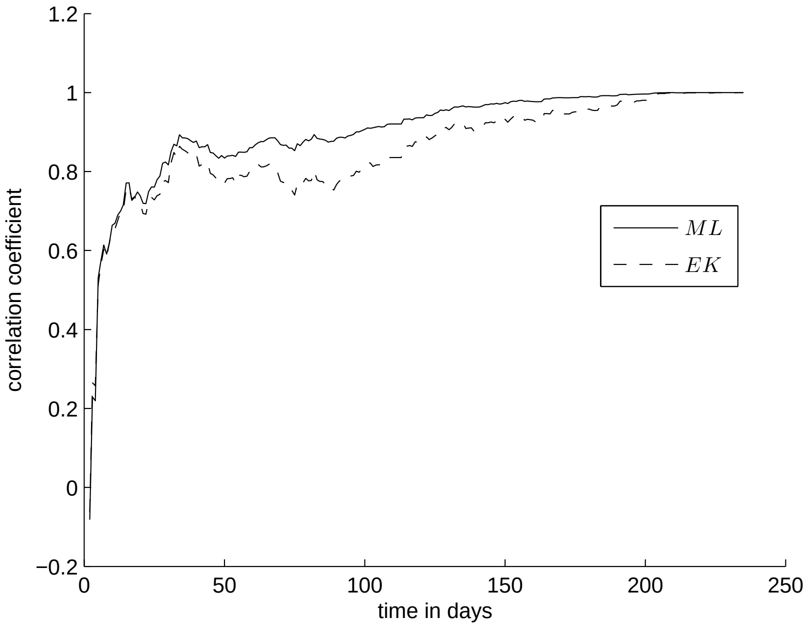

FIGURE 14. Coefficient of linear correlation between the state estimates $\mathbf{x_{t|t}}$ and $\mathbf{x_{T|T}}$ with parameter vectors $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ for the NBA season 2009-2010.

series, is 0.862. In the end, LAL wins the game 7 and the series $4 - 3$. Overall, the Kalman filter algorithm predicts five of the seven finals games correctly.[13]

Finally, it is important to remark that the ratings of the NBA finalists LAL and BOS are only the third and seventh best in the NBA at the end of 2009-2010 season – even though the finalists are undoubtedly the two best teams in the NBA. Apparently, the statistical analysis of point differentials does not tell the entire story about the progress of a NBA season. This shortcoming is addressed in the following quote from Bradford Doolittle and Kevin Pelton, the authors of Basketball Prospectus 2010-11 (Doolittle and Pelton 2010), one of the most respected compilations of advanced NBA statistics:

> The moral of the 2010 postseason, after all, was more about misleading statistics. By no measure did the Boston Celtics look like an elite team after limping into the playoffs with an 11-11 record in March and April. The defending champion Los Angeles Lakers claimed the

[13]Yet, one should keep in mind that identical predictions about the outcomes of games would have been made by simply selecting the home team as the favorite.

Western Conference's top seed, meanwhile, but did so with a point differential that ranked third in the West.[14] Ultimately, those numbers did not matter during the playoffs. The Celtics and Lakers both found additional gears, demonstrating that their regular-season performance painted an inaccurate picture of their ability. While we weren't fooled by the Lakers' middling numbers, both of us picked against Boston prior to the Celtics' upset wins over Cleveland and Orlando.

The lesson here is not to abandon statistics entirely. After all, the Celtics' run was noteworthy precisely because it was so unlikely and unexpected. The NBA's grueling seven-game series leave relatively little room for the kind of fluke outcomes that are commonplace in the MLB (Major League Baseball) and NFL post seasons; more often than not teams perform to their regular-season level during the playoffs. Instead, the takeaway is a healthy humility about the limitations of statistical analysis, which cannot account for every external factor like health or even motivation.

3.5. **Final Team Ratings for the NBA Season 2009-2010.** The final team ratings for the NBA season 2009-2010 calculated using parameter vectors $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ are presented in Table 7. The table also includes the teams winning percentages, point differentials and point differentials over last 60 days, for comparison. The team ratings are followed by P-values in parenThesis.

The P-values presented in Table 7 show the statistical significance of the differences of team ratings. The teams are compared to the highest rated team (ORL) using as discussed in Subsection 2.6.2. The teams have been ordered in a descending order with respect to the ratings given by parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$. Therefore, the test statistic used in the comparison should increase along the rows resulting in monotonously decreasing P-values, i.e., stronger evidence on the difference between the teams. This is indeed the case for most teams. The few exceptions follow from the correlations between the teams' ratings. Across the board, the P-values related to the ratings produced by parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ are larger than those given by $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$. This is follows from the larger values of variance parameters in $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$. In effect,

[14]One should note that the Kalman ratings are highly correlated with the cumulative point differentials referred to by Doolittle and Pelton as is seen in Section 3.5.

Rank	Team	W	L	W%	Point diff.	Point diff. (60 days)	Team rating, $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$	Team rating, $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$
1.	ORL	69	27	0.719	7.66	10.05	8.76 (-)	10.23 (-)
2.	PHO	64	34	0.653	4.86	5.96	6.56 (0.1995)	8.11 (0.3275)
3.	LAL	73	32	0.695	4.52	3.64	5.55 (0.1269)	7.01 (0.2761)
4.	CLE	65	25	0.722	6.01	4.27	5.30 (0.1075)	5.25 (0.1714)
5.	UTA	57	35	0.620	4.63	2.54	4.75 (0.0791)	3.94 (0.0460)
6.	SAS	54	38	0.587	4.16	3.73	4.52 (0.0651)	4.75 (0.1484)
7.	BOS	64	41	0.610	3.19	1.14	3.87 (0.0443)	5.15 (0.1743)
8.	DEN	55	33	0.625	3.74	2.18	3.48 (0.0123)	2.82 (0.0663)
9.	OKC	52	36	0.591	3.03	2.10	3.40 (0.0300)	3.41 (0.1059)
10.	DAL	57	31	0.648	2.63	3.14	3.04 (0.0169)	4.02 (0.1287)
11.	ATL	57	36	0.613	3.28	0.55	2.32 (0.0119)	0.99 (0.0445)
12.	POR	52	36	0.591	2.38	1.96	2.29 (0.0048)	2.13 (0.0687)
13.	MIA	48	39	0.552	1.68	2.31	1.81 (0.0072)	1.91 (0.0260)
14.	MIL	49	40	0.551	1.27	1.59	1.47 (0.0044)	1.16 (0.0471)
15.	CHA	44	42	0.512	0.92	1.53	1.13 (0.0036)	1.16 (0.0418)
16.	HOU	42	40	0.512	-0.37	-0.97	-0.63 (0.0003)	-1.19 (0.0186)
17.	CHI	42	45	0.483	-2.07	-4.45	-1.13 (0.0003)	-0.65 (0.0231)
18.	IND	32	50	0.390	-3.01	0.33	-1.53 (0.0002)	0.44 (0.0276)
19.	MEM	40	42	0.488	-1.50	-2.29	-1.75 (0.0001)	-2.98 (0.0080)
20.	NOH	37	45	0.451	-2.37	-4.68	-2.36 (0.0000)	-2.92 (0.0086)
21.	TOR	40	42	0.488	-1.96	-4.33	-2.41 (0.0000)	-2.94 (0.0043)
22.	GSW	26	56	0.317	-3.60	-3.10	-2.99 (0.0000)	-3.01 (0.0050)
23.	PHI	27	55	0.329	-3.94	-7.33	-4.00 (0.0000)	-4.48 (0.0035)
24.	NYK	29	53	0.354	-3.48	-5.53	-4.12 (0.0000)	-4.09 (0.0012)
25.	SAC	25	57	0.305	-4.37	-5.37	-4.52 (0.0000)	-5.41 (0.0021)
26.	WAS	26	56	0.317	-4.73	-5.63	-4.71 (0.0000)	-4.79 (0.0030)
27.	DET	27	55	0.329	-5.11	-5.39	-5.08 (0.0000)	-5.10 (0.0017)
28.	LAC	29	53	0.354	-6.35	-9.90	-6.86 (0.0000)	-8.43 (0.0002)
29.	NJN	12	70	0.146	-9.13	-4.73	-7.42 (0.000)	-6.11 (0.0002)
30.	MIN	15	67	0.183	-9.60	-11.59	-8.73 (0.0000)	-9.55 (0.0000)

Table 7. Team rankings for NBA season 2009-2010 with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$.

the larger values of variance parameters lead to less compelling conclusions when the teams are compared.

The data presented in Table 7 is also illustrated in Figures 15, 16, and 17. Figure 15 shows the dependence between the final ratings produced by two alternative parameter estimates, viz., $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$. The red line in the figure denotes the

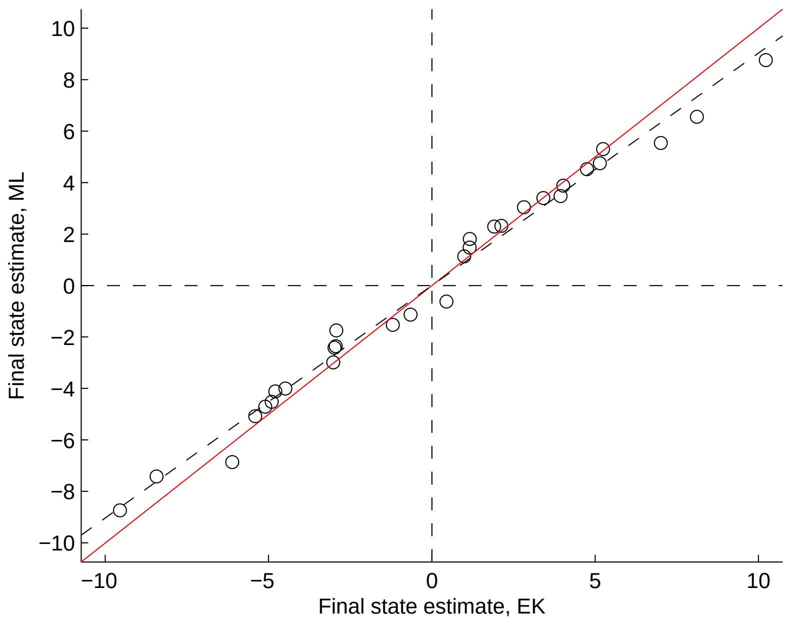

FIGURE 15. Scatterplot presenting the final team ratings with parameter vectors $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$.

line where $x = y$. Clearly, the ratings are consistent with one another, i.e., the teams' ratings are more or less equal with both parameter estimates. This is also represented by the linear regression model fitted to the data (black dashed line). In the regression model, the constant term is 0.000 and the slope is 0.903. The fit is almost perfect giving the coefficient of determination $R^2 = 0.986$, and P-value for the regressor is 0.000.

The ratings produced by $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ are plotted with other measures of performance in Figure 16. The figure shows that there is considerable positive correlation between the performance measures. Interestingly, the teams' ratings match their point differentials. That is, when the full seasons scores are used to determine the ML estimates for the model parameters, the model reproduces the ranking given by the point differentials. Figure 17, similar comparison is done for the ratings based on $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$. Now, the positive dependence is also observed, but the ratings do not match the point differentials as exactly as in Figure 16. Overall, the four discussed performance measures result in very similar rankings for the teams.

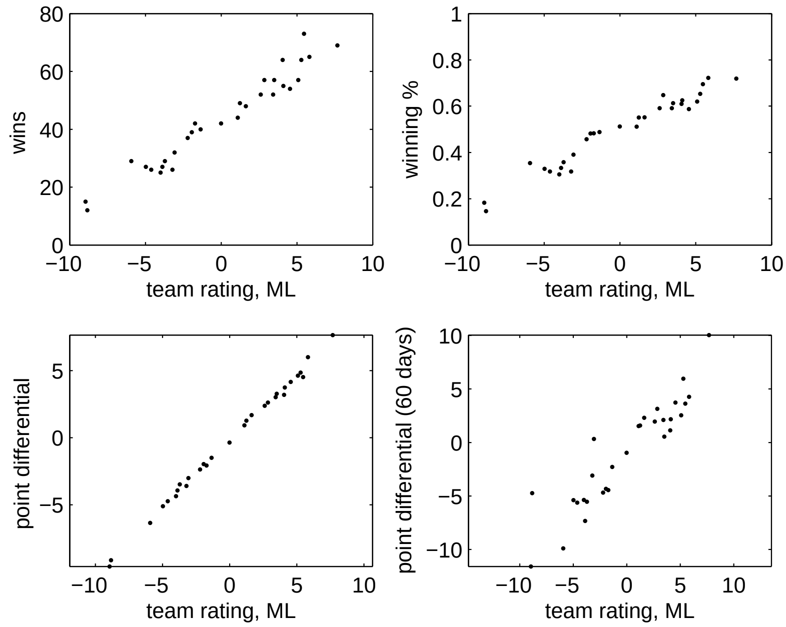

FIGURE 16. Scatterplot presenting the final team ratings with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ and other measures of performance.

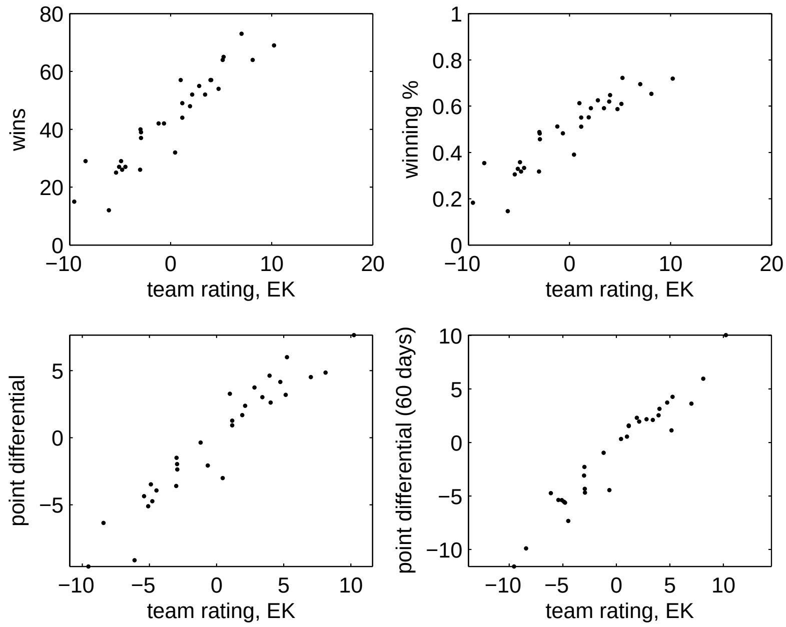

FIGURE 17. Scatterplot presenting the final team ratings with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$ and other measures of performance.

4. Alternative Models and Topics for Future Research

The state-space model presented in this Thesis is intentionally minimal and simplistic and it could be developed further in various ways. In the following, a few ideas for future research are introduced. First, alternative state-space models are proposed. Then, some ideas for other applications of the state-space model and the Kalman filter algorithm are given.

4.1. **Alternative Models.**

4.1.1. *Additional Factors.* Many additional factors may also be included in the observation equation of the state-space model, i.e., equation (7). Examples of such factors include, among others, the effects of rest, traveling, and roster changes. These factors are included in the following model:

$$\mathbf{y_t}(k) = \mathbf{x_t}(i) - \mathbf{x_t}(j) + \mu + \nu + \xi + \psi + \mathbf{w_t}(k), \tag{29}$$

where μ denotes the home court advantage, ν is the effect of rest, ξ represents the traveling schedule, and ψ represents the roster changes. The effect of home court advantage, denoted by μ, is already discussed in Chapter 2.2.

The amount of rest is a notable factor in the NBA as the teams play 82 regular season games within approximately 170 days (Entine and Small 2008) and, therefore, its effect should be included in the observation equation. On average, for seasons 2004-05 and 2005-06 home teams scored 3.24 more points than the visitors. This effect could be divided into to 0.31 points due to smaller amount of rest for visiting teams in NBA schedule, and 2.93 due to other factors, e.g., the home court advantage. Thus, the inclusion of rest would yield a more detailed representation of the home court advantage resulting in improved ratings and better predictive capability. In this approach, if visiting team has played the previous night but home team has not, a positive term is added to the expected value of the game score. If home team is playing a back-to-back game, a negative term is used. Naturally, if both teams have played the previous night, these effects can neutralize each other.

The traveling schedule is also a real, although not dramatic factor contributing to the NBA' home court advantage (Entine and Small 2008). That is, the travel itself makes a team playing in its third consecutive away game more fatigued than a team with the same game dates, but playing at home. Therefore, the effect of the length of the current road trip should perhaps be included in equation (6).

The effect of roster changes, i.e., the absence of key players due to injuries or suspensions, could also be included in the state-space model. For example, if the visiting team is missing one of its key players, a positive factor can be added to the expected score. However, the effects of home court, rest, and traveling are easily estimated from observed game results, but the effect of roster changes is not as straightforward. First, expert knowledge is needed in determining who the "key players" are for each team. For example, the team's best scorer may not be its most important player and the number of key players may vary from team to team. Second, the existing data has to studied to determine whether teams are playing in full strength or not. This may be tedious, unless some easily screened definition for the key players is given. Third, the effect of missing players may not be additive or linear, i.e., the availability of a capable substitute may remove the effect altogether or the simultaneous absence of several players may render the team completely incapacitated. This is an interesting topic for future research that definitely warrants further attention.

4.1.2. *Team Specific Model Parameters.* The state-space model presented in this Thesis assumes identical and time invariant variance parameters μ, σ_0^2, σ_w^2, and σ_v^2 for all the teams. Alternatively, it would be possible to have different parameters for each team. For example, each team could have its own home court advantage representing, e.g., the fanaticism of the team's supporters, the atmosphere of the team's home arena, and the playing style of the team. Similarly, different values of variance parameters would represent team specific uncertainties and variations.

4.1.3. *Time Variant Model Parameters.* The values of the parameters could also be time variant. This would enable the model to adjust to possible changes in the system dynamics and observation equation. For example, the effect of playoff atmosphere could also be tackled by having different home court advantages for regular season and playoffs. On the other hand, it should be kept in mind that models with parameters that follow complicated time dynamics may become unstable and over-fit into inconsequential variations in the observations.

4.1.4. *Non-Linear Models.* The dependence between the difference of team strengths and the observed game outcomes need not be linear as in equation (6). For example, the observation equation could be replaced with a non-linear alternative such as

$$\mathbf{y_t}(k) = f\left(\mathbf{x_t}(i), \mathbf{x_t}(j)\right) + \mu + \mathbf{w_t}(k),$$

where $f(\mathbf{x_t}(i), \mathbf{x_t}(j))$ is non-linear function of the team strengths. The function could also be made discontinuous in order to exclude the possibility of a tied game. The state of this kind of non-linear model could then be estimated using, e.g., the extended Kalman filter (e.g., Harvey 1989, pp. 160).

4.1.5. *Bimodal Probability Distribution for Point Differential.* In the presented analyses, the ML estimates of the model parameters are calculated based on point differentials that are assumed normally distributed. However, the observed point differentials follow a bimodal probability distribution that deviates from the model assumptions. Therefore, the state-space model could possibly be improved by replacing the normal distribution with some bimodal alternative. For example, a suitable non-linear mapping applied to the predictions given by the Kalman filter algorithm could produce predictions more in line with the observed distributions.

4.2. **Other Applications.**

4.2.1. *Model for Total Points.* In over/under betting, the bettor is supposed to guess whether the total number of points scored in a game surpasses a limit set by the bookmaker, viz. the over/under line. In practice, this involves the bettor's view on the tempo of the game (fast pace results in more scoring opportunities) and the playing styles of the teams (high scoring vs. defensive minded). The state-space model introduced in this Thesis could be easily modified for modeling the playing styles of the teams. That is, instead of the point differential, the vector $\mathbf{y_t}$ would represent the total number of points scored in a game resulting in the following observation equation:

$$\mathbf{y_t}(k) = \mathbf{x_t}(i) + \mathbf{x_t}(j) + \mathbf{w_t}(k).$$

This would give a tool for predicting the total number of points scored in a game and assessing over/under -type bets.

4.2.2. *Other Predictions.* In this Thesis, the Kalman filter algorithm is used to predict the results of individual games, viz., $\hat{\mathbf{y}}_{\mathbf{t+1}}$. Similar predictions could also be made with a longer time horizon, e.g., $\hat{\mathbf{y}}_{\mathbf{t+2}}, \ldots, \hat{\mathbf{y}}_{\mathbf{T}}$. Thus, the current team ratings could be combined with league schedule to produce predictions for the rest of the season. These predictions could be collected into the implied final standings for the regular season. Furthermore, the final standings could then used to determine the seeding of the teams and the team matchups for the playoffs. Naturally, the algorithm could also produce predictions for all playoff series and, ultimately, the

NBA Finals. Thus, the Kalman filter algorithm could be used to estimate each teams' probability of making the playoffs or winning the NBA championship, at any time during the season. In practice, such calculations would be easiest to perform using some kind of Monte Carlo simulation (e.g., Fishman 1996).

4.2.3. *Optimal Smoothing.* In the presented analyses, the team ratings are given ex ante, i.e., the estimate of system state $\hat{\mathbf{x}}_{\mathbf{t|t}}$ at time t is based on the games played by that date. In an retrospective approach, the team ratings could be calculated after the season has been played. In such ex post analysis, optimal smoothing could be used to obtain the smoothed estimates of system state, i.e., $\hat{\mathbf{x}}_{\mathbf{t|T}}$, that use all the available information at time T (e.g., Harvey 1989, pp. 149).

4.2.4. *Measuring Competitiveness.* The ratings produced by the Kalman filter could also be used to measure the level of competitiveness of the league, i.e., the parity between the strengths of the teams. A suitable measure for the competitiveness would be given by the variance of the team ratings where smaller variance implies higher level of competitiveness. For a more detailed discussion of competitiveness, see, (Yilmaz and Chatterjee 2000).

5. Conclusions

This Thesis presents a state-space model for a basketball league and the estimation of the system state using a Kalman filter algorithm. Now, the system state represents the relative strengths of the teams. In the state-space model, the teams' strengths follow a discrete time random walk and the state of the system is observed through the game results. The Kalman filter algorithm is used to update the estimate of the system state in accordance to the observed scores and this estimate is referred to as the rating of the teams. The rating of the teams gives the ordering of the teams from the best to the worst, viz., the ranking of the teams. The ratings can also be used to predict the scores of the future games.

Based on the example analyses related to the National Basketball Association (NBA) that are presented in this Thesis, the team ratings given by the Kalman filter algorithm give an accurate and informative representation for the time evolution of the basketball season and the assumptions behind the model appear more or less valid.[15] In the presented analyses, the algorithm predicts correctly $65 - 70\%$ of the outcomes of the NBA games which is slightly better than the simplest performance measures excluding the point differential. If only the second half of the season is considered, the performance improves to $66 - 73\%$. Additionally, the team ratings explain approximately 16.5% of the total variation in observed point differentials. Most importantly, on average the predictions given by the algorithm appear unbiased and independent of the residuals. This holds for both the predicted point differentials and the predicted winning probabilities. Therefore, the algorithm can be said to produce accurate and correct predictions for the results of the NBA games.

In this Thesis, the parameters of the underlying state-space model are estimated using maximum likelihood (ML) method and other means. In the ML estimation, the numerical results provided by the Kalman filter are utilized for increased efficiency. When alternative estimates for the parameters of the state-space model were compared, it was found that the ML estimates give the best predictions. This is to be expected as the ML estimation exploits the observed scores for the entire season, unlike the parameter estimates calibrated by the expert. However, this finding is far from conclusive as the performance of the algorithm does not decline significantly if expert analysis or previous season's scores are used in parameter estimation. The

[15]This holds true also for numerous other analyses conducted by the author. These analyses are not reported in this Thesis, however.

ML estimates discussed in this Thesis are based on point differentials following a normal distribution but this is not the case in the observed scores of the NBA games. Therefore, the model could be improved by using some kind of bimodal distribution for the point differentials. On the other hand, the parameter estimates given by the expert are overtly conservative, i.e., the variances are evaluated as too large. This increases rapid changes in the ratings and slows down the convergence of the ratings. If future studies are conducted based on expert analyses, it might be a good idea to employ smaller estimates for the variance parameters.

As a criticism of the Kalman filter algorithm, one could claim that the it does not produce radically better results than the simple calculation of average point differentials due to the high correlation between the two performance measures. This criticism is countered by pointing out that this Thesis introduces a state-space model for the basketball league and that the Kalman filter algorithm is only used to estimate its state. Thus, the ratings produced by the Kalman filter algorithm are based on a dynamic and probabilistic model of a system instead of calculating arbitrary differences in scoring averages. That is, the objective of a basketball team in a given game is to score more points than its opponent – not to outscore the opponent by highest margin possible. Therefore, one could also articulate that the state-space model and Kalman filter algorithm give a mathematical foundation and justification to the use of point differential as a measure of teams' level of performance.

As a final note, there is more to winning in the NBA, and particularly in the NBA playoffs, than the point differentials achieved during the regular season (as is discussed in Section 3.4). In a seven game playoff series, the winner is not determined by the past performance but by the ability elevate one's game at the critical moment. To quote legendary NBA coach Rudy Tomjanovich, "don't ever underestimate the heart of a champion" (Feinstein 2002). On the other hand, keeping this caveat in mind, the Kalman filter algorithm produces believable ratings and accurate predictions for the duration of the regular season. Therefore, it is to be considered a useful tool for rating of the teams and following the progress of an ongoing season.

References

J. Arkes. Revisiting the hot hand theory with free throw data in a multivariate framework. *Journal of Quantitative Analysis in Sports*, 6(1):1–10, 2010.

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

G. W. Bassett, Jr. Point spreads versus odds. *The Journal of Political Economy*, 89(4):752–768, Aug. 1981.

G. W. Bassett, Jr. Robust sports ratings based on least absolute error. *The American Statistician*, 51(2):99–107, May 1997.

G. W. Bassett, Jr. Quantile regression for rating teams. *Statistical Modeling*, 7(4): 301–313, Dec. 2007.

G. W. Bassett, Jr. and W. Hurley. The effects of alternative home-away sequences in a best-of-seven playoff series. *The American Statistician*, 52(1):51–53, Feb. 1998.

D. J. Berri. Who is 'most valuable'? measuring the player's production of wins in the National Basketball Association. *Managerial and Decision Economics*, 20(8): 411–427, 1999.

D. J. Berri, M. B. Schmidt, and S. L. Brook. *The Wages of Wins: Taking Measure of the Many Myths in Modern Sport*. Stanford University Press, Stanford, CA, 2007.

D. P. Bertsekas. *Dynamic Programming & Optimal Control*, volume I. Athena Scientific, Nashua, NH, 2005.

B. L. Boulier and H. O. Stekler. Are sports seedings good predictors?: An evaluation. *International Journal of Forecasting*, 15(1):83–91, Feb. 1999.

B. L. Boulier and H. O. Stekler. Predicting the oucomes of National Football League games. *International Journal of Forecasting*, 19(2):257–270, Apr.-Jun. 2003.

B. P. Carlin. Improved NCAA basketball tounament modeling via point spread and team strength information. *The American Statistician*, 50(6):39–43, Feb. 1996.

G. Casella and R. Berger. *Statistical Inference*. Duxbury Resource Center, 2001.

C. R. Cassady, L. M. Maillart, and S. Salman. Ranking sports teams: A customizable quadratic assignment approach. *Interfaces*, 35(6):497–510, Nov.-Dec. 2005.

S. B. Caudill and N. H. Godwin. Heterogeneous skewness in binary choice models: predicting outcomes in the men's NCAA basketball tournament. *Journal of Applied Statistics*, 29(7):991–1001, Sep. 2002.

S. Chatterjee and M. R. Yilmaz. The NBA as an evolving multivariate system. *The American Statistician*, 53(3):257–262, Aug. 1999.

B. J. Coleman. Minimizing game score violations in college football rankings. *Interfaces*, 35(6):483–496, Nov.-Dec. 2005.

B. J. Coleman and A. K. Lynch. NCAA tournament games: The real nitty-gritty. *Journal of Quantitative Analysis in Sports*, 5(3):1–27, 2009.

L. L. Colquitt, N. H. Godwin, and R. T. Shortridge. The effects of uncertainty on market prices: Evidence in coaching changes in the NBA. *Journal of Business Finance and Accounting*, 34(5–6):861–871, Jun.-Jul. 2007.

K. S. Courneya and A. V. Carron. The home advantage in sport competitions: A literature review. *Journal of Sport and Exercise Psychology*, 14:13–27, 1992.

R. Davidson and J. G. MacKinnon. *Estimation and inference in econometrics*. Oxford University Press, New York, NY, 1993.

R. Davidson and J. G. MacKinnon. *Econometric Theory and Methods*. Oxford University Press, New York, NY, 2004.

B. Doolittle and K. Pelton. *Pro Basketball Prospectus 2010-11*, volume 2. CreateSpace, 2010.

O. Entine and D. Small. The role of rest in the NBA home-court advantage. *Journal of Quantitative Analysis in Sports*, 4(2):6, 2008.

P. Fearnhead and B. M. Taylor. Calculating strength of schedule, and choosing teams for march madness. Unpublished, 2009.

J. Feinstein. *The Punch: One Night, Two Lives, and the Fight That Changed Basketball Forever*. Little, Brown, 2002.

FIBA. Fédération internationale de basketball, home page, 2010. http://www.fiba.com, April 12th, 2011.

G. S. Fishman. *Monte Carlo: Concepts, Algorithms, and Applications (Springer Series in Operations Research)*. Springer-Verlag New York, Inc., New York, NY, 1996.

J. Galinsky. FOX Sports NBA power rankings, 2010. http://msn.foxsports.com/nba/powerRankings, Accessed September 27th, 2010.

J. M. Gandar, W. H. Dare, C. R. Brown, and R. A. Zuber. Informed traders and price variations in the betting market for professional basketball games. *The Journal of Finance*, 53(1):385–401, Feb. 1998.

J. Gibbs. Point shaving in the NBA: An economic analysis of the national basketball association's point spread betting market. Master's thesis, Stanford, 2007.

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

M. E. Glickman and H. S. Stern. A state-space model for National Football League scores. *The Journal of American Statistical Association*, 93(441):25–35, Mar. 1998.

C. Gourieroux, A. Monfort, and A. Trognon. Pseudo maximum likelihood methods: Theory. *Econometrica*, 52(3):681–700, May 1984.

A. Y. Govan. *Ranking Theory with Application to Popular Sports*. PhD thesis, North Carolina State University, Raleigh, North Carolina, 2009.

A. Y. Govan, A. N. Langville, and C. D. Meyer. Offense-defense approach to ranking team sports. *Journal of Quantative Analysis in Sports*, 5(1):1–19, 2009.

J. D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

A. C. Harvey. *Forecasting, Structural Time Series Models, and the Kalman Filter*. Cambridge University Press, 1989.

D. A. Harville. The use of linear-model methodology to rate high school or college football teams. *Journal of the American Statistical Association*, 72(358):278–289, Jun. 1977.

D. A. Harville. Predictions for National Football League games via linear-model methodology. *Journal of American Statistical Association*, 75(371):516–524, Sep. 1980.

D. A. Harville. The selection or seeding of college basketball or football teams for post-season competition. *Journal of American Statistical Association*, 98(461): 17–27, Mar. 2003.

J. Hollinger. Hollinger power rankings, 2011. http://espn.go.com/nba/hollinger/powerrankings, Accessed April 12th, 2011.

F. Hu and J. V. Zidek. Forecasting NBA basketball playoff outcomes using the weighted likelihood. *IMS Lecture Note-Monograph Series*, 45:385–395, 2004.

M. B. Jones. Home advantage in the NBA as a game-long process. *Journal of Quantitative Analysis in Sports*, 3(4):1–14, 2007.

M. B. Jones. A note on team-specific home advantage in the NBA. *Journal of Quantitative Analysis in Sports*, 4(3):1–15, 2008.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

L. Knorr-Held. Dynamic rating of sports teams. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 49(2):261–276, 2000.

R. Koenker and G. W. Bassett Jr. March madness, quantile regression bracketology, and the Hayek hypothesis. *Journal of Business and Economic Statistics*, 28(1):

26–35, Jan. 2010.

M. W. Kraus, C. Huang, and D. Keltner. Tactile communication, cooperation, and performance: An ethological study of the NBA. *Emotion*, 2010. To appear.

S. Kuper and S. Szymanski. *Soccernomics: Why England Loses, Why Germany and Brazil Win, and Why the U.S., Japan, Australia, Turkey–and Even Iraq–Are Destined to Become the Kings of the World's Most Popular Sport*. Nation Books, 2009.

P. Kvam and J. S. Sokol. A logistic regression/Markov chain model for NCAA basketball. *Naval Research Logistics*, 53(8):788–803, Dec. 2006.

M. Lewis. *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton & Company, 2003.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

N. Marttinen. Creating a profitable betting strategy for football using statistical modelling. Master's thesis, Trinity College, Dublin, Ireland, 2001.

K. Massey. Massey ratings. Web site, http://www.masseyratings.com, 1995. Accessed February 17th, 2010.

K. Massey. Statistical models applied to the rating of sports teams. Honors project: Mathematics, Bluefield College, Bluefield, Virginia, 1997.

Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, Jun. 1947.

T. J. Moskowitz and L. J. Wertheim. *Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won*. Crown Archetype, 2011.

S. Mustonen. A measure for total variability in multivariate normal distribution. *Computational Statistics And Data Analysis*, 23:321–334, 1997.

NBA. National basketball association, home page, 2011. http://www.nba.com, Accessed April 12th, 2011.

NCAA. National collegiate athletic association, home page, 2011. http://www.ncaa.com, April 12th, 2011.

J. B. Niemi. Identifying and evaluating contrarian strategies for NCAA tournament pools. Master's thesis, University of Minnesota, 2005.

D. Oliver. *Basketball on Paper: Rules and Tools for Performance Analysis*. Potomac Books, Dulles, VA, 2004.

J. Park and M. Newman. A network-based ranking system for US college football. *Journal of Statistical Mechanics: Theory and Experiment*, Oct. 2005.

R. J. Parker. Modeling basketball's points per possession with application to predicting the outcome of college basketball games. Bachelor's essay, College of Charleston, May 2010.

K. Pomeroy. Pomeroy ratings. Web site, http://www.kenpom.com, 2003. Accessed April 12th, 2011.

A. P. Rotshtein, M. Posner, and A. Rakityanskaya. Football predictions based on a fuzzy model with genetic and neural tuning. *Cybernetics and Systems Analysis*, 41(4):619–630, July 2005.

J. Sagarin. Sagarin ratings. Web site, http://www.usatoday.com/sports/sagarin.htm, 1985. Accessed April 12th, 2011.

S. Sanders. Cheap ticket to the dance: Systematic bias in college basketball. *Economics Bulletin*, 4(34):1–7, 2007.

J. Schuhmann. NBA power rankings, April 2010. http://www.nba.com/2010/news/powerrankings/04/12/week24/index.html, Accessed September 27th, 2010.

N. C. Schwertman, T. A. McCready, and L. Howard. Probability models for the NCAA regional basketball tournaments. *The American Statistician*, 45(1):35–38, Feb. 1991.

N. C. Schwertman, K. L. Schenk, and B. C. Holbrook. More probability models for the NCAA regional basketball tournaments. *The American Statistician*, 50(1): 34–38, Feb. 1996.

S. L. Shapiro, J. Drayer, B. Dwyer, and A. L. Morse. Punching a ticket to the big dance: A critical analysis of at-large selection into the NCAA division i men's basketball tournament. *Journal of Issues in Intercollegiate Athletics*, 2:46–63, 2009.

T. Smith and N. C. Schwertman. Can the NCAA basketball tournament seeding be used to predict margin of victory? *The American Statistician*, 53(2):94–98, May 1999.

B. P. Soebbing and B. R. Humphreys. Do gamblers think that teams tank? Evidence from the NBA. Working paper, 2011.

M. Spann and B. Skiera. Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters. Accepted for publication in Journal of Forecasting, 2008.

R. T. Stefani. Improved least squares football, basketball, and soccer predictions. *Transactions on Systems, Man, and Cybernetics*, 10(2):116–123, Feb. 1980.

M. Stein. ESPN power rankings, 2010. http://espn.go.com/nba/powerrankings, Accessed September 27th, 2010.

H. Stern. On the probability of winning a football game. *The American Statistician*, 45(3):179–183, Aug. 1991.

B. T. West. A simple and flexible rating method for predicting success in the NCAA basketball tournament. *Journal on Quantitative Analysis in Sports*, 2(3): 1–16, 2006.

M. R. Yilmaz and S. Chatterjee. Patterns of NBA team performance from 1950 to 1998. *Journal of Applied Statistics*, 27(5):555–566, July 2000.

T. Zimmer and T. Kuethe. Major conference bias and the NCAA men's basketball tournament. *Economics Bulletin*, 17(12):1–6, 2008.

Appendix A

Rank	Team	W	L	W%	Point diff.	Point diff. (60 days)	Team rating, $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$	Team rating, $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$
1.	ORL	69	27	0.719	7.66	10.05	8.76 (-)	10.23 (-)
2.	PHO	64	34	0.653	4.86	5.96	6.56 (0.1995)	8.11 (0.3275)
3.	LAL	73	32	0.695	4.52	3.64	5.55 (0.1269)	7.01 (0.2761)
4.	CLE	65	25	0.722	6.01	4.27	5.30 (0.1075)	5.25 (0.1714)
5.	BOS	64	41	0.610	3.19	1.14	3.87 (0.0443)	5.15 (0.1743)
6.	SAS	54	38	0.587	4.16	3.73	4.52 (0.0651)	4.75 (0.1484)
7.	DAL	57	31	0.648	2.63	3.14	3.04 (0.0169)	4.02 (0.1287)
8.	UTA	57	35	0.620	4.63	2.54	4.75 (0.0791)	3.94 (0.0460)
9.	OKC	52	36	0.591	3.03	2.10	3.40 (0.0300)	3.41 (0.1059)
10.	DEN	55	33	0.625	3.74	2.18	3.48 (0.0123)	2.82 (0.0663)
11.	POR	52	36	0.591	2.38	1.96	2.29 (0.0048)	2.13 (0.0687)
12.	MIA	48	39	0.552	1.68	2.31	1.81 (0.0072)	1.91 (0.0260)
13.	MIL	49	40	0.551	1.27	1.59	1.47 (0.0044)	1.16 (0.0471)
14.	CHA	44	42	0.512	0.92	1.53	1.13 (0.0036)	1.16 (0.0418)
15.	ATL	57	36	0.613	3.28	0.55	2.32 (0.0119)	0.99 (0.0445)
16.	IND	32	50	0.390	-3.01	0.33	-1.53 (0.0002)	0.44 (0.0276)
17.	CHI	42	45	0.483	-2.07	-4.45	-1.13 (0.0003)	-0.65 (0.0231)
18.	HOU	42	40	0.512	-0.37	-0.97	-0.63 (0.0003)	-1.19 (0.0186)
19.	NOH	37	45	0.451	-2.37	-4.68	-2.36 (0.0000)	-2.92 (0.0086)
20.	TOR	40	42	0.488	-1.96	-4.33	-2.41 (0.0000)	-2.94 (0.0043)
21.	MEM	40	42	0.488	-1.50	-2.29	-1.75 (0.0001)	-2.98 (0.0080)
22.	GSW	26	56	0.317	-3.60	-3.10	-2.99 (0.0000)	-3.01 (0.0050)
23.	NYK	29	53	0.354	-3.48	-5.53	-4.12 (0.0000)	-4.09 (0.0012)
24.	PHI	27	55	0.329	-3.94	-7.33	-4.00 (0.0000)	-4.48 (0.0035)
25.	WAS	26	56	0.317	-4.73	-5.63	-4.71 (0.0000)	-4.79 (0.0030)
26.	DET	27	55	0.329	-5.11	-5.39	-5.08 (0.0000)	-5.10 (0.0017)
27.	SAC	25	57	0.305	-4.37	-5.37	-4.52 (0.0000)	-5.41 (0.0021)
28.	NJN	12	70	0.146	-9.13	-4.73	-7.42 (0.000)	-6.11 (0.0002)
29.	LAC	29	53	0.354	-6.35	-9.90	-6.86 (0.0000)	-8.43 (0.0002)
30.	MIN	15	67	0.183	-9.60	-11.59	-8.73 (0.0000)	-9.55 (0.0000)

Table 8. Team rankings for NBA season 2009-2010 with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$.

Appendix B

Figure 18, Figures 19 and 20 present the ratings of NBA teams during season 2009-2010 with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$.

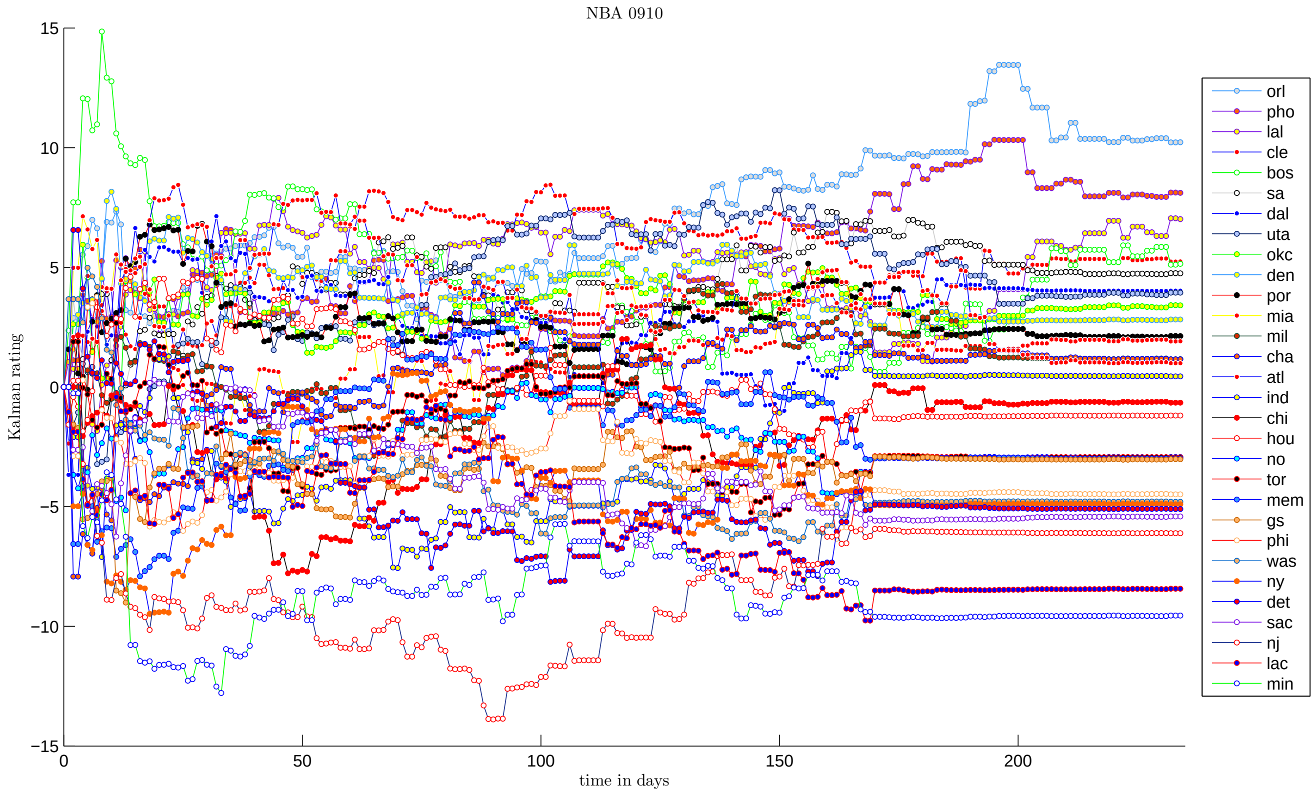

FIGURE 18. Team ratings for the NBA teams during season 2009-2010 with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$.

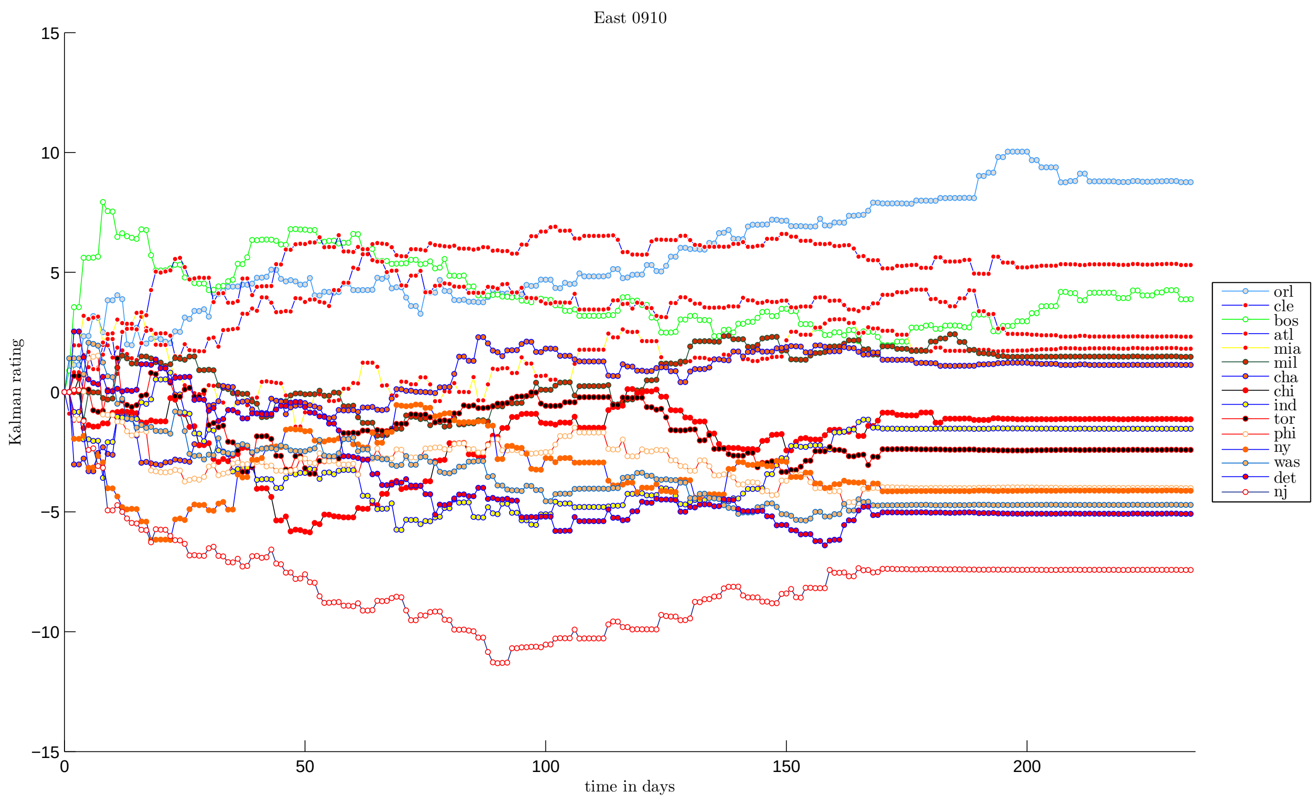

FIGURE 19. Team ratings for the Eastern Conference teams during season 2009-2010 with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$.

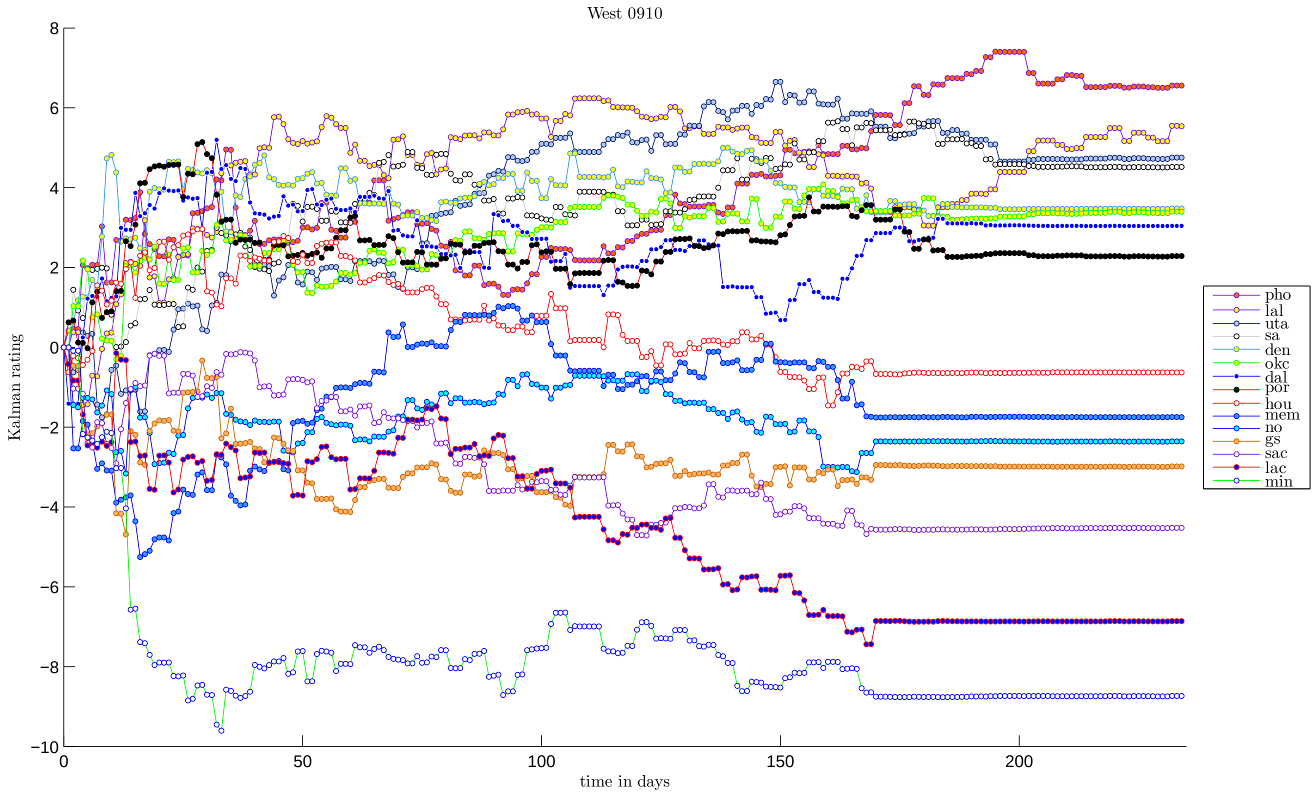

FIGURE 20. Team ratings for the Western Conference teams during season 2009-2010 with parameter vector $\hat{\boldsymbol{\theta}}_{\boldsymbol{EK}}$.