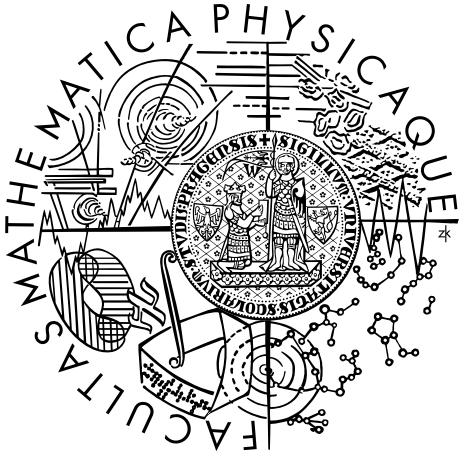
Charles University in Prague
Faculty of Mathematics and Physics

MASTER THESIS



Filip Šimsa

Analysis and prediction of league games results

Department of Probability and Mathematical Statistics

Supervisor of the master thesis:	RNDr. Tomáš Hanzák, Ph.D.
Study programme:	Mathematics
Specialization:	Probability, Mathematical Statistics and Econometrics

Prague 2015

I would like to thank my supervisor, RNDr. Tomáš Hanzák, Ph.D., for his guidance, strong interest in the topic and his friendly approach. I am especially grateful for his prompt responses and constructive suggestions. His active and engaged supervision was paramount to finalize my thesis.

I would also like to express my gratitude to my parents who always supported me during my studies.

I declare that I carried out this diploma thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In Prague on 7th May 2015 Filip Šimsa

Název práce: Analýza a predikce výsledků ligových utkání

Autor: Filip Šimsa

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Tomáš Hanzák, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Práce se zabývá analýzou hokejových utkání v rámci nejvyšší české hokejové soutěže v sezónách 1999/2000 až 2014/2015 a predikcí následujících zápasů. Popisuje a následně aplikuje teorii Kalmanova filtru, kde formy týmů představují nepozorovatelný stavový vektor a výsledky zápasů slouží jako pozorování. Jako vhodná transformace výsledku zápasu jsou identifikovány gólové rozdíly. Ty použijeme jako vysvětlovanou proměnnou také v lineární regresi pro nalezení vhodných prediktorů. Pro předpověď výsledku zápasu je zkonstruován ordinální model s těmito prediktory. Pomocí zobecněného Giniho koeficientu srovnáme diverzifikační schopnost tohoto modelu a kurzů, které nabízí sázkové kanceláře. V závěru využijeme informaci o znalosti kurzů před zápasem a společně v kombinaci s dalšími vysvětlujícími proměnnými vytvoříme predikční model. Tento model je použit pro identifikaci ziskových sázek.

Klíčová slova: Kalmanův filtr, ordinální proměnná, Giniho koeficient, predikce výsledku hokejových utkání

Title: Analysis and prediction of league game results

Author: Filip Šimsa

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Tomáš Hanzák, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The thesis is devoted to an analysis of ice hockey matches results in the highest Czech league competition in seasons 1999/2000 to 2014/2015 and to prediction of the following matches. We describe and apply Kalman filter theory where forms of teams represent an unobservable state vector and results of matches serve as measurements. Goal differences are identified as a suitable transformation of a match result. They are used as a dependent variable in a linear regression to find significant predictors. For a prediction of a match result we construct an ordinal model with those predictors. By using generalized Gini coefficient, we compare a diversification power of this model with betting odds, which are offered by betting companies. At the end, we combine knowledge of odds before a match with other predictors to make a prediction model. This model is used to identify profitable bets.

Keywords: Kalman filter, ordinal variable, Gini coefficient, prediction of ice hockey match results

Contents

Introduction

The aim of this thesis is to make an econometric analyses of ice hockey matches results in the highest Czech league competition. For that reason, we use available data from the main part of Extraliga in seasons 1999/2000 till 2014/2015. Another objective is to examine time dependency and the possibility of state space modelling with forms of teams as an unobservable state vector and results of matches as measurements. Those findings are to be used to construct a suitable prediction model that is to be compared with odds given by betting companies.

The data of ice hockey matches are paired comparisons because we observe a result of a match between two teams. In a particular season they can be seen as longitudinal observations with teams as subjects and rounds as time steps that are multivariate since there are seven matches in every round. Important predictor is a current form of a team, which is an unobservable variable and may change over time. The result of a match is given either by goals of a home team and goals of its opponent or as an ordinal variable with categories win, draw or loss. The discrete form is used by betting companies to set odds, therefore, we are interested in estimating probabilities of win, draw and loss. Those characteristics are taken into account and we develop several models to cope with them.

The thesis is divided into a theoretical part and a practical part. First three chapters are devoted to the theoretical part, which outlines several statistical methods for an analysis of ice hockey matches results and provides a link between them. Those findings are important for better understanding of the topic and some methods used in the practical part. However, a reader that is rather practically oriented in the topic can flip through those chapters and seek practical results in the last two chapters.

in the first chapter we summarise different types of regression models that can be used for modelling longitudinal data. There are briefly described main characteristics and similarities among them. We continue with the concept of Bayesian inference, which is a theoretical background for estimating random parameters and, therefore, for derivation of the Kalman filter. In the last section of the chapter it is defined several goodness of fit criteria. We mainly focus on generalizations of a Gini coefficient, which are used to compare categorical models in later chapters.

The second chapter is devoted to dynamic linear models and to the Kalman filter as a tool for estimating a state vector. We derive the optimal estimator of a state vector in a dynamic linear model and its iterative estimation. We outline two methods for an estimation of hyperparameters and in the last section we explain how to treat missing observations.

in the third chapter we present some generalizations of linear models with stress on a categorical variable as a response because we are interested in predic-

tion of a match result in terms of win, draw and loss. We call this type of a result a plain result throughout the thesis. We show that a categorical variable belongs to an exponential family and briefly describe how Kalman filter could be generalized in case that measurements come from an exponential family. We also present proportional odds model as a suitable model for predicting an ordinal variable.

The practical part is divided into two chapters. In the first one we describe some characteristics of the highest Czech league ice hockey competition Extraliga and some specifics of the analysed data. We examine what is a suitable one-dimensional transformation of an outcome and further, we identify an effect of possible predictors for a result of a match. We analyse the effect of home advantage, result of the previous match, tiredness from the previous match and history of mutual matches. We also suggest several measures of teams performance that could be used to access the current form of a team. The main candidate is presented as the last one and it is modelling of team form as an unobservable state variable in a dynamic linear model with the usage of the Kalman filter. We compare those measures based on the generalized Gini coefficients.

In the last chapter we compare a diversification power of the analytical model, which uses only quantitative information, with odds offered by betting companies. We discuss the possibility and the advantage of using odds as an additional predictor for predicting a result of a match. We construct a final model that is used to identify profitable bets and formulate a strategy used for betting. The strategy is evaluated on the whole dataset and its profitability is discussed.

Chapter 1

Theoretical background

In this chapter we give an overview of several aspects that we use both for theoretical derivations and in the practical analysis. We shortly describe several regression models with stress on their assumptions (the main reference is Fahrmeir and Tutz, 1994). Further, we deal with Bayesian statistics, which is characterized by the principle that parameters are assumed to be random and their estimation is based on loss functions. It is mainly based on books Hušková (1985) and Robert (2007). In the last section we define Gini coefficient and its generalizations as suitable criteria to access goodness of fit for regression models for categorical variables.

1.1 Overview of regression models for longitudinal data

In this section we present various types of regression models that can be applied to longitudinal data. Longitudinal or panel data are observations measured on the same subjects at multiple points in time. Regression technique can be used in order to distinguish effects of certain explanatory variables on the response or to predict future realization of the response being given values of explanatory variables. In the following suppose that we observe data

$$(Y_{it}, \boldsymbol{z}_{it}^\top)^\top,\ i = 1, \dots, n,\ t = 1, \dots, T$$

where Y_{it} is a response and $\boldsymbol{z}_{it}$ is a vector of regressors. Some models take advantage of multivariate definition, stacking together either different subjects i or different time points t. Therefore, we denote

$$\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{it})^\top, \qquad \boldsymbol{Y}_t = (Y_{1t}, \dots, Y_{nt})^\top.$$

Further, we present regression models that have different assumptions about dependence of mean of Y_{it} on regressors $\boldsymbol{z}_{it}$, distribution of Y_{it}, regressor parameters and about the dependence structure of Y_{it} and regressor parameters. In the following we assume that $\boldsymbol{z}_{it}$ is deterministic at time t.

Linear model (LM)

Linear model is specified by the following assumptions

(i) $\mathsf{E}\, Y_{it} = \boldsymbol{z}_{it}^\top \boldsymbol{\beta}$,

(ii) $Y_{it}|\boldsymbol{\beta} \sim \mathsf{N}(\boldsymbol{z}_{it}^\top \boldsymbol{\beta}, \sigma^2)$,

(iii) $\boldsymbol{\beta}$ is a constant (fixed effect),

(iv) $Y_{it},\ i = 1, \ldots, n,\ t = 1, \ldots, T$ are independent.

This model is the fundamental regression model. In this text we suppose that the reader is familiar with its usage and properties. Parameter estimation, dealing with violation of certain assumptions and goodness of fit criteria can be found in Anděl (2011), Cipra (2008) or Zvára (2008).

Generalized linear model (GLM)

Generalized linear model is specified by the following assumptions

(i) $g(\mathsf{E}\, Y_{it}) = \boldsymbol{z}_{it}^\top \boldsymbol{\beta}$,

(ii) $Y_{it}|\boldsymbol{\beta}$ comes from an exponential family (see definition 6),

(iii) $\boldsymbol{\beta}$ is a constant (fixed effect),

(iv) $Y_{it},\ i = 1, \ldots, n,\ t = 1, \ldots, T$ are independent.

The assumption of linear dependence of mean of Y_{it} on $\boldsymbol{z}_{it}$ is generalized to dependence through a function g and distribution comes from a broader family, including alternative or Poisson distributions. Its definition and parameter estimation is outlined in Farhmeir and Tutz (1994, chapter 2). Its multivariate extension is briefly presented in section (3.1).

Linear mixed effect model (LMM)

Linear mixed effect model is specified by the following assumptions

(i) $\mathsf{E}\,[Y_{it}|\boldsymbol{\alpha}_i] = \boldsymbol{z}_{it}^{1\top} \boldsymbol{\beta} + \boldsymbol{z}_{it}^{2\top} \boldsymbol{\alpha}_i$,

(ii) $Y_{it}|\boldsymbol{\alpha}_i \sim \mathsf{N}(\boldsymbol{z}_{it}^{1\top} \boldsymbol{\beta} + \boldsymbol{z}_{it}^{2\top} \boldsymbol{\alpha}_i, \sigma^2)$,

(iii) $\boldsymbol{\beta}$ is a constant (fixed effect), $\boldsymbol{\alpha}_i \overset{\text{iid}}{\sim} \mathsf{N}(\boldsymbol{0}, Q)$ (random effect),

(iv) $\boldsymbol{Y}_i, \boldsymbol{\alpha}_i,\ i = 1, \ldots, n$ are mutually independent.

This model releases the assumption of independent observations Y_{it} by letting the regression coefficient $\boldsymbol{\alpha}_i$ be a random variable. The correlation between two measurements on one subject are $\mathrm{cov}(Y_{it+h}, Y_{it}) = \boldsymbol{z}_{it+h}^{2\top} Q \boldsymbol{z}_{it}^2$, so the dependence is the same for all times. More detailed description and more general definition can be found in Laird (1982).

Generalized linear mixed effect model (GLMM)

Generalized linear mixed effect model is specified by the following assumptions

(i) $g(\mathsf{E}[Y_{it}|\boldsymbol{\alpha}_i]) = \boldsymbol{z}_{it}^{1\top}\boldsymbol{\beta} + \boldsymbol{z}_{it}^{2\top}\boldsymbol{\alpha}_i$,

(ii) $Y_{it}|\boldsymbol{\beta}, \boldsymbol{\alpha}_i$ comes from an exponential family (see definiton 6),

(iii) $\boldsymbol{\beta}$ is a constant (fixed effect), $\boldsymbol{\alpha}_i \overset{\text{iid}}{\sim} \mathsf{N}(\mathbf{0}, Q)$,

(iv) $\boldsymbol{Y}_i, \boldsymbol{\alpha}_i,\ i = 1, \ldots, n$ are mutually independent.

This model is a mixture of GLM and LMM, conditionally on the random effect the model satisfies a GLM and in the case of normal distribution of Y_{it} and g being identity, it becomes LMM. If we set $Q = \mathbf{0}$ then $\boldsymbol{\alpha}_i = \mathbf{0}$ and if independence assumption holds then random effect model simplifies to LM or GLM.

Dynamic linear model (DLM)

Dynamic linear model is specified by the following assumptions

(i) $\mathsf{E}[Y_{it}|\boldsymbol{\alpha}_t] = \boldsymbol{z}_{it}^{\top}\boldsymbol{\alpha}_t$,

(ii) $Y_{it}|\boldsymbol{\alpha}_t \sim \mathsf{N}(\boldsymbol{z}_{it}^{\top}\boldsymbol{\alpha}_t, \sigma^2)$,

(iii) $\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1} \overset{\text{iid}}{\sim} \mathsf{N}(F_t\boldsymbol{\alpha}_{t-1}, Q_t)$, $\boldsymbol{\alpha}_0 \sim \mathsf{N}(\boldsymbol{a}_0, Q_0)$ (random process)

(iv) $\boldsymbol{Y}_t|\boldsymbol{\alpha}_t,\ t = 1, \ldots, T$ are independent.

This model assumes that random effect is a process that changes throughout the time. The model for random effect (iii) is called *Gaussian transition model*. The covariance between two measurements on one subject is $\operatorname{cov}(Y_{it+h}, Y_{it}) = \boldsymbol{z}_{t+h}^{\top}\operatorname{cov}(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t+h})\boldsymbol{z}_t = \boldsymbol{z}_{t+h}^{\top}F_{t+h}F_{t+h-1}\ldots F_tQ_t\boldsymbol{z}_t$, so the dependence may change over time. There are no restrictions on variance matrix Q_t, therefore, there might be fixed effects as in LMM or GLMM, it is sufficient to set variance of a certain component $\boldsymbol{\alpha}_t$ to zero.

Dynamic generalized linear model (DGLM)

Dynamic generalized linear model is specified by the following assumptions

(i) $g(\mathsf{E}[Y_{it}|\boldsymbol{\alpha}_t]) = \boldsymbol{z}_{it}^{\top}\boldsymbol{\alpha}_t$,

(ii) $Y_{it}|\boldsymbol{\alpha}_t$ comes from an exponential family (see definition 6),

(iii) $\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1} \overset{\text{iid}}{\sim} \mathsf{N}(F_t\boldsymbol{\alpha}_{t-1}, Q_t)$, $\boldsymbol{\alpha}_0 \sim \mathsf{N}(\boldsymbol{a}_0, Q_0)$ (random process),

(iv) $\boldsymbol{Y}_t|\boldsymbol{\alpha}_t,\ t = 1, \ldots, T$ are independent.

This model is a mixture of GLM and DLM. A special case of DGLM with g identity and normal distribution of Y_{it} is DLM. If we set $Q_t = \mathbf{0}$ and $F_t = I$ then we loose dynamic structure and dynamic models are simplified to LM or GLM.

LMM and GLMM are oriented on subjects rather than time dependency, those models do not need all observations to be observed at the same time points and the number of observations for each subject might differ. Dynamic models DLM and DGLM are, in contrast, based on time continuity and we assume all subjects to be measured at the same times (distinct time measurements can be handled but it complicates its structure).

Our aim is to model dynamic system of ice hockey matches where measurements are results of matches (win/draw/loss or even number of goals of both teams) and subjects are teams in a given season. Therefore, we describe statistical inference using dynamic systems in better detail.

The main difference between LM or GLM and other models is the concept of another random variable that enters the regression equation as a parameter. This approach is specific for Bayesian statistics, we briefly describe the concept and parameter estimation.

1.2 Bayesian inference

Suppose we observe random vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T$ with density $p(\boldsymbol{y}|\boldsymbol{\alpha})$. Unlike in common (frequentist) statistics we suppose that parameter $\boldsymbol{\alpha}$ is a random vector and for inference we use not only observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T$ but also the knowledge of distribution $\pi(\boldsymbol{\alpha})$ (called *prior* distribution). In practical situations both densities $p(\boldsymbol{y}|\boldsymbol{\alpha})$ and $\pi(\boldsymbol{\alpha})$ are assumed to be known. For inference about parameter $\boldsymbol{\alpha}$ we use Bayes theorem.

Theorem 1. *Suppose $\boldsymbol{\alpha}$ is a random vector with density $\pi(\boldsymbol{\alpha})$ with respect to σ-finite measure λ on $(\Theta, \mathbb{B}(\Theta))$ and $\boldsymbol{Y}_T^* = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_T^\top)^\top$ is a random vector with conditional density $p(\boldsymbol{y}|\boldsymbol{\alpha})$ for given $\boldsymbol{\alpha}$ with respect to σ-finite measure ν_T. Then for the conditional density $\pi(\boldsymbol{\alpha}|\boldsymbol{y})$ with given $\boldsymbol{Y}_T^* = \boldsymbol{y}$ holds*

$$\pi(\boldsymbol{\alpha}|\boldsymbol{y}) = \begin{cases} \frac{p(\boldsymbol{y}|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha})}{\int_\Theta p(\boldsymbol{y}|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha}) d\,\lambda(\boldsymbol{\alpha})}, & \int_\Theta p(\boldsymbol{y}|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha}) d\,\lambda(\boldsymbol{\alpha}) \neq 0, \\ 0, & else. \end{cases} \tag{1.1}$$

Proof. See Hušková (1985, page 11). □

The inference is based on posterior distribution $\pi(\boldsymbol{\alpha}|\boldsymbol{y})$. The idea is that our prior information of $\boldsymbol{\alpha}$ is refined by observations $\boldsymbol{y}$. The choice of prior distribution $\pi(\boldsymbol{\alpha})$ influences the resulted posterior distribution significantly, therefore it is either based on prior objective information (earlier study) or it should be vaguely informative (called *diffuse prior*). An extreme case is called *improper prior* when $\pi(\boldsymbol{\alpha})$ is a constant on the whole domain – it might not even be a density (for details of prior distributions see Hušková (1985, Chapter 2) or Robert (2007, Chapter 3).

There might be two aims of an analysis, to estimate the next outcome $\boldsymbol{Y}_{T+1}$ or estimate the parameter $\boldsymbol{\alpha}$. In both cases the estimates are based on posterior distribution given $\boldsymbol{y}$. For estimating the parameter $\boldsymbol{\alpha}$ we use (1.1) and for prediction the following equation.

$$p(\boldsymbol{y}_{T+1}|\boldsymbol{y}) = \int_\Theta p(\boldsymbol{y}_{T+1}, \boldsymbol{\alpha}|\boldsymbol{y}) d\,\lambda(\boldsymbol{\alpha}) = \int_\Theta p(\boldsymbol{y}_{T+1}|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha}|\boldsymbol{y}) d\,\lambda(\boldsymbol{\alpha}). \tag{1.2}$$

Notice that we have simplified $p(\boldsymbol{y}_{T+1}|\boldsymbol{\alpha}, \boldsymbol{y})$ to $p(\boldsymbol{y}_{T+1}|\boldsymbol{\alpha})$ as this is a characteristic of widely used hierarchical models. Some examples of hierarchical models are DLM or DGLM.

Loss functions and decision making

To receive an estimate of either $\boldsymbol{\alpha}$ or $\boldsymbol{y}_{T+1}$ we must define some criteria. This is done using loss functions as measures of distance of our estimate and the true value. In Bayesian statistics the concept of decision making unifies parameter estimations and hypothesis testing. Set of possible decisions is denoted as $\mathcal{D}$, in case of estimating the parameter $\boldsymbol{\alpha}$ the set $\mathcal{D} = \Theta$.

Definition 1. *A loss function is any function* $L : \Theta \times \mathcal{D} \to [0, +\infty)$.

Note. The loss function for the true value of a parameter $\boldsymbol{\alpha}$ and its estimate measures the error. An example of a loss function is a quadratic loss function

$$L_2(\boldsymbol{\alpha}, \widehat{\boldsymbol{\alpha}}) = (\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}})^\top (\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}). \tag{1.3}$$

The goal is to minimize the loss function, but it is typically impossible to minimize it uniformly. Therefore, we want to find a function δ that for realizations $\boldsymbol{y}$ gives the optimal estimate. Function δ is called an estimator while the value $\delta(\boldsymbol{y})$ is called an estimate of $\boldsymbol{\alpha}$. In Bayesian view the problem of minimizing $L(\boldsymbol{\alpha}, \delta(\boldsymbol{y}))$ is random in both $\delta(\boldsymbol{y})$ and $\boldsymbol{\alpha}$ but realizations $\boldsymbol{y}$ are supposed to be known. The minimization problem has an objective expected loss function, defined as

$$r(\pi, \delta) = \mathsf{E}\ L(\boldsymbol{\alpha}, \delta) = \int_\Theta \int_{\mathbb{R}^T} L(\boldsymbol{\alpha}, \delta(\boldsymbol{y})) p(\boldsymbol{y}|\boldsymbol{\alpha}) \pi(\boldsymbol{\alpha}) d\,\nu_T(\boldsymbol{y}) d\,\lambda(\boldsymbol{\alpha}). \tag{1.4}$$

The function δ is not necessary to be found for any realizations $\boldsymbol{y}$, so one may want to minimize only conditional expectation for given realizations

$$\rho(\pi, \widehat{\boldsymbol{\alpha}}) = \mathsf{E}\left[L(\boldsymbol{\alpha}, \widehat{\boldsymbol{\alpha}})|\boldsymbol{y}\right] = \int_\Theta L(\boldsymbol{\alpha}, \widehat{\boldsymbol{\alpha}}) \pi(\boldsymbol{\alpha}|\boldsymbol{y}) d\,\lambda(\boldsymbol{\alpha}). \tag{1.5}$$

Actually both concepts are equivalent as they lead to same decisions (see Robert, 2007, Theorem 2.3.2). By minimizing the formula (1.5) for different values of $\boldsymbol{y}$, we receive the optimal estimator δ. A Bayes estimator δ^* is a solution of minimizing the expression (1.4) with respect to admissible set of functions δ and the value $r(\pi, \delta^*)$ is called Bayes risk. We provide solution of a special case with the quadratic loss function.

Theorem 2. *The optimal Bayes estimator and its Bayes risk for the quadratic loss function* $L_2(\boldsymbol{\alpha}, \delta) = (\boldsymbol{\alpha} - \delta(\boldsymbol{y}))^\top (\boldsymbol{\alpha} - \delta(\boldsymbol{y}))$ *in terms of risk function* (1.4) *are*

$$\delta^*(\boldsymbol{y}) = \mathsf{E}\left[\boldsymbol{\alpha}|\boldsymbol{y}\right], \tag{1.6}$$

$$r(\pi, \delta^*) = \mathsf{E}\ \mathrm{var}(\boldsymbol{\alpha}|\boldsymbol{Y}). \tag{1.7}$$

Proof. Derivations can be found in Hušková (1985, Theorem 4.1.). □

This result is used in deriving the optimal estimator in DLM system via Kalman filter. It represents one of rare cases where the optimal estimator can be found analytically and even easily updated with future observations. Mostly, the computation of $\pi(\boldsymbol{\alpha}|\boldsymbol{y})$ is impossible. However, in most situations a sampling method that provides a sampling from a given distribution could be used. Traditional sampling method is Monte Carlo that provides a random sample by using hierarchical structure of given distribution. Nevertheless, in our setting it cannot be sampled from $\pi(\boldsymbol{\alpha}|\boldsymbol{y})$ directly. However, more general approach might be applied – creating a Markov chain with limiting distribution $\pi(\boldsymbol{\alpha}|\boldsymbol{y})$. This method is called *Markov chain Monte Carlo* (abbreviated as MCMC). Methods of sampling and treatment of MCMC are presented in Robert (2007, Chapter 6). This approach might be applied to receive smoothed estimates of team forms in a particular season in Extraliga. We implemented a hierarchical model in OpenBUGS 3.2.3 but observed poor convergence of MCMC and high dependency on initial values. For those reasons, we do not include this approach in the practical part.

1.3 Goodness of fit criteria

Every model encounters some difficulties due to some of the following reasons. It might assume some distributional form of given variable(s) and need to estimate parameters. Our choice of distribution is based on type of data that we observe but usually also on type of analysis that we are going to use. Small number of parameters do not provide flexible form and require that we are close to the true data generating process. On the other hand too many parameters can lead to overparametrization. If we use non-parametric model, there are usually less assumptions but less stability in the solution. Another typical assumption is some kind of independence structure that might not correspond with the reality.

To judge if our model corresponds with observed data, we need a criterion that would be used across all models and would indicate to which extent our model fits the data. In linear regression, the most used criterion is coefficient of determination R^2. Its definition and some of its properties can be found in either Anděl (2011), Zvára (2008) or Cipra (2008). It takes values between 0 and 1, higher value signalizing better fit of the model. Using OLS method for estimating parameters is formulated as minimization problem of a residual sum of squares but it is equivalent to maximization of R^2. For testing a submodel with only an intercept, we use F-test statistics which can also be rewritten in terms of R^2 (see Zvára, 2008, page 37). Its usage is so spread that it is usually given as a part of a summary after running linear regression. Throughout the thesis, we use R^2 for evaluating linear models and their generalizations.

For GLM models there is no ultimate goodness of fit criterion. Different types of criteria are used for different types of dependent variable. We focus on a multinomial explanatory variable, at first with two different values (having alternative distribution) and afterwards with ordinal (nominal) values. We define Gini coefficient and its generalization. They are introduced in Šimsa (2012, Chapter 2), which is the main reference of the following section.

Gini coefficient

Suppose Y_i has an alternative distribution with probability of being one equal to π_i, hence $Y_i \sim Alt(\pi_i)$, $i = 1, \ldots, n$. We let $\pi_i = \mathsf{P}(Y_i = 1)$ depend on index i and do not specify the dependency closer. The usual way for an estimation is based on the fact that alternative distribution belongs to an univariate exponential family (see section 3.1 and realize that alternative distribution is a special case of multinomial), therefore GLM theory can be applied.

Suppose $\widehat{\pi}_i$ are estimates of the probability that Y_i equals one. In a perfect case we would get that all estimated probabilities are higher for realized values one than for realized values zero, in such a case there would be a value which would serve as the threshold. Estimated probabilities smaller than this value would be assigned zero and bigger one. In practical situations, it is never such a perfect case but it is natural that the fewer interjections of estimated probabilities between groups with y values zero and one the better the model. Upon this principles is defined the Gini coefficient. Proper definition requires an introduction of terms *sensitivity* and *specificity*.

Definition 2. *Suppose $Y_i \sim Alt(\pi_i)$ and $\widehat{\pi}_i$ is an estimation of π_i, $i = 1, \ldots, n$ then*

$$h_1(c) = \mathsf{P}(\widehat{Y}_i(c) = 1 | Y_i = 1) \text{ is called sensitivity},$$
$$h_0(c) = \mathsf{P}(\widehat{Y}_i(c) = 0 | Y_i = 0) \text{ is called specificity}$$

where

$$\widehat{Y}_i(c) = \begin{cases} 1, & \widehat{\pi}_i > c \\ 0, & else \end{cases} \quad \text{and } c \in [0, 1]\,.$$

Note. The probability in definition of sensitivity is the probability that if I choose a unit out of observed variables with values one its estimated probability is greater than c and analogously for specificity. Therefore, those values are

$$h_1(c) = \frac{\#[\widehat{Y}_i(c) = 1 \cap Y_i = 1]}{\#Y_i = 1} = \frac{1}{n_1} \sum_{i, y_i = 1} \mathbb{1}_{[\widehat{\pi}_i > c]}, \tag{1.8}$$

$$h_0(c) = \frac{\#[\widehat{Y}_i(c) = 0 \cap Y_i = 0]}{\#Y_i = 0} = \frac{1}{n_0} \sum_{i, y_i = 0} \mathbb{1}_{[\widehat{\pi}_i \leq c]} \tag{1.9}$$

where $n_1 = \sum_{i=1}^{n} y_i$ and $n_0 = n - n_1$.

The sensitivity and specificity are piecewise linear, left-continuous, with values between 0 and 1. Sensitivity is non-increasing and specificity is non-decreasing. The higher is the value c the lower *sensitivity* and the higher *specificity*. Marginal cases are $h_1(1) = h_0(0) = 0$ and $h_1(0) = h_0(1) = 1$. The relation between *sensitivity* and *specificity* is usually graphically depicted using either receiver operating characteristic (ROC) or Lorenz curve. The first mentioned plots sensitivity versus one minus specificity (points $[1 - h_0(c), h_1(c)], c \in [0, 1]$) whereas the second one plots one minus sensitivity versus specificity (points $[h_0(c), 1 - h_1(c)], c \in [0, 1]$). It holds that both curves for better models are more bowed and closer to x and y axis.

Definition 3. *Gini coefficient for pairs* $(Y_i, \widehat{\pi}_i)^\top$, $i = 1, \ldots, n$ *is defined as*

$$G = 1 - 2 \int_0^1 (1 - h_1(c)) d\, h_0(c)$$

where $h_1(c)$ *and* $h_0(c)$ *are estimates of sensitivity and specificity as defined in equations* (1.8) *and* (1.9)*.*

Note. Gini coefficient is closely related to Lorenz curve. Gini coefficient equals:

$$G = 1 - 2AUC$$

where AUC is the area under the Lorenz curve which equals the integral in the definition.

Gini coefficient takes values from -1 to 1. Negative values correspond to reversed model and higher values signify better predictive ability of the model. G equals one if all estimated probabilities for cases where $y_i = 1$ are higher than estimated probabilities for $y_i = 0$. The more interjections between estimated probabilities there are the lower its value. This can be easily seen from the following theorem, which provides computationally more suitable formula.

Lemma 3. *Denote* $R_i, i = 1, \ldots, n$ *ranks for* $\widehat{\pi}_i, i = 1, \ldots, n$ *that are sorted (from lowest to biggest). Then*

$$G = \frac{S - S_0}{S_M - S_0} \tag{1.10}$$

where

$$\begin{aligned} S &= \sum_{i=1}^{n} R_i Y_i, \\ S_M &= \sum_{i=n_0+1}^{n} i = \frac{n_1}{2}(2n - n_1 + 1), \\ S_m &= \sum_{i=1}^{n_1} i = \frac{n_1}{2}(n_1 + 1), \\ S_0 &= \frac{S_M + S_m}{2}. \end{aligned}$$

Proof. The proof is rather technical, so we omit it. The proof in more general setting can be found in (Šimsa, 2012, pages 18-19). □

Note. In lemma 3 Gini coefficient is calculated via ranks. Because ranks for $\widehat{\pi}_i$ are the same as for $f(\widehat{\pi}_i)$ with a real, increasing function f, Gini coefficient remains the same as well. This property enables to calculate G without knowing the exact probability estimates $\widehat{\pi}_i$; it is sufficient to know their increasing transformation. This is usually done in scoring models, where transformations $s_i = f(\widehat{\pi}_i)$ are called scores.

Generalizations of Gini coefficient

In this section we present four generalizations of Gini coefficient for a multinomial variable Y_i, one in case of a nominal variable and three in case of an ordinal one. All of them keep the same values from -1 to 1 and coincide if the variable Y_i has alternative distribution and equal G. Other important properties are that they can be expressed as weighted average of partial Gini coefficients (specified on the following pages) and are invariant to increasing transformations of estimated probabilities or scores. For motivation and more detailed derivation see Šimsa (2012).

Gini coefficient for multinomial distribution

Let us suppose we observe a categorical variable Y_i with nominal values coded with numbers $\{0, 1, \ldots, r\}$. We denote the probability of belonging to a group q as $\pi_i^q = P(Y_i = q)$. The output of a non-specified multinomial model are estimated probabilities $\widehat{\boldsymbol{\pi}}_i = (\widehat{\pi}_i^0, \widehat{\pi}_i^1, \ldots, \widehat{\pi}_i^r)^\top$.

The idea for the following definition is that if we observe $Y_i = q$ and $Y_j \neq q$ then the probability π_i^q should (in most cases) be higher than π_j^q. This is a generalization of the same idea as for G, for which we get value one if all probabilities $\widehat{\pi}_i$ for realizations $Y_i = 1$ are higher than for realizations $Y_j = 0$. In nominal concept, we get the value one if the same principle holds for all values $q = 0, \ldots, r$ and its estimated probabilities.

Definition 4. *Suppose $(Y_i, \widehat{\boldsymbol{\pi}}_i)$ are pairs of nominal variables and estimated probabilities of belonging to a certain category. Then*

$$G_n = \frac{1}{n} \sum_{q=0}^{r} \sum_{i, Y_i = q} \sum_{j, Y_j \neq q} \frac{sgn(\widehat{\pi}_i^q - \widehat{\pi}_j^q)}{n - n_q} \tag{1.11}$$

where $n_q = \sum_{i=1}^{n} \mathbb{1}_{[Y_i = q]}$. G_n is called Gini coefficient for a nominal variable.

Another useful formula is that G_n can be expressed in terms of partial Gini coefficients. A categorical variable Y_i carries the same information as a set of alternative variables $\widetilde{Y}_i^1, \ldots, \widetilde{Y}_i^r$ where $\widetilde{Y}_i^q = \mathbb{1}_{[Y_i = q]}$. For every set of pairs $(\widetilde{Y}_i^q, \widehat{\pi}_i^q)^\top$, $i = 1, \ldots, n$ we can calculate G according to definition 3. Connection between G_n and those partial Gini coefficients give the following theorem.

Lemma 4. *Suppose $(Y_i, \widehat{\boldsymbol{\pi}}_i)$, $i = 1, \ldots, n$ are pairs of nominal variables with categories $0, \ldots, r$ and estimated probabilities. Then*

$$G_n = \frac{1}{n} \sum_{q=0}^{r} n_q G_q \tag{1.12}$$

where G_q is a partial Gini coefficient, i.e. Gini coefficient for pairs $(\widetilde{Y}_i^q, \widehat{\pi}_i^q)^\top$.

Proof. Straightforward derivation (see Šimsa, 2012, page 24). □

We see that the Gini coefficient for a nominal variable is a weighted average of partial Gini coefficients.

Gini coefficient for ordinal categorical variable

Suppose the same setting as in the preceding section but with ordinal categories of Y_i. Further we suppose that there exists a common linear predictor enabling to define scores. Scores accumulate the information of estimated probabilities such that higher values of scores signify expected higher category of Y_i. One type of model that enables this is the proportional odds model (see Cipra, 2008, pages 171-175).

Our data consist of pairs $(Y_i, s_i)^\top$, $i = 1, \ldots, n$ where s_i denotes score for a variable Y_i (e.g. $s_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$). In a perfect setting if we sort pairs (Y_i, s_i) by s_i, values of Y_i will be also sorted. Every violation of this principle will lead to a decrease of the ability to diversify between values Y_i. The total number of those violations can be counted as a sum of violations between each pairs of categories. This concept unifies all generalizations of the Gini coefficient for an ordinal categorical variable and enables their common definition.

Definition 5. *Let us suppose $(Y_i, s_i)^\top$, $i = 1, \ldots, n$ are pairs of ordinal variables with categories $0, \ldots, r$ and their corresponding scores. We define following measures of goodness of fit*

$$G_1 = \frac{1}{C_1} \sum_{k=1}^{r} \sum_{q=0}^{k-1} n_k n_q (k - q) G_{q,k}, \tag{1.13}$$

$$G_2 = \frac{1}{C_2} \sum_{k=1}^{r} \sum_{q=0}^{k-1} n_k n_q (U_k - U_q) G_{q,k}, \tag{1.14}$$

$$G_C = \frac{1}{C_3} \sum_{k=1}^{r} \sum_{q=0}^{k-1} n_k n_q G_{q,k} \tag{1.15}$$

where $G_{q,k}$ is partial Gini coefficient, i.e. Gini coefficient for pairs $(\overline{Y}_i, s_i)^\top$,

$$\overline{Y}_i = \begin{cases} 0, & \text{if } Y_i = q, \\ 1, & \text{if } Y_i = k, \end{cases} \quad i \in \{l : Y_l \in \{q, k\}\},$$

variables $U_k = \sum_{q=0}^{k-1} n_q + \frac{1}{2}(n_k + 1)$ and C_i are normalizing constants such that $G_1, G_2, G_C \in [-1, 1]$.

Note. Normalizing constants are

$$C_1 = \sum_{k=1}^{r} \sum_{q=0}^{k-1} n_k n_q (k - q), \qquad C_2 = \sum_{k=1}^{r} \sum_{q=0}^{k-1} n_k n_q (U_k - U_q), \qquad C_3 = \sum_{k=1}^{r} \sum_{q=0}^{k-1} n_k n_q.$$

The first two coefficients G_1 and G_2 are motivated by generalization of a definition of the Lorenz curve. G_1 is based on values of Y_i whereas G_2 is based on ranks (U_k) of those values. Same values of Y_i have the same ranks. Both of them can be computed in an alternative way similarly as in the lemma 3. In case that number of observations in each category is the same then $G_1 = G_2$. Definition of G_C generalizes the relationship of Gini coefficient and C-statistics. For more detail see Šimsa (2012).

We see that measures G_1 and G_2 have bigger weights for partial Gini coefficients between more distant groups whereas G_C treats them only based on the number of observations within each group.

For a comparison of different models for an ordinal variable, it is recommended to use measures G_1, G_2 and G_C because they take the ordinal scale into account. However, for prediction purposes it is more advisable to use G_n even for an ordinal variable. If we overestimate probability of win, we suffer the same loss if the realized outcome was draw or loss.

Chapter 2

Dynamic linear model

This section is based on Anderson and Moore (1979), Welch and Bishop (2006), Fahrmeir and Tutz (1994) and Cipra (2008).

A dynamic linear model or a state space linear model was briefly presented in section 1.1. The main specific is that time series observations $\boldsymbol{y}_t$[1] are related to an unobserved *state vector* $\boldsymbol{\alpha}_t$[2], which carries the dependence between observations at different time points and both observations and states are normally distributed.

We suppose that the current value of the state vector is determined by the state vector in the previous time $\boldsymbol{\alpha}_{t-1}$ and some random effect $\boldsymbol{w}_t$. Moreover, we suppose that the system is *linear*, meaning that $\boldsymbol{\alpha}_t$ depends on stated variables linearly.

As said before, we do not observe values of the state vector, but only of some other variable – *measurements* $\boldsymbol{y}_t$, which depend linearly on $\boldsymbol{\alpha}_t$ and on a random effect $\boldsymbol{v}_t$. The system is defined as follows

$$\boldsymbol{y}_t = Z_t\boldsymbol{\alpha}_t + \boldsymbol{v}_t, \tag{2.1}$$
$$\boldsymbol{\alpha}_t = F_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{w}_t, \tag{2.2}$$

$$\boldsymbol{\alpha}_t \in \mathbb{R}^n,\ \boldsymbol{y}_t \in \mathbb{R}^m,\ F_t \in \mathbb{R}^{n\times n},\ Z_t \in \mathbb{R}^{m\times n}. \tag{2.3}$$

Further, the random effects with multivariate normal distributions and independent of each other

$$\boldsymbol{\alpha}_0 \sim \mathsf{N}(\boldsymbol{a}_0, Q_0),\ \boldsymbol{v}_t \overset{\text{iid}}{\sim} \mathsf{N}(0, R_t),\ \boldsymbol{w}_t \overset{\text{iid}}{\sim} \mathsf{N}(0, Q_t), \boldsymbol{\alpha}_0 \perp \boldsymbol{w}_t \perp \boldsymbol{v}_s\, \forall t, s \in \mathbb{N}. \tag{2.4}$$

$$\boldsymbol{a}_0 \in \mathbb{R}^n,\ \boldsymbol{w}_t \in \mathbb{R}^n,\ Q_t \in \mathbb{R}^{n\times n},\ \boldsymbol{v}_t \in \mathbb{R}^m,\ R_t \in \mathbb{R}^{m\times m}.$$

The matrix F_t relates the state at the current time step t to the next state and the matrix Z_t relates the state $\boldsymbol{\alpha}_t$ to the measurement $\boldsymbol{y}_t$. It takes role as a regression matrix and may depend on covariates or past measurements.

[1] In literature about Kalman filter are random variables denoted with small letters (see any reference above), even though it confuses the difference between a random variable and its realization we keep this traditional notation.

[2] More typical notation in presenting the Kalman filter is using $\boldsymbol{x}_t$ for a state variable. However, we want to stress the similarity with regression models and its role as regression parameters.

In the simple case all system matrices Z_t, F_t, R_t, Q_t and initial values $\boldsymbol{a}_0, Q_0$ are assumed to be deterministic and known at the given time period t. However, in a practical situation they might depend on unknown hyperparameters $\boldsymbol{\theta}$, such that

$$Z_t = Z_t(\boldsymbol{\theta}), F_t = F_t(\boldsymbol{\theta}), R_t = R_t(\boldsymbol{\theta}), Q_t = Q_t(\boldsymbol{\theta}), \boldsymbol{a}_0 = \boldsymbol{a}_0(\boldsymbol{\theta}), Q_0 = Q_0(\boldsymbol{\theta}). \quad (2.5)$$

It should be remarked that covariance matrices are allowed to be singular. This enables to include constant parameters into a state vector $\boldsymbol{\alpha}_t$.

For our application, dynamic linear system is composed of measurements represented by the results of ice hockey matches and the state vector is represented by forms of teams, it might include some fixed parameters as home advantage. Variances of matches and forms of teams will be unknown and intended to be estimated. However, in the following we assume deterministic matrices; the estimation of hyperparameters is presented in section 2.2.

2.1 Kalman filter

The aim of Kalman filter is to provide estimates of state vector $\boldsymbol{\alpha}_t$ in a dynamic linear model. As our state estimate is a random variable, the approach is based on the posterior distribution as in Bayesian statistics (see section 1.2).

Throughout the text by the symbol $\mathcal{L}(\boldsymbol{\alpha}|\boldsymbol{y})$ we understand the conditional distribution of $\boldsymbol{\alpha}$ given realized values of random vector $\boldsymbol{y}$ unless it is specified differently, i.e. a conditional mean $\mathsf{E}[\boldsymbol{\alpha}|\boldsymbol{y}]$ is understood as a deterministic value and not as a random vector. Past values of random vectors $\boldsymbol{y}_t$ up to a time t are denoted by

$$\boldsymbol{y}_t^* = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_t^\top)^\top.$$

A random variable $\boldsymbol{\alpha}_t|\boldsymbol{y}_s^*$ will be denoted by $\boldsymbol{\alpha}_{t|s}$ as is common in time series for filtered values using history of the observed variable.

In praxis two main tasks are to be performed. We want either to estimate the current state $\boldsymbol{\alpha}_t$ or predict its future value having all the information up to time t. Therefore, the tasks are

- to use only information about measurements prior to time t, $\boldsymbol{y}_{t-1}^*$, so we want to estimate $\boldsymbol{\alpha}_{t|t-1}$ and $\boldsymbol{y}_{t|t-1}$.
- to use information of measurements including the current measurement, $\boldsymbol{y}_t^*$, which means to estimate $\boldsymbol{\alpha}_{t|t}$.

As stressed above, the first one can be used for predicting the value of $\boldsymbol{y}_t$ using the estimated value of the state vector $\boldsymbol{\alpha}_{t|t-1}$, which will be our main focus. The second one is better estimate of the current value of the state vector $\boldsymbol{\alpha}_t$ since it uses more measurements.

In the following, we want to find an optimal Bayes estimator for $\boldsymbol{\alpha}_t$ in terms of a quadratic loss function (1.3) that could be computed iteratively accounting for new observations. At a given time the solution is conditional mean (see theorem 2). We define $\widehat{\boldsymbol{\alpha}}_{t|t-1}$ to be our *a priori* state estimate at the time t

given the information $\boldsymbol{y}^*_{t-1}$ and $\widehat{\boldsymbol{\alpha}}_{t|t}$ to be our *a posteriori* state estimate given the information $\boldsymbol{y}^*_t$. So the optimal estimates are

$$\widehat{\boldsymbol{\alpha}}_{t|t-1} = \mathsf{E}\left[\boldsymbol{\alpha}_t|\boldsymbol{y}^*_{t-1}\right] = \mathsf{E}\ \boldsymbol{\alpha}_{t|t-1},$$
$$\widehat{\boldsymbol{\alpha}}_{t|t} = \mathsf{E}\left[\boldsymbol{\alpha}_t|\boldsymbol{y}^*_t\right] = \mathsf{E}\ \boldsymbol{\alpha}_{t|t}.$$

Further we denote $P_{t|t-1}$ the variance matrix of the posterior distribution given information $\boldsymbol{y}^*_{t-1}$ and $P_{t|t}$ to be the variance matrix of the posterior distribution given information $\boldsymbol{y}^*_t$. Variances are

$$P_{t|t-1} = \mathrm{var}[\boldsymbol{\alpha}_t|\boldsymbol{y}^*_{t-1}] = \mathrm{var}(\boldsymbol{\alpha}_{t|t-1}),$$
$$P_{t|t} = \mathrm{var}[\boldsymbol{\alpha}_t|\boldsymbol{y}^*_t] = \mathrm{var}(\boldsymbol{\alpha}_{t|t}).$$

We present one rather technical theorem, which will be useful for later derivations.

Theorem 5. *Suppose* $\boldsymbol{W} = (\boldsymbol{Y}^\top, \boldsymbol{X}^\top)^\top$ *has a regular normal distribution* $\mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Then the conditional distribution of* $\boldsymbol{Y}$ *with given* $\boldsymbol{X} = \boldsymbol{x}$ *is*

$$\mathsf{N}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}_2\right), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}),$$

where $\mathsf{E}\,\boldsymbol{Y} = \boldsymbol{\mu}_1, \mathsf{E}\,\boldsymbol{X} = \boldsymbol{\mu}_2, \mathrm{var}(\boldsymbol{Y}) = \boldsymbol{\Sigma}_{11}, \mathrm{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{22}, \mathrm{cov}(\boldsymbol{Y}, \boldsymbol{X}) = \boldsymbol{\Sigma}_{12}$.

Proof. The proof can be found in Anděl (2011, Theorem 4.12). □

Note. The preceding theorem assumes a regular joint distribution but this assumption can be excluded; pseudo-inverse would replace the inverse (see Anderson and Moore, 1979, page 39).

Assuming the linear system for a non-observable state vector $\boldsymbol{\alpha}_t$ (equation (2.2)) results in inheriting the normal distribution from random effects $\boldsymbol{w}_t$, so the prior distribution of $\boldsymbol{\alpha}_t$ is normal. Our goal is to determine the conditional distribution of the state $\boldsymbol{\alpha}_t$ given measurements $\boldsymbol{y}^*_t$. By assuming the normal distribution of random effects $\boldsymbol{v}_t$ in equation (2.1), we receive that the conditional distribution is also normal, which is the content of the next theorem.

However, for now we do not calculate either the mean or the variance of $\boldsymbol{\alpha}_{t|t}$ (it is a result of the theorem 9) but only the mean and the variance for $\boldsymbol{\alpha}_{t|t-1}$ as functions of corresponding characteristics of $\boldsymbol{\alpha}_{t-1|t-1}$.

Theorem 6. *Suppose the setting given by equations* (2.1) *and* (2.2) *and by the distributional conditions* (2.3).

Then distributions of random vectors $\boldsymbol{\alpha}_{t|t}$, $\boldsymbol{\alpha}_{t|t-1}$ *and* $\boldsymbol{y}_{t|t-1}$ *are*

$$\boldsymbol{\alpha}_{t|t} \sim \mathsf{N}(\widehat{\boldsymbol{\alpha}}_{t|t}, P_{t|t}),$$
$$\boldsymbol{\alpha}_{t|t-1} \sim \mathsf{N}(\widehat{\boldsymbol{\alpha}}_{t|t-1}, P_{t|t-1}),$$
$$\boldsymbol{y}_{t|t-1} \sim \mathsf{N}(\widehat{\boldsymbol{y}}_{t|t-1}, Z_t P_{t|t-1} Z_t^\top + R_t)$$

where $\widehat{\boldsymbol{\alpha}}_{t|t-1} = F_t\widehat{\boldsymbol{\alpha}}_{t-1|t-1}, P_{t|t-1} = F_t P_{t-1|t-1} F_t^\top + Q_t$ *and* $\widehat{\boldsymbol{y}}_{t|t-1} = Z_t\widehat{\boldsymbol{\alpha}}_{t|t-1}$.

Proof. Firstly, we show that $\boldsymbol{\alpha}_{t|t}$, $\boldsymbol{\alpha}_{t|t-1}$ and $\boldsymbol{y}_{t|t-1}$ are normally distributed. We know that $\boldsymbol{\alpha}_1$ is normally distributed since it is a sum of normally distributed vectors $\boldsymbol{\alpha}_0$ and $\boldsymbol{w}_0$. Further $\boldsymbol{\alpha}_t$ and $\boldsymbol{y}_t$ have normal distributions because they are linear combinations of two normal distributions. Theorem 5 proves the normality. Because we denoted $\widehat{\boldsymbol{\alpha}}_{t|t} = \mathsf{E}\ \boldsymbol{\alpha}_{t|t}$, we just define $P_{t|t} = \operatorname{var}(\boldsymbol{\alpha}_{t|t})$ and the first part is done.

We calculate the mean and the variance of $\boldsymbol{\alpha}_{t|t-1}$ using the fact that $\boldsymbol{w}_t$ and $\boldsymbol{y}^*_{t-1}$ are independent.

$$\widehat{\boldsymbol{\alpha}}_{t|t-1} = \mathsf{E}\left[\boldsymbol{\alpha}_t|\boldsymbol{y}^*_{t-1}\right] = \mathsf{E}\left[F_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{w}_t|\boldsymbol{y}^*_{t-1}\right] = F_t\widehat{\boldsymbol{\alpha}}_{t-1|t-1} + \mathsf{E}\ \boldsymbol{w}_t = F_t\widehat{\boldsymbol{\alpha}}_{t-1|t-1},$$
$$P_{t|t-1} = \operatorname{var}[\boldsymbol{\alpha}_t|\boldsymbol{y}^*_{t-1}] = \operatorname{var}[F_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{w}_t|\boldsymbol{y}^*_{t-1}] = F_tP_{t-1|t-1}F_t^\top + Q_t.$$

Similarly we calculate the mean and the variance of $\boldsymbol{y}_{t|t-1}$

$$\widehat{\boldsymbol{y}}_{t|t-1} = \mathsf{E}\left[\boldsymbol{y}_t|\boldsymbol{y}^*_{t-1}\right] = \mathsf{E}\left[Z_t\boldsymbol{\alpha}_t + \boldsymbol{v}_t|\boldsymbol{y}^*_{t-1}\right] = Z_t\widehat{\boldsymbol{\alpha}}_{t|t-1} + \mathsf{E}\ \boldsymbol{v}_t = Z_t\widehat{\boldsymbol{\alpha}}_{t|t-1},$$
$$\operatorname{var}[\boldsymbol{y}_t|\boldsymbol{y}^*_{t-1}] = \operatorname{var}[Z_t\boldsymbol{\alpha}_t + \boldsymbol{v}_t|\boldsymbol{y}^*_{t-1}] = Z_tP_{t|t-1}Z_t^\top + R_t.$$

□

The preceding theorem shows not only how to estimate the current state using the history of measurements $\boldsymbol{y}^*_{t-1}$ but also how to estimate the following measurement itself with the knowledge of $\boldsymbol{y}^*_{t-1}$, which is the main focus in our case – we are more interested in the prediction of the outcome of the next game rather than an estimate of teams forms.

The following lemma is only a supporting proposition of inverting block symmetrical matrices.

Lemma 7. *Suppose that* $\boldsymbol{K} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{D} \end{pmatrix}$ *is a positive definite matrix and blocks* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *are square matrices. Then*

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{D} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{Q}^{-1} & -\boldsymbol{Q}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{B}^\top\boldsymbol{Q}^{-1} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{B}^\top\boldsymbol{Q}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \end{pmatrix}$$

where $\boldsymbol{Q} = \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{B}^\top$

Proof. The proof can be found in Anděl (2011, Theorem A.10) □

In theorem 6 we have shown how to predict the next state and the next observation knowing the estimate of the previous current state. It was quite straightforward, only calculating its moments and using distributional assumptions. The more complicated problem is to derive the distribution of $\boldsymbol{\alpha}_{t|t}$. It is not clear if using a new measurement can be absorbed without the need of deriving the distribution from the beginning. Following theorem shows that in case of normally distributed errors it can be done.

This theorem is not included in the usual books used as references of the Kalman filter (not even in Anderson and Moore (1979)) but only verbally formulated. Therefore, despite its clumsy formulation, we provide it in better detail.

Theorem 8. *Suppose $\boldsymbol{X}$ is normally distributed and $\boldsymbol{W} = (\boldsymbol{Y}^\top, \boldsymbol{Z}^\top)$ has a joint regular normal distribution. Then*

$$\mathcal{L}\left(\boldsymbol{X}|(\boldsymbol{Y}^\top, \boldsymbol{Z}^\top)^\top = (\boldsymbol{y}^\top, \boldsymbol{z}^\top)^\top\right) \;=\; \mathcal{L}\left((\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y})|(\boldsymbol{Z}|\boldsymbol{Y} = \boldsymbol{y}) = \boldsymbol{z}\right).$$

To simplify the notation we denote $\boldsymbol{X}(\boldsymbol{y}) = \boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}$, so the theorem states that

$$\mathcal{L}\left(\boldsymbol{X}(\boldsymbol{y}, \boldsymbol{z})\right) = \mathcal{L}\left(\boldsymbol{X}(\boldsymbol{y})|\boldsymbol{Z}(\boldsymbol{y}) = \boldsymbol{z}\right).$$

Proof. Using the theorem 5 gives that $\boldsymbol{X}(\boldsymbol{y}, \boldsymbol{z})$ and $\boldsymbol{X}(\boldsymbol{y})|\boldsymbol{Z}(\boldsymbol{y}) = \boldsymbol{z}$ are normally distributed. In the following we show that means and variances of both conditional distributions are the same, which implies the same distribution as the normal distribution is determined by mean and variance.

We will prove only that variances are the same, for means we would use the same steps. Theorem 5 gives following

$$V_{\boldsymbol{X}(\boldsymbol{w})} = \operatorname{var}[\boldsymbol{X}|\boldsymbol{W} = \boldsymbol{w}] = V_{\boldsymbol{X}} - V_{\boldsymbol{X}\boldsymbol{W}} V_{\boldsymbol{W}}^{-1} V_{\boldsymbol{W}\boldsymbol{X}}.$$

We need to determine the inverse of $V_{\boldsymbol{W}}$. We know that it is a regular positive definite matrix with squared blocks $V_{\boldsymbol{Y}}$ and $V_{\boldsymbol{Z}}$; therefore, we can use the lemma 7.

$$\begin{aligned} V_{\boldsymbol{W}}^{-1} &= \begin{pmatrix} V_{\boldsymbol{Y}} & V_{\boldsymbol{Y}\boldsymbol{Z}} \\ V_{\boldsymbol{Z}\boldsymbol{Y}} & V_{\boldsymbol{Z}} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} V_{\boldsymbol{Y}(\boldsymbol{z})}^{-1} & -V_{\boldsymbol{Y}(\boldsymbol{z})}^{-1} V_{\boldsymbol{Y}\boldsymbol{Z}} V_{\boldsymbol{Z}}^{-1} \\ -V_{\boldsymbol{Z}}^{-1} V_{\boldsymbol{Z}\boldsymbol{Y}} V_{\boldsymbol{Y}(\boldsymbol{z})}^{-1} & V_{\boldsymbol{Z}}^{-1} + V_{\boldsymbol{Z}}^{-1} V_{\boldsymbol{Z}\boldsymbol{Y}} V_{\boldsymbol{Y}(\boldsymbol{z})}^{-1} V_{\boldsymbol{Y}\boldsymbol{Z}} V_{\boldsymbol{Z}}^{-1} \end{pmatrix} \end{aligned}$$

We multiply the following result from left by $V_{\boldsymbol{X}\boldsymbol{W}}$ and after some algebra we receive:

$$V_{\boldsymbol{X}\boldsymbol{W}} V_{\boldsymbol{W}}^{-1} = \begin{pmatrix} V_{\boldsymbol{X}(\boldsymbol{z})\boldsymbol{Y}(\boldsymbol{z})} V_{\boldsymbol{Y}(\boldsymbol{z})}^{-1} \\ V_{\boldsymbol{X}\boldsymbol{Z}} V_{\boldsymbol{Z}}^{-1} - V_{\boldsymbol{X}(\boldsymbol{z})\boldsymbol{Y}(\boldsymbol{z})} V_{\boldsymbol{Y}(\boldsymbol{z})}^{-1} V_{\boldsymbol{Y}\boldsymbol{Z}} V_{\boldsymbol{Z}}^{-1} \end{pmatrix}^\top \tag{2.6}$$

So far we have shown that

$$V_{\boldsymbol{X}(\boldsymbol{w})} = V_{\boldsymbol{X}} - \begin{pmatrix} V_{\boldsymbol{X}(\boldsymbol{z})\boldsymbol{Y}(\boldsymbol{z})} V_{\boldsymbol{Y}(\boldsymbol{z})}^{-1} \\ V_{\boldsymbol{X}\boldsymbol{Z}} V_{\boldsymbol{Z}}^{-1} - V_{\boldsymbol{X}(\boldsymbol{z})\boldsymbol{Y}(\boldsymbol{z})} V_{\boldsymbol{Y}(\boldsymbol{z})}^{-1} V_{\boldsymbol{Y}\boldsymbol{Z}} V_{\boldsymbol{Z}}^{-1} \end{pmatrix}^\top V_{\boldsymbol{W}\boldsymbol{X}}.$$

Now we focus on the second variance

$$\begin{aligned} \operatorname{var}[\boldsymbol{X}(\boldsymbol{z})|\boldsymbol{Y}(\boldsymbol{z}) = \boldsymbol{y})] &= V_{\boldsymbol{X}(\boldsymbol{z})} - V_{\boldsymbol{X}(\boldsymbol{z})\boldsymbol{Y}(\boldsymbol{z})} V_{\boldsymbol{Y}(\boldsymbol{z})}^{-1} V_{\boldsymbol{Y}(\boldsymbol{z})\boldsymbol{X}(\boldsymbol{z})} \\ &= V_{\boldsymbol{X}} - V_{\boldsymbol{X}\boldsymbol{Z}} V_{\boldsymbol{Z}}^{-1} V_{\boldsymbol{Z}\boldsymbol{X}} - V_{\boldsymbol{X}(\boldsymbol{z})\boldsymbol{Y}(\boldsymbol{z})} V_{\boldsymbol{Y}(\boldsymbol{z})}^{-1} (V_{\boldsymbol{Y}\boldsymbol{X}} - V_{\boldsymbol{Y}\boldsymbol{Z}} V_{\boldsymbol{Z}}^{-1} V_{\boldsymbol{Z}\boldsymbol{X}}) \\ &= V_{\boldsymbol{X}} - \begin{pmatrix} V_{\boldsymbol{X}(\boldsymbol{z})\boldsymbol{Y}(\boldsymbol{z})} V_{\boldsymbol{Y}(\boldsymbol{z})}^{-1} \\ V_{\boldsymbol{X}\boldsymbol{Z}} V_{\boldsymbol{Z}}^{-1} - V_{\boldsymbol{X}(\boldsymbol{z})\boldsymbol{Y}(\boldsymbol{z})} V_{\boldsymbol{Y}(\boldsymbol{z})}^{-1} V_{\boldsymbol{Y}\boldsymbol{Z}} V_{\boldsymbol{Z}}^{-1} \end{pmatrix}^\top \begin{pmatrix} V_{\boldsymbol{Y}\boldsymbol{X}} \\ V_{\boldsymbol{Z}\boldsymbol{X}} \end{pmatrix} \\ &= V_{\boldsymbol{X}} - \begin{pmatrix} V_{\boldsymbol{X}(\boldsymbol{z})\boldsymbol{Y}(\boldsymbol{z})} V_{\boldsymbol{Y}(\boldsymbol{z})}^{-1} \\ V_{\boldsymbol{X}\boldsymbol{Z}} V_{\boldsymbol{Z}}^{-1} - V_{\boldsymbol{X}(\boldsymbol{z})\boldsymbol{Y}(\boldsymbol{z})} V_{\boldsymbol{Y}(\boldsymbol{z})}^{-1} V_{\boldsymbol{Y}\boldsymbol{Z}} V_{\boldsymbol{Z}}^{-1} \end{pmatrix}^\top V_{\boldsymbol{W}\boldsymbol{X}}. \end{aligned}$$

We see above that $V_{\boldsymbol{X}(\boldsymbol{w})} = \operatorname{var}[\boldsymbol{X}(\boldsymbol{z})|\boldsymbol{Y}(\boldsymbol{z}) = \boldsymbol{y})]$. □

The previous theorem ensures that the distribution of $\boldsymbol{\alpha}_{t|t}$ can be obtained using $\boldsymbol{\alpha}_{t|t-1}$ and the measurement $\boldsymbol{y}_t$. We have prepared all necessary analytical apparatus to formulate update equations.

Theorem 9. *Suppose the setting given by equations* (2.1) *and* (2.2) *and by the distributional conditions* (2.3).

Mean and variance of $\boldsymbol{\alpha}_t|\boldsymbol{y}_t^*$ *are*

$$\widehat{\boldsymbol{\alpha}}_{t|t} = \widehat{\boldsymbol{\alpha}}_{t|t-1} + K_t(\boldsymbol{y}_t - Z_t\widehat{\boldsymbol{\alpha}}_{t|t-1}), \tag{2.7}$$
$$P_{t|t} = (I - K_t Z_t)P_{t|t-1} \tag{2.8}$$

where $K_t = P_{t|t-1}Z_t^\top(Z_t P_{t|t-1}Z_t^\top + R_t)^{-1}$.

Proof. Using the theorem 8 gives

$$\mathcal{L}(\boldsymbol{\alpha}_{t|t}) = \mathcal{L}(\boldsymbol{\alpha}_{t|t-1}|\boldsymbol{y}_{t|t-1} = \boldsymbol{y}_t).$$

Further we use the theorems 6 and 5

$$\mathcal{L}(\boldsymbol{\alpha}_{t|t-1}|\boldsymbol{y}_{t|t-1} = \boldsymbol{y}_t) = \mathsf{N}(\widehat{\boldsymbol{\alpha}}_{t|t-1} + K_t(\boldsymbol{y}_t - \widehat{\boldsymbol{y}}_{t|t-1}), P_{t|t-1} - K_t \operatorname{cov}(\boldsymbol{y}_{t|t-1}, \boldsymbol{\alpha}_{t|t-1})),$$

where $K_t = \operatorname{cov}(\boldsymbol{\alpha}_{t|t-1}, \boldsymbol{y}_{t|t-1}) \operatorname{var}(\boldsymbol{y}_{t|t-1})^{-1}$. Now we calculate the covariance using definition of a measurement $\boldsymbol{y}_t = Z_t\boldsymbol{\alpha}_t + \boldsymbol{v}_t$ and independence of $\boldsymbol{v}_t$ and $\boldsymbol{y}_{t-1}^*$ and independence of $\boldsymbol{v}_t$ and $\boldsymbol{\alpha}_t$, so

$$\operatorname{cov}(\boldsymbol{\alpha}_{t|t-1}, \boldsymbol{y}_{t|t-1}) = \operatorname{cov}(\boldsymbol{\alpha}_{t|t-1}, Z_t\boldsymbol{\alpha}_{t|t-1}) + \operatorname{cov}(\boldsymbol{\alpha}_{t|t-1}, \boldsymbol{v}_t) = P_{t|t-1}Z_t^\top.$$

We plug this into the formulas for K_t and $P_{t|t}$ with the variance of $\boldsymbol{y}_{t|t-1}$ calculated in the theorem 6 to get desired forms.

$$K_t = \operatorname{cov}(\boldsymbol{\alpha}_{t|t-1}, \boldsymbol{y}_{t|t-1}) \operatorname{var}(\boldsymbol{y}_{t|t-1})^{-1} = P_{t|t-1}Z_t^\top(Z_t P_{t|t-1}Z_t^\top + R_t)^{-1},$$
$$P_{t|t} = P_{t|t-1} - K_t Z_t P_{t|t-1} = (I - K_t Z_t)P_{t|t-1}.$$

□

Note. Notice that the variable $\boldsymbol{y}_t$ is used as a realized observation but in the term $\boldsymbol{y}_{t|t-1}$ as a random variable.

The matrix $K_t \in \mathbb{R}^{n\times m}$ is called *gain* and is chosen such that it minimizes the a posteriori covariance error, because we could have formulated the definition of the optimal estimator in terms of the error $\boldsymbol{e}_t = \boldsymbol{\alpha}_t - \widehat{\boldsymbol{\alpha}}_{t|t}$ with the following property

$$\operatorname{var}(\boldsymbol{e}_t|\boldsymbol{y}_t^*) = \min_f \operatorname{var}(\boldsymbol{\alpha}_t - f(\boldsymbol{y}_t^*)|\boldsymbol{y}_t^*).$$

In case of our optimal estimator $\widehat{\boldsymbol{\alpha}}_{t|t}$ we have

$$\mathsf{E}\left[\boldsymbol{e}_t|\boldsymbol{y}_t^*\right] = \mathbf{0},$$
$$\operatorname{var}(\boldsymbol{e}_t|\boldsymbol{y}_t^*) = \mathsf{E}\left[\boldsymbol{e}_t\boldsymbol{e}_t^\top|\boldsymbol{y}_t^*\right] = P_{t|t}.$$

So, the mean of error conditioned on given information is zero (we do not over- or underestimate the state in the mean) and the conditional variance is minimal in the sense of positive definite matrices (the difference between any other variance of the estimate of the state and our estimator is a positive definite matrix).

The Kalman filter uses all information up to the given time. Firstly, we want to estimate the state in the future at the time t. For that we use *a priori* estimates and equations (2.9), (2.10), which are called *time update* equations.

$$\widehat{\boldsymbol{\alpha}}_{t|t-1} = F_t \widehat{\boldsymbol{\alpha}}_{t-1|t-1}, \tag{2.9}$$

$$P_{t|t-1} = F_t P_{t-1|t-1} F_t^\top + Q_t. \tag{2.10}$$

After we observe the measurement at a time t, we correct our estimates using equations (2.11), (2.12) and (2.13), which are called *measurement update* equations.

$$K_t = P_{t|t-1} Z_t^\top (Z_t P_{t|t-1} Z_t^\top + R_t)^{-1}, \tag{2.11}$$

$$\widehat{\boldsymbol{\alpha}}_{t|t} = \widehat{\boldsymbol{\alpha}}_{t|t-1} + K_t(\boldsymbol{y}_t - Z_t \widehat{\boldsymbol{\alpha}}_{t|t-1}), \tag{2.12}$$

$$P_{t|t} = (I - K_t Z_t) P_{t|t-1}. \tag{2.13}$$

The *a posteriori* state estimate in equation (2.12) is a linear combination of the *a priori* state estimate and the difference of a measurement $\boldsymbol{y}_t$ and its estimate $Z_t \widehat{\boldsymbol{\alpha}}_{t|t-1}$. The stress put on the difference is given by *gain* K_t, which is more naturally interpreted if we rewrite equation (2.12) to the form $\widehat{\boldsymbol{\alpha}}_{t|t} = (I - K_t Z_t)\widehat{\boldsymbol{\alpha}}_{t|t-1} + K_t \boldsymbol{y}_t$. Now we see that the matrix K_t determines how dependent is *a posteriori* state estimate on new measurement $\boldsymbol{y}_t$. If we suppose that $m = n$ and Z_t is regular then

$$\lim_{R_t \to 0} K_t = Z_t^{-1},$$

so $\widehat{\boldsymbol{\alpha}}_{t|t} = Z_t^{-1} \boldsymbol{y}_t$. That is natural, because when the variance of the random effect in equation (2.1) is zero and the matrix Z_t is regular then we can exactly derive that $\boldsymbol{\alpha}_t = Z_t^{-1} \boldsymbol{y}_t$ from equation (2.2).

The other extreme is when the variance of the state estimate goes to zero, then

$$\lim_{P_{t|t-1} \to 0} K_t = 0, \tag{2.14}$$

which means that $\widehat{\boldsymbol{\alpha}}_{t|t} = \widehat{\boldsymbol{\alpha}}_{t|t-1}$, so we do not get new information with a new observation $\boldsymbol{y}_t$. This occurs, for example, when $Q_t = \mathbf{0}$.

The state vector $\boldsymbol{\alpha}_t$ may include some fixed parameters that do not vary during the time. We simply set the corresponding components of the variance matrix Q_t to zero. However, in some applications it might be more convenient to exclude them from the state vector and rewrite equation (2.1) using a design matrix $\mathbb{X}_t$, which might be time-dependent, and fixed parameters $\boldsymbol{\beta}$:

$$\boldsymbol{y}_t = Z_t \boldsymbol{\alpha}_t + \mathbb{X}_t \boldsymbol{\beta} + \boldsymbol{v}_t. \tag{2.15}$$

If covariates are non-random or independent of state vectors $\boldsymbol{\alpha}_t^*$ and since $\boldsymbol{\beta}$ is non-random it does not affect the previous derivations of the posterior state variances and posterior state means are only shifted, e.g.

$$\widehat{\boldsymbol{\alpha}}_{t|t-1} = F_t \widehat{\boldsymbol{\alpha}}_{t-1|t-1} + \mathbb{X}_t \boldsymbol{\beta}.$$

2.2 Estimation of hyperparameters

To use the Kalman filter we need to know all values of matrices Z_t, F_t, R_t and Q_t and the fixed parameter $\boldsymbol{\beta}$. However, in practical situations they usually depend on some unknown hyperparameters $\boldsymbol{\theta}$. In this section we discuss their estimation based on maximum likelihood. The main reference for this section is Durbin and Koopman (2001) and Farhmeir and Tutz (1994). Maximum likelihood method is based on the joint distribution of all observations. The *direct* method defines likelihood as follows

$$L(\boldsymbol{\theta}) = p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T|\boldsymbol{\theta})$$

where $p(.)$ is used to denote a continuous density.

In most applications measurements are independent and the joint density can be rewritten as a product of densities $p(\boldsymbol{y}_t|\boldsymbol{\theta})$. However, in our application measurements are not independent as they are derived from state vectors, which are clearly dependent (see equation (2.2)). Nevertheless, we can use properties of conditional probabilities (see Anděl (2011)) iteratively and obtain

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{T} p(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}^*, \boldsymbol{\theta})$$

where in a special case $t = 0$ we define $p(\boldsymbol{y}_1|\boldsymbol{\theta}, \boldsymbol{y}_0^*) = p(\boldsymbol{y}_1|\boldsymbol{\theta})$. In theorem 6, we have derived the conditional distribution of $p(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}^*)$. It is normal with mean

$$\widehat{\boldsymbol{y}}_{t|t-1} = \mathsf{E}\left[\boldsymbol{y}_t|\boldsymbol{y}_{t-1}^*\right] = Z_t\widehat{\boldsymbol{\alpha}}_{t|t-1}$$

and variance

$$S_t = \mathrm{var}[\boldsymbol{y}_t|\boldsymbol{y}_{t-1}^*] = Z_t P_{t|t-1} Z_t^\top + R_t.$$

We plug this into a density of normal distribution and receive

$$p(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}^*) = (2\pi)^{-\frac{m}{2}}|S_t|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}_t - \widehat{\boldsymbol{y}}_{t|t-1})^\top S_t^{-1}(\boldsymbol{y}_t - \widehat{\boldsymbol{y}}_{t|t-1})\right).$$

Remark that for the initialization step $t = 1$ we assume that hyperparameters $\boldsymbol{a}_0$ and Q_0 are known.

Now we derive the form of the log-likelihood

$$\ell(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta})) = -\frac{mT}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\left(\log|S_t| + (\boldsymbol{y}_t - \widehat{\boldsymbol{y}}_{t|t-1})^\top S_t^{-1}(\boldsymbol{y}_t - \widehat{\boldsymbol{y}}_{t|t-1})\right).$$

Parameters might be included in a matrix S_t and in measurement predictions $\widehat{\boldsymbol{y}}_{t|t-1}$. The task is to maximize the log-likelihood with respect to the parameter $\boldsymbol{\theta}$. To perform such a maximization, numerical methods are used. Most of them are based on a Newton's method, which we omit to present but only refer to literature (see for example Durbin and Koopman, 2001). The practical approach of this method is that calculating the log-likelihood can be easily computed as a by-product of running the Kalman filter.

Another type of estimating parameters is called *indirect*. It is based on the complete data and the algorithm used is called EM algorithm. The complete log-likelihood is

$$L_c(\boldsymbol{\theta}) = p(\boldsymbol{y}_T^*, \boldsymbol{\alpha}_T^* | \boldsymbol{\theta}).$$

Derivation of the joint density takes advantage of the hierarchical structure of the dynamic linear model. The joint density uses iteratively the step:

$$p(\boldsymbol{y}_T^*, \boldsymbol{\alpha}_T^* | \boldsymbol{\theta}) = p(\boldsymbol{y}_T | \boldsymbol{\alpha}_T^*, \boldsymbol{y}_{T-1}^*, \boldsymbol{\theta}) p(\boldsymbol{\alpha}_T | \boldsymbol{\alpha}_{T-1}^*, \boldsymbol{y}_{T-1}^*, \boldsymbol{\theta}) p(\boldsymbol{y}_{T-1}^*, \boldsymbol{\alpha}_{T-1}^* | \boldsymbol{\theta}).$$

From the distributional assumptions both conditional densities can be simplified by letting only the last observation of a state vector in the condition, e.g. $p(\boldsymbol{y}_T | \boldsymbol{\alpha}_T^*, \boldsymbol{y}_{T-1}^*, \boldsymbol{\theta}) = p(\boldsymbol{y}_T | \boldsymbol{\alpha}_T, \boldsymbol{\theta})$. The log-likelihood is, apart from additive constants not containing $\boldsymbol{\theta}$, given by

$$\begin{aligned}
\ell_c(\boldsymbol{\theta}) = &- \frac{1}{2} \sum_{t=1}^{T} \left(\log |R_t| + (\boldsymbol{y}_t - Z_t \boldsymbol{\alpha}_t)^\top R_t^{-1} (\boldsymbol{y}_t - Z_t \boldsymbol{\alpha}_t) \right) \\
&- \frac{1}{2} \sum_{t=1}^{T} \left(\log |Q_t| + (\boldsymbol{\alpha}_t - F_t \boldsymbol{\alpha}_{t-1})^\top Q_t^{-1} (\boldsymbol{\alpha}_t - F_t \boldsymbol{\alpha}_{t-1}) \right) \\
&- \frac{1}{2} (\log |Q_0| + \left(\boldsymbol{\alpha}_0 - \boldsymbol{a}_0 \right)^\top Q_0^{-1} (\boldsymbol{\alpha}_0 - \boldsymbol{a}_0)) \, .
\end{aligned}$$

The kth E-step of the algorithm lies in computing

$$M(\boldsymbol{\theta} | \boldsymbol{\theta}^{(k)}) = \mathsf{E} \left[\ell_c(\boldsymbol{\theta}) | \boldsymbol{y}_T^*, \boldsymbol{\theta}^{(k)} \right].$$

To compute the conditional mean of the complete log-likelihood is not straightforward and one need to compute smoothed values of state vectors, e.g. $\mathsf{E} \left[\boldsymbol{\alpha}_t | \boldsymbol{y}_T^* \right]$. For this purpose the Kalman smoother fixed at $\boldsymbol{\theta}^{(k)}$ can be used and we can then solve the maximization of $M(\boldsymbol{\theta} | \boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}$ (see Farhmeir and Tutz, 1994, pages 268-269).

The direct approach is based on the so-called integrated likelihood because

$$p(\boldsymbol{y}_t | \boldsymbol{y}_{t-1}^*) = \int_{\mathbb{R}^m} p(\boldsymbol{y}_t, \boldsymbol{\alpha}_t | \boldsymbol{y}_{t-1}^*) \, d \, \boldsymbol{\alpha}_t.$$

In most situations calculating this integral is analytically impossible but in case of normal distributions it was computed easily. However, in more general cases (such as DGLM) the indirect approach is more handful as it is rather based on hierarchical structure than on distributional assumptions.

2.3 Missing observations

This section deals with a common situation of missing observations in measurements $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$. The advantage of the state space approach is the ease how missing observations can be handled. In this section we focus on a situation when some components of $\boldsymbol{y}_t$ are missing. It is a problem that we encounter in ice hockey matches. On specific day, some teams might not have scheduled any match, which can be understood as a missing observation. Days when no

matches are planned are omitted. However, in case of needed estimation of the state vector with no observation at the given time one uses the same idea (see Durbin and Koopman, 2001, pages 92-93).

Suppose that at a given time t some elements of the measurement vector $\boldsymbol{y}_t$ are missing. We denote $\widetilde{\boldsymbol{y}}_t$ vector with only non-missing observations that has dimension $\widetilde{m}$. For computational purposes it is convenient to define a matrix W_t with dimension $\widetilde{m} \times m$. It is the identity matrix with omitted rows where elements of the vector $\boldsymbol{y}_t$ are missing. It holds that

$$\widetilde{\boldsymbol{y}}_t = W_t \boldsymbol{y}_t.$$

To realize the fact that missing observations do not represent a major problem, one should realize the following

$$p(\boldsymbol{\alpha}_t|\boldsymbol{y}_t) = p(\boldsymbol{\alpha}_t|\widetilde{\boldsymbol{y}}_t).$$

It only states that the same information is included in vectors $\widetilde{\boldsymbol{y}}_t$ and $\boldsymbol{y}_t$. It means that we need to rewrite equation (2.1) for only non-missing observations. Hence,

$$\widetilde{\boldsymbol{y}}_t = \widetilde{Z}_t \boldsymbol{\alpha}_t + \widetilde{\boldsymbol{v}}_t$$

where

$$\widetilde{Z}_t = W_t Z_t, \quad \widetilde{\boldsymbol{v}}_t = W_t \boldsymbol{v}_t, \quad \mathrm{var}(\widetilde{\boldsymbol{v}}_t) = \widetilde{R}_t = W_t R_t W_t^\top.$$

Equation for the state vector (2.2) does not have to be changed. Filtering with the Kalman filter works the same only vector $\boldsymbol{y}_t$ and matrices Z_t, R_t are replaced by $\widetilde{\boldsymbol{y}}_t, \widetilde{Z}_t$ and $\widetilde{R}_t$. For the Kalman gain it means that corresponding columns are omitted and its dimension is $n \times \widetilde{m}$.

For a practical evaluation it means that *time update* equations (2.9) and (2.10) remain unchanged and in *measurement update* equations (2.11), (2.12) and (2.13) we omit components with the missing observations. However, we don't even need to change dimensions of $Z_t, \boldsymbol{y}_t$ and R_t but simply in places of the missing observations we set corresponding elements of Z_t and $\boldsymbol{y}_t$ to zero. It guarantees the same result.

Chapter 3

Generalizations of linear models

The linear system defined by equations (2.1) and (2.2) has many applications and can be applied in various situations, some fields mentioned in Anderson and Moore (1979) are economics, bioengineering or operations research. Nevertheless, the assumption of normal distribution for measurements $\boldsymbol{z}_t$ are, in some cases, unrealistic. To those cases belong any non-continuous distributions, e.g. Poisson, Bernoulli or multinomial. Those three mentioned are essential for modelling responses of counts, probability of success or probabilities with more than two categories.

Identical need for relaxing assumption of normal distribution in linear regression has led to establishing benchmark for distributions in exponential family. Models with response variable in exponential family are called generalized linear models (GLM). To this family belong (among others) Poisson,Bernoulli and multinomial distributions. The framework contains mainly an algorithm for estimating regression coefficients via iterated weighted least squares procedure.

The natural extension of the Kalman filter would be to allow measurements $\boldsymbol{z}_t$ to be in the class of exponential distributions. This approach is taken in article Fahrmeir (1992), which is the basis for our next section.

3.1 Multivariate generalized linear models

This section is based on Fahrmeir and Tutz (1994) and notes from class NMST432 Advanced regression models taught by Doc. Mgr. Michal Kulich, PhD. Generalized linear models represent a generalization of classical linear models in two ways. Firstly, they allow for wider class of distributions for a response variable, namely exponential family distributions. Secondly, they relax the dependence between the mean of the response and covariates from linear to linear after some transformation.

Exponential family

The exponential family is a class of probability distributions sharing a certain form, specified below. Into the exponential family belong, among others, normal, exponential, gamma as well as Bernoulli and multinomial distributions. The last two are the most important for our purposes. Bernoulli distribution is relevant if we are interested in matches with two possible outcomes – loss or win and

multinomial distribution in case of three categories – loss, draw or win. All given examples are univariate distributions but multinomial.

Definition 6. *We say that a r-dimensional random variable $\boldsymbol{Y}$ comes from an exponential family if its probability density function with respect to a σ-finite measure can be expressed in the form*

$$p(\boldsymbol{y}|\boldsymbol{\theta},\varphi) = \exp\left(\frac{\boldsymbol{y}^\top\boldsymbol{\theta} - b(\boldsymbol{\theta})}{\varphi} + c(\boldsymbol{y},\varphi)\right) \tag{3.1}$$

where $b(\boldsymbol{\theta}), c(\boldsymbol{y},\varphi)$ are known real functions.

The r-dimensional parameter $\boldsymbol{\theta}$ is called a *canonical parameter* and φ is called a *dispersion parameter* for which we demand to be positive.

Distributions in form (3.1) are sometimes called an *exponential dispersion family* as it allows different structure for the dispersion parameter. In the following lemma, we calculate mean and variance using a moment generating function within this class of distributions.

Lemma 10. *Suppose a r-dimensional random variable $\boldsymbol{Y}$ comes from an exponential family with density having the form* (3.1) *with respect to a σ-finite measure ν. Its canonical parameter $\boldsymbol{\theta}$ lies in an open, convex and non-empty space $\Theta \subset \mathbb{R}^r$ and the real function $b(.)$ is twice continuously differentiable. Then*

$$M_{\boldsymbol{Y}}(\boldsymbol{t}) = \exp\left(\frac{b(\boldsymbol{\theta}+\varphi\boldsymbol{t}) - b(\boldsymbol{\theta})}{\varphi}\right),$$

$$\mathsf{E}\ \boldsymbol{Y} = \mu(\boldsymbol{\theta}) = \frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{3.2}$$

$$\operatorname{var}(\boldsymbol{Y}) = \Sigma(\boldsymbol{\theta}) = \varphi\frac{\partial^2 b(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}. \tag{3.3}$$

Proof. We calculate a moment generating function for $\boldsymbol{t}$ such that $\boldsymbol{\theta}+\varphi\boldsymbol{t} \in \Theta$

$$\begin{aligned}
M_{\boldsymbol{Y}}(\boldsymbol{t}) &= \mathsf{E}\exp(\boldsymbol{t}^\top\boldsymbol{Y}) = \int_{\mathbb{R}^r} \exp(\boldsymbol{t}^\top\boldsymbol{Y}) p(\boldsymbol{y}|\boldsymbol{\theta},\varphi) d\nu(\boldsymbol{y}) \\
&= \int_{\mathbb{R}^r} \exp\left(\frac{\boldsymbol{y}^\top(\boldsymbol{\theta}+\varphi\boldsymbol{t}) - b(\boldsymbol{\theta}+\varphi\boldsymbol{t}) + b(\boldsymbol{\theta}+\varphi\boldsymbol{t}) - b(\boldsymbol{\theta})}{\varphi} + c(\boldsymbol{y},\varphi)\right) d\nu(\boldsymbol{y}) \\
&= \int_{\mathbb{R}^r} p(\boldsymbol{y}|\boldsymbol{\theta}+\varphi\boldsymbol{t},\varphi) d\nu(\boldsymbol{y}) \exp\left(\frac{b(\boldsymbol{\theta}+\varphi\boldsymbol{t}) - b(\boldsymbol{\theta})}{\varphi}\right) \\
&= \exp\left(\frac{b(\boldsymbol{\theta}+\varphi\boldsymbol{t}) - b(\boldsymbol{\theta})}{\varphi}\right).
\end{aligned}$$

Using fundamental properties of a moment generating function we calculate mean and variance as:

$$\mathsf{E}\ \boldsymbol{Y} = \left.\frac{\partial M_{\boldsymbol{Y}}(\boldsymbol{t})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{t}=\boldsymbol{0}} = \frac{\partial b(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}.$$

$$\operatorname{var}(\boldsymbol{Y}) = \left.\frac{\partial^2 M_{\boldsymbol{Y}}(\boldsymbol{t})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right|_{\boldsymbol{t}=\boldsymbol{0}} - (\mathsf{E}\ \boldsymbol{Y})^2 = \varphi\frac{\partial^2 b(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}.$$

□

As a variance matrix is always positive semi-definite (see Anděl, 2011, page 32), from equation (3.3) we can conclude that the Hessian matrix of b is also positive semi-definite (we know that $\varphi > 0$) and therefore b is convex. Typically the variance $\text{var}(\boldsymbol{Y})$ is a regular matrix, and therefore b is strictly convex and b' is strictly increasing, so $\mu(\boldsymbol{\theta}) : \Theta \to M = \mu(\Theta)$ is injective and the variance is a function of $\boldsymbol{\mu}$:

$$\text{var}(\boldsymbol{Y}) = \varphi \frac{\partial^2 b(\boldsymbol{\theta}(\boldsymbol{\mu}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \varphi \frac{\partial^2 b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\bigg|_{\boldsymbol{\theta}=(\nabla b)^{-1}(\boldsymbol{\mu})} = \varphi V(\boldsymbol{\mu}). \tag{3.4}$$

Moreover, the exponential distribution for given functions $b(.)$ and $c(.)$ is determined by $\boldsymbol{\mu}$ and φ, so we could rewrite the definition 6 using $p(\boldsymbol{y}|\boldsymbol{\mu}, \varphi)$.

Multinomial distribution as a member of exponential family

Our main focus is on categorical variables, which can be described by probabilities of occurrence for given categories. Let Y denote a categorical variable with probability for category q denoted $\pi_q = P(Y = q)$ where $q = 0, \ldots, r$. The distribution of Y is determined by probabilities $\pi_1, \ldots, \pi_r$ and can be described as a multinomial distribution with one trial. This is done by transforming Y using dummy variables $Y_q = \mathbb{1}_{[Y=q]}$ and stacking them into a vector. We introduce a random vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_r)^\top$ which contains the same information as Y itself and has a multinomial distribution. We consider a multinomial distribution with one category omitted as in Fahrmeir and Tutz (1994). It means

$$\boldsymbol{Y} \sim Mult(1, (\pi_1, \ldots, \pi_r)^\top), \quad \sum_{q=0}^{r} \pi_q = 1, \quad Y_0 = 1 - \sum_{q=1}^{r} Y_q.$$

Density mass function can be expressed as follows

$$\begin{aligned} p(\boldsymbol{y}|\pi_1, \ldots, \pi_r) = \Pi_{q=0}^{r} \pi_q^{y_q} &= \exp\left(\sum_{q=0}^{r} y_q \log(\pi_q)\right) \\ &= \exp\left(\sum_{q=1}^{r} y_q \log(\pi_q) + \left(1 - \sum_{q=1}^{r} y_q\right) \log(\pi_0)\right) \\ &= \exp\left(\sum_{q=1}^{r} y_q \log\left(\frac{\pi_q}{\pi_0}\right) + \log(\pi_0)\right) \\ &= \exp\left(\boldsymbol{y}^\top \boldsymbol{\theta} - b(\boldsymbol{\theta})\right). \end{aligned}$$

We have defined the parameter $\boldsymbol{\theta}$ with components

$$\theta_q = \log\left(\frac{\pi_q}{\pi_0}\right), \quad q = 1, \ldots, r. \tag{3.5}$$

Definition of function $b(.)$ is more hidden. But we can easily derive its form

$$1 + \sum_{q=1}^{r} \exp(\theta_q) = 1 + \frac{\sum_{q=1}^{r} \pi_q}{\pi_0} = \frac{1}{\pi_0}. \tag{3.6}$$

Special case of a multinomial distribution with $r = 1$ is bivariate (Bernoulli), and then function of π_1 in (3.5) is called a *logit* function.

By combining equations (3.5) and (3.6) we get

$$b(\boldsymbol{\theta}) = -\log(\pi_0) = \log\left(\frac{1}{\pi_0}\right) = \log\left(1 + \sum_{q=1}^{r} \exp(\theta_q)\right). \tag{3.7}$$

We have shown that multinomial distribution belongs to the exponential family with a dispersion parameter $\varphi = 1$, function $c(.) \equiv 1$ and function $b(.)$ given by (3.7). The formula for π_q depending on θ_q can be derived easily using the definition of θ_q and formula (3.6).

$$\pi_q = \frac{\exp(\theta_q)}{1 + \sum_{i=1}^{r} \exp(\theta_i)}, \quad q = 1, \ldots, r,$$
$$\pi_0 = \frac{1}{1 + \sum_{i=1}^{r} \exp(\theta_i)}.$$

For computing mean and variance we can use lamma 10.

$$\mathsf{E}\, Y_q = \frac{\partial b(\boldsymbol{\theta})}{\partial \theta_q} = \frac{\exp(\theta_q)}{1 + \sum_{i=1}^{r} \exp(\theta_i)} = \pi_q,$$
$$\operatorname{var} Y_q = \frac{\partial^2 b(\boldsymbol{\theta})}{\partial \theta_q \partial \theta_q} = \frac{\exp(\theta_q)(1 + \sum_{i \neq j} \exp(\theta_i))}{(1 + \sum_{i=1}^{r} \exp(\theta_i))^2} = \pi_q(1 - \pi_q), \tag{3.8}$$
$$\operatorname{cov}(Y_q, Y_s) = \frac{\partial^2 b(\boldsymbol{\theta})}{\partial \theta_q \partial \theta_s} = -\frac{\exp(\theta_q)\exp(\theta_s)}{(1 + \sum_{i=1}^{r} \exp(\theta_i))^2} = -\pi_q \pi_s, \quad q \neq s. \tag{3.9}$$

From (3.8) and (3.9) we can conclude that the variance matrix of a multinomial distribution $\boldsymbol{Y}$ is

$$\operatorname{var}(\boldsymbol{Y}) = \Sigma(\boldsymbol{\pi}) = \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_r \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_r\pi_1 & -\pi_r\pi_2 & \cdots & \pi_r(1-\pi_r) \end{pmatrix}.$$

Definition 7. *Suppose observations* $(\boldsymbol{Y}_i^\top, \boldsymbol{X}_i^\top)^\top$, $i = 1, \ldots, n$ *where* $\boldsymbol{Y}_i$ *are* r*-dimensional responses and* $\boldsymbol{X}_i$ *are covariates. Multivariate generalized linear model is characterized by the following structure:*

(i) $\boldsymbol{Y}_1, \ldots \boldsymbol{Y}_n$ *are independent and distribution of* $\boldsymbol{Y}_i$ *depends on* $\boldsymbol{X}_i$ *through regression coefficients* $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ *that come from an admissible open set* $B \subset \mathbb{R}^p$.

(ii) The conditional distribution of $\boldsymbol{Y}_i$ *given* $\boldsymbol{X}_i$ *is exponential with* $b(.)$ *being twice continuously differentiable and* $\boldsymbol{\theta}_i$ *depends on covariates* $\boldsymbol{X}_i$ *and* $\boldsymbol{\beta}$ *through linear predictor*

$$\boldsymbol{\eta}_i = \mathbb{Z}_i \boldsymbol{\beta}$$

where $\mathbb{Z}_i = \boldsymbol{Z}(\boldsymbol{X}_i)$ *is* $(q \times p)$ *- design matrix.*

(iii) There exists a strictly monotone and twice continuously differentiable link function $g : \mathbb{R}^r \to \mathbb{R}^r$ such that

$$g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i$$

where $\boldsymbol{\mu}_i = \mathsf{E}\ \boldsymbol{Y}_i$.

For some theoretical purposes it is convenient to relate *linear predictor* $\boldsymbol{\eta}_i$ to the natural parameter $\boldsymbol{\theta}_i$. For that we can use equation (3.2) and then holds

$$\boldsymbol{\theta}_i = u(\boldsymbol{\eta}_i) = (g \circ \nabla b)^{-1}(\boldsymbol{\eta}_i) = (g \circ \nabla b)^{-1}(\mathbb{Z}_i\boldsymbol{\beta}).$$

Special case represent link functions that are chosen in a way that $\boldsymbol{\theta}_i = \boldsymbol{\eta}_i$, they are called *natural* links. For a given exponential distribution (known $b(.)$ and $c(.)$) and with $\nabla b(.)$ having inverse, choosing $g(\boldsymbol{\alpha}) = (\nabla b)^{-1}(\boldsymbol{\alpha})$ we get the natural link.

In the following, we make an inference on the regression parameter $\boldsymbol{\beta}$, so for shorter notation we will understand parameters $\boldsymbol{\theta}_i$, $\boldsymbol{\mu}_i$ as functions of $\boldsymbol{\beta}$.

MLE estimates

Estimation of parameters for a GLM model is done by maximizing the likelihood. Since observations $(\boldsymbol{Y}_i^\top, \boldsymbol{X}_i^\top)^\top, i = 1, \ldots, n$ that follow a GLM model are independent the log-likelihood can be expressed as a sum of log-likelihood contributions as follows

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\boldsymbol{Y}_i^\top \boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)}{\varphi}.$$

The derivative of the likelihood with respect to $\boldsymbol{\beta}$ is called a score vector and its maximization yields the desired estimate. The maximum is not reached analytically but via iteratively reweighted least squares procedure. In case of the natural link function the score vector is concave, which simplifies the maximization.

3.2 Ordinal paired comparison data

This section is based mainly on the article Agresti (1992) by Alan Agresti. In this section we present a non-dynamic model for a categorical variable coming from pairwise comparisons.

Let Y_{ij} denote the categorical variable of interest that represents the result of a comparison between teams i and j. Possible outcomes of Y_{ij} are ordinal categories $0, \ldots, r$, where $q = 0$ is the least favourable for team i and $q = r$ is the most favourable one for i. Notice that in our setting the order of indexes i and j matters. Further, we suppose that the scale is symmetric, such that $Y_{ij} = q$ is equivalent to $Y_{ji} = r - q$ for all q. As in the usual ordinal model we suppose that there is an underlying latent continuous variable Y_{ji}^* and cutpoints $\gamma_0 < \gamma_1 < \ldots < \gamma_{r-1}$ such that the following holds

$$\mathsf{P}(Y_{ij} = q) = \mathsf{P}(\gamma_{q-1} < Y_{ij}^* \leq \gamma_q), \quad \forall q \in 0, \ldots, r.$$

For easier notation we have defined $\gamma_{-1} = -\infty$ and $\gamma_r = \infty$. The latent variable Y_{ij}^* can be decomposed for interpretation into three components, such that $Y_{ij}^* = \mu_i - \mu_j + \varepsilon_{ij}$. Two of them represent the team performance in the match, which are supposed to be non-random, and the last one can be interpreted as randomness. About the random component we suppose that it follows the same distribution for any match, namely:

$$\varepsilon_{ij} \sim F \text{ ; strictly increasing, continuous.} \tag{3.10}$$

It is a straightforward calculation to derive the probability of $Y_{ij} = q$.

$$\begin{aligned}
\mathsf{P}(Y_{ij} = q) &= \mathsf{P}(\gamma_{q-1} < Y_{ij}^* \leq \gamma_q) \\
&= \mathsf{P}(\gamma_{q-1} < \mu_i - \mu_j + \varepsilon_{ij} \leq \gamma_q) \\
&= \mathsf{P}(\gamma_{q-1} - \mu_i + \mu_j < \varepsilon_{ij} \leq \gamma_q - \mu_i + \mu_j) \\
&= F(\gamma_q - \mu_i + \mu_j) - F(\gamma_{q-1} - \mu_i + \mu_j).
\end{aligned}$$

If we broaden the assumption of a symmetrical scale also for latent variables Y_{ij}^*, such that $Y_{ij}^* = -Y_{ji}^*$, it ensures that $\varepsilon_{ij} = -\varepsilon_{ji}$. Using that and (3.10) put constraint on a distribution function F

$$F(x) = \mathsf{P}(\varepsilon_{ij} \leq x) = \mathsf{P}(-\varepsilon_{ij} \geq -x) = 1 - \mathsf{P}(\varepsilon_{ji} \leq -x) = 1 - F(-x). \tag{3.11}$$

Important possible options for the distribution function F are logistic and normal.

The assumption of symmetry for Y_{ij} and equation (3.11) leads to a constraint of cutpoints

$$\begin{aligned}
\mathsf{P}(Y_{ij} \leq q) &= \mathsf{P}(Y_{ji} \geq r - q) \\
F(\gamma_q - \mu_i + \mu_j) &= 1 - F(\gamma_{r-q-1} + \mu_i - \mu_j) \\
F(\gamma_q - \mu_i + \mu_j) &= F(-\gamma_{r-q-1} - \mu_i + \mu_j) \\
\gamma_q &= -\gamma_{r-q-1}.
\end{aligned} \tag{3.12}$$

Model identifiability is ensured by an assumption such as $\mu_1 = 0$ or $\sum \mu_i = 0$.

Cumulative logit model

The preceding section was formed in a general setting. For our purposes we present a specific model for response (result of a match) with only three categories, such that

$$Y_{ij} = \begin{cases} 0 & \text{team } i \text{ loses,} \\ 1 & \text{draw,} \\ 2 & \text{team } i \text{ wins.} \end{cases}$$

The assumption of symmetry of the scale holds because the probability of win of the team i equals the probability of loss of the team j, etc.

As the distribution function F from (3.10) we use logistic distribution function with the known inverse called logit, which is

$$F^{-1}(x) = \text{logit}(x) = \log\left(\frac{x}{1-x}\right).$$

In the preceding section we have not specified performances μ_i. For some applications it could be interpreted straightforward as a form of unit i (e.g. chess competition). But since we are interested in applications for ice hockey a major factor besides the current form could be a home advantage. So the performance can be decomposed into components an ability α_i and a home advantage (HA_i). As in Knorr-Held (2000) we assume that the home advantage is the same for each team i:

$$HA_i = HA, \ \forall i = 1, \ldots, n.$$

This assumption is not rejected in the practical part. If the game is played at a stadium of the team i then $\mu_i = \alpha_i + HA$ and $\mu_j = \alpha_j$. From the assumption of a symmetrical treatment we have received the constraint (3.12), which in our setting means that $\gamma_0 = -\gamma_1$ and $\gamma_1 > 0$. The model now yields

$$\begin{aligned} \text{logit}(\mathsf{P}(Y_{ij} \leq 0)) &= -\gamma_1 - \alpha_i - HA + \alpha_j, \\ \text{logit}(\mathsf{P}(Y_{ij} \leq 1)) &= \gamma_1 - \alpha_i - HA + \alpha_j. \end{aligned}$$

Remark that in this parametrization higher ability of α_i indicates lower probability of loss. Therefore, the interpretation corresponds with the usual usage. Now we set $\theta_0 = -\gamma_1 - HA$ and $\theta_1 = \gamma_1 - HA$.

Cumulative logit model for a three categorical variable Y_{ij} with a home advantage for team i has the form

$$\text{logit}(\mathsf{P}(Y_{ij} \leq q)) = \theta_q - \alpha_i + \alpha_j, \quad q = 0, 1 \text{ and } \theta_0 < \theta_1. \tag{3.13}$$

Notice that we have lost the symmetry of Y_{ij} in our new notation. Now the order of indexes ij determines that the home advantage has the team i. Another specification of this model is in terms of odds, which are defined in the following definition.

Definition 8. *Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space and $A \in \mathcal{A}$ satisfies $\mathsf{P}(A) < 1$. Then the ratio*

$$\text{odds}(A) = \frac{\mathsf{P}(A)}{1 - \mathsf{P}(A)}$$

is called odds for event A.

Note. The typical usage is with a categorical variable Y and its probabilities. This enables us also easier definition of logit:

$$\text{logit}(\mathsf{P}(Y_{ij} \leq 1)) = \log(\text{odds}(\mathsf{P}(Y_{ij} \leq 1))).$$

In the model (3.13) parameters α_i and α_j do not depend on q, this means that log odds of the response being under q differ only by θ_q and the model assumes the same effect across all levels. This gives another name for the model (3.13) – *proportional odds model*.

Interpretation of parameters

To get an interpretation for parameters of ability, α_i, is straightforward. Suppose $\widetilde{\alpha}_i = \alpha_i + x$, it is the ability of team i increased by x. Then

$$\log(\text{odds}(\mathsf{P}(Y_{ij} \leq 0|\widetilde{\alpha}_i, \alpha_j))) - \log(\text{odds}(\mathsf{P}(Y_{ij} \leq 1|\alpha_i, \alpha_j))) = -\alpha_i + \widetilde{\alpha}_i = -x,$$
$$\log(\text{odds}(\mathsf{P}(Y_{ij} \leq 1|\widetilde{\alpha}_i, \alpha_j))) - \log(\text{odds}(\mathsf{P}(Y_{ij} \leq 2|\alpha_i, \alpha_j))) = -\alpha_i + \widetilde{\alpha}_i = -x.$$

With increase x of ability α_i the logarithm of odds for loss and draw of the team i are decreased by x.

Another parameter of interest is the home advantage.

$$\begin{aligned}\log\left(\frac{\text{odds}(\mathsf{P}(Y_{ij} = 2))}{\text{odds}(\mathsf{P}(Y_{ji} = 0))}\right) &= \log(\text{odds}(\mathsf{P}(Y_{ij} = 2))) - \text{logit}(\mathsf{P}(Y_{ji} \leq 0)) \\ &= -\log(\text{odds}(\mathsf{P}(Y_{ij} \leq 1))) - \text{logit}(\mathsf{P}(Y_{ji} \leq 0) \\ &= -\theta_2 - \theta_1 = 2HA.\end{aligned}$$

From the preceding derivation, we can say that ratio of odds to win at home and odds to win out equals $\exp(2HA)$. The higher is the home advantage the higher are odds to win home compared to odds to win out.

Lemma 11. *Let $(Y_i, \boldsymbol{x}_i^\top)^\top$, $i = 1, \ldots, n$ be a random sample where Y_i is a categorical variable with categories $q = 0, \ldots, r$ satisfying cumulative model*

$$P(Y_i \leq q|\boldsymbol{x}_i) = F(\theta_q - \boldsymbol{x}_i^\top \boldsymbol{\alpha}), \quad q = 0, \ldots, r$$

where $-\infty < \theta_0 < \theta_1 < \ldots < \theta_r = \infty$, a strictly increasing distribution function F and $\boldsymbol{x}_i$ vector of covariates. Then $(\boldsymbol{Y}_i^\top, \boldsymbol{x}_i^\top)^\top$, $i = 1, \ldots, n$ follow a multivariate generalized linear model with

$$\begin{aligned}g_q(\pi_0, \ldots, \pi_{r-1}) &= F^{-1}(\pi_0 + \ldots + \pi_q), \quad q = 0, \ldots, r-1, \\ \mathbb{Z}_i &= \begin{pmatrix} 1 & & \cdots & & -\boldsymbol{x}_i^\top \\ & 1 & & & -\boldsymbol{x}_i^\top \\ & & \ddots & & \vdots \\ & & & 1 & -\boldsymbol{x}_i^\top \end{pmatrix}, \quad i = 1, \ldots, n \\ \boldsymbol{\beta} &= (\theta_0, \ldots, \theta_{r-1}, \boldsymbol{\alpha})\end{aligned}$$

where $\boldsymbol{Y}_i = (\mathbb{1}_{[Y_i=0]}, \ldots, \mathbb{1}_{[Y_i=r-1]})$.

Proof. We have shown that the categorical variable belongs to an exponential family. From the form of a cumulative model, we have

$$F^{-1}(\pi_0 + \ldots + \pi_{q-1}) = \theta_{q-1} - \boldsymbol{x}_i^\top \alpha = \boldsymbol{z}_{iq}\boldsymbol{\beta}, \quad q = 1, \ldots, r$$

where $\boldsymbol{z}_{iq}$ is qth row of a design matrix $\mathbb{Z}_i$. The preceding equations give a form of the function g and the design matrix $\mathbb{Z}_i$. □

The preceding lemma ensures that for the cumulative link model (3.13) the theory for GLM models applies.

3.3 Dynamic generalized linear model

The overall setting is a generalization of linear state space with a non-observable *state vectors* $\boldsymbol{\alpha}_t$ and *measurements* $\boldsymbol{z}_t$. The update equation (2.1) for a state parameter $\boldsymbol{\alpha}_t$ remains the same. The modification lies in the connection between the state vector $\boldsymbol{\alpha}_t$ and the current measurement $\boldsymbol{z}_t$. We assume that the conditional density $p(\boldsymbol{z}_t|\boldsymbol{\alpha}_t)$ is in the exponential class. Moreover, $\boldsymbol{z}_t$ together with $\boldsymbol{\alpha}_t$ as a regression coefficient and non-specified covariates form a multivariate generalized linear model (according to a definition 7).

Multivariate dynamic generalized linear model is specified by structural conditions:

$$\begin{aligned} \boldsymbol{\alpha}_t &= F_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{w}_t, \\ \boldsymbol{\mu}_t &= \mathsf{E}\left[\boldsymbol{z}_t|\boldsymbol{\alpha}_t\right] = h(\boldsymbol{\eta}_t) = h(\mathbb{Z}_t\boldsymbol{\alpha}_t), \end{aligned} \tag{3.14}$$

by distributional conditions

$$\boldsymbol{w}_t \overset{\text{iid}}{\sim} \mathsf{N}(0, Q_t), \quad \boldsymbol{\alpha}_0 \sim \mathsf{N}(\boldsymbol{a}_0, Q_0), \tag{3.15}$$

$$\ell_t(\boldsymbol{\alpha}_t) = \log(p(\boldsymbol{z}_t|\boldsymbol{\alpha}_t^*, \boldsymbol{z}_{t-1}^*)) = \log(p(\boldsymbol{z}_t|\boldsymbol{\alpha}_t, \boldsymbol{z}_{t-1}^*)) = \boldsymbol{z}_t^\top\boldsymbol{\theta}_t - b(\boldsymbol{\theta}_t) + c(\boldsymbol{z}_t),$$

$$\boldsymbol{\alpha}_t \in \mathbb{R}^n,\ F_t \in \mathbb{R}^{n\times n},\ Q_t > 0,\ \boldsymbol{z}_t, \boldsymbol{\mu}_t, \boldsymbol{\theta}_t \in \mathbb{R}^m, \quad \mathbb{Z}_t \in \mathbb{R}^{m\times n},\ g : \mathbb{R}^m \to \mathbb{R}^m$$

and further technical conditions as in definition 7.

Conditions (3.14) and (3.15) ensure that process $\{\boldsymbol{\alpha}_t\}$ has Markov property:

$$p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}^*) = p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}). \tag{3.16}$$

The equation (3.16) means that for current state the information in the whole vector of previous states is the same as the information within the last observation.

Posterior mode estimation

This section is based on articles Fahrmeir (1992) and Farhmeir and Kaufmann (1991). The goal is to get filtered estimation of the state vector $\boldsymbol{\alpha}_t$. In a linear state space model, we use several properties of normal distribution, which enable to derive exact distribution of $\boldsymbol{\alpha}_{t|t-1}$ and $\boldsymbol{\alpha}_{t|t}$. Moreover, the mean of $\boldsymbol{\alpha}_{t|t}$ is only a linear combination of new observation $\boldsymbol{z}_t$ and its previous estimation $\widehat{\boldsymbol{\alpha}}_{t|t-1}$. However, if $\boldsymbol{z}_t$ do not have normal distribution the derivation of the posterior distribution $\boldsymbol{\alpha}_{t|t}$ is not possible without integration. In the articles is proposed to estimate a posterior mode because it does not require integration.

Theorem 12. *Suppose $\boldsymbol{\alpha}_t$ and $\boldsymbol{z}_t$ follow a multivariate dynamic generalized linear model. Then*

$$p(\boldsymbol{\alpha}_t^*|\boldsymbol{z}_t^*) \propto \prod_{s=1}^{t} p(\boldsymbol{z}_s|\boldsymbol{\alpha}_s) \prod_{s=1}^{t} p(\boldsymbol{\alpha}_s|\boldsymbol{\alpha}_{s-1}),$$

$$\ell_t^*(\boldsymbol{\alpha}_t^*) = \log(p(\boldsymbol{\alpha}_t^*|\boldsymbol{z}_t^*)) = \sum_{s=1}^{t} \ell_s(\boldsymbol{\alpha}_s) + a_t(\boldsymbol{\alpha}_t^*) + c$$

where c is a constant and

$$a_t(\boldsymbol{\alpha}_t^*) = -\frac{1}{2}\sum_{s=1}^{t}(\boldsymbol{\alpha}_s - F_t\boldsymbol{\alpha}_{s-1})^\top Q_s^{-1}(\boldsymbol{\alpha}_s - F_t\boldsymbol{\alpha}_{s-1}) - \frac{1}{2}(\boldsymbol{\alpha}_0 - \boldsymbol{a}_0)^\top Q_0^{-1}(\boldsymbol{\alpha}_0 - \boldsymbol{a}_0).$$

Proof. The proof is based on Bayes theorem 1.1 and Markovian property (3.16):

$$\begin{aligned} p(\boldsymbol{\alpha}_t^*|\boldsymbol{z}_t^*) &\propto p(\boldsymbol{z}_t|\boldsymbol{\alpha}_t^*, \boldsymbol{z}_{t-1}^*)p(\boldsymbol{\alpha}_t^*, \boldsymbol{z}_{t-1}^*) = p(\boldsymbol{z}_t|\boldsymbol{\alpha}_t)p(\boldsymbol{\alpha}_t^*, \boldsymbol{z}_{t-1}^*) \\ &= p(\boldsymbol{z}_t|\boldsymbol{\alpha}_t)p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}^*, \boldsymbol{z}_{t-1}^*)p(\boldsymbol{\alpha}_{t-1}^*, \boldsymbol{z}_{t-1}^*) = p(\boldsymbol{z}_t|\boldsymbol{\alpha}_t)p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1})p(\boldsymbol{\alpha}_{t-1}^*, \boldsymbol{z}_{t-1}^*). \end{aligned}$$

We get the desired form using iteratively the same steps. To derive form of $\ell_t^*(\boldsymbol{\alpha}_t^*)$ we use the preceding result and definition of normal distribution. □

Note. The criterion $\ell_t^*(\boldsymbol{\alpha}_t^*)$ can be interpreted as a penalized log-likelihood. In frequentist statistics we would be interested in maximizing likelihood $\sum_{s=1}^{t} \ell_s(\boldsymbol{\alpha}_s)$. In our setting, there is an additional term $a_t(\boldsymbol{\alpha}_t^*)$ serving as a smoothness prior.

For a symmetric unimodal distribution its mode coincides with its mean. Hence, in dynamic linear model the estimation of mean coincides with the estimation of mode. In Farhmeir and Kaufmann (1991) is proposed to estimate posterior mode by solving $\max_{\boldsymbol{\alpha}_t^*} \ell_t^*(\boldsymbol{\alpha}_t^*)$ using numerical Gauss-Newton or Fisher-Scoring algorithm. Four algorithms for the numerical solution are presented and compared and all of them simplifies into the Kalman filter in the linear Gaussian model. The last performs only a single step of a numerical estimation and it is convenient for the fact that can be calculated iteratively. It was called generalized Kalman filter as it resembles the Kalman filter like a generalized linear model resembles a normal linear model.

The generalized Kalman filter for a numerical estimation of a posterior mode runs in two steps. Firstly, the prediction step remains the same as for the Kalman filter, so *time update equations* are

$$\begin{aligned} \widehat{\boldsymbol{\alpha}}_{t|t-1} &= F_t \widehat{\boldsymbol{\alpha}}_{t-1|t-1}, \\ P_{t|t-1} &= F_t P_{t-1|t-1} F_t^\top + Q_t, \end{aligned}$$

followed by *measurement update equations*

$$\begin{aligned} K_t &= P_{t|t-1} Z_t H_t \left(H_t^\top Z_t^\top P_{t-1|t-1} Z_t H_t + \Sigma_t \right), \\ \widehat{\boldsymbol{\alpha}}_{t|t} &= \widehat{\boldsymbol{\alpha}}_{t|t-1} + K_t(\boldsymbol{y}_t - h(Z_t \widehat{\boldsymbol{\alpha}}_{t|t-1})), \\ P_{t|t} &= (\mathbb{I} - K_t H_t Z_t) P_{t|t-1} \end{aligned}$$

where $\Sigma_t(\boldsymbol{\alpha}_t) = \mathrm{var}[\boldsymbol{z}_t|\boldsymbol{\alpha}_t]$, $H_t(\boldsymbol{\alpha}_t) = \frac{\partial}{\partial \boldsymbol{\eta}_t} h(\boldsymbol{\eta}_t)$. Matrices Σ_t and H_t are those matrices evaluated at $\widehat{\boldsymbol{\alpha}}_{t|t-1}$ and $\mathbb{I}$ denotes an identity matrix.

This algorithm was applied in article Farhmeir and Tutz (1994) to ordered paired comparison systems. We may use the concept to ice hockey matches with measurements $\boldsymbol{z}_t$ as a result of a game and $\boldsymbol{\alpha}_t$ as a vector of unobserved forms. However, in the practical part was found out that using goal differences is more informative than using plain result (see section 4.9). For that reason, this approach was not implemented.

Chapter 4

Analysis of Czech extraliga 1999-2014

Ice hockey overview

Ice hockey is a team sport played on ice in which two teams of skaters tries to place a puck into opponent's net using wooden sticks. Each team has one goaltender and five players – usually three forwards and two defencemen on ice. Goaltender has a special gear, blocker, catch glove, and leg pads, which he uses to prevent the puck from getting into h is team's net. In contrast to rest of the players, one goaltender usually plays the whole match; the coach switches him only in case of injury or poor performance. The other players take turns in irregular intervals but time spent on ice is usually less than one minute. The game is played in three thirds and each lasts 20 minutes. The clock is stopped whenever the game is halted.

Since ice hockey is a full contact sport many injuries happen during the season. Some are minor that prevent the injured from continuing in the following match but some are major, which disable the player from continuing in the running season for months or even forever. Estimated injury rate in Extraliga season 2010/2011 was 57.4 injuries in 1000 matches based on a survey (see Šulcová, 2011) where further statistics can be found.

Czech extraliga

Ice hockey has a long tradition in the Czech Republic. There are several competitions and most of them are run by Czech Ice Hockey Association (see `http://www.cslh.cz/`). The three major leagues are

- Tipsport Extraliga,
- 1. liga ČR (Premier league),
- 2. liga ČR (Second league).

Czech extraliga is the highest ice hockey league in the Czech republic. It was created in 1993 when the Czechoslovak First Ice Hockey League split following the breakup of Czechoslovakia. The official name has changed several times depending on the main sponsor.

Official names are:

- 1999–2000 — Staropramen Extraliga,
- 2001–2002 — Český Telecom Extraliga,
- 2003–2006 — Tipsport Extraliga,
- 2007–2010 — O2 Extraliga,
- 2010–current — Tipsport Extraliga.

System of extraliga

Main part of the league is played from September to February, within which all 14 teams meet four times, twice on the home field and twice as a guest team, so there are 52 rounds in one season. Until season 2005/2006 top 8 teams were qualified for playoff. The last team had to play against the first team in First League about the place in Extraliga. Since 2006/2007 top 6 teams have been qualified for the play-offs, teams finished as seventh to ninth play a play-in to determine, which two teams will join the top 6 into the play-off quarter-finals. In play-in play seventh against tenth and eighth against ninth on three wins.

Then a classic play-off follows. At the beginning, the following matches are played: the first against the eight, the second against the seventh, the third against the sixth and the fourth against the fifth. Those and following matches are played on 4 wins; the 4 winning teams continue into semifinal and winners into final. The winner of the finals becomes the league champion.

Eleventh till fourteenth teams after the main part play in a play-out group to maintain in Extraliga. Each team meets with one another four times and points from the main part count as well. Teams finishing as last two have to play against two best teams from First league about the participation in Extraliga in the following season.

Scoring system

Until the season 2005/2006 matches could end up with three types of results – win, draw or loss. Giving 3 points in case of victory, 1 point for a draw and 0 points to the defeated team.

Since the 2006/2007 season, there is no match without a winner. Every match (even in main part) has to be decided in a regular time or in the following overtime (5 minutes during main part and 20 minutes in play-off). If the game is still undecided after the overtime, penalty shootout follows. Awarding system is 3 points for a regulation win and 2 points for an overtime (penalty) victory, while the defeated team in an overtime (penalty) gets 1 point. Therefore, three points are given in every match. Hence, there are four possible results – win, win in overtime, loss in overtime and loss.

Main points important for data usage

Rounds for the main part are scheduled in September, but matches might be postponed. Therefore, some matches scheduled for example in tenth round might be played later than matches in the eleventh round.

Basic rules for ice hockey remain the same for every season but some changes are made regularly every four years by Czech Ice Hockey Association (see [1]). Those modifications might have an influence on the game itself, leading to more goals in a match. As some experts argue about the changes in rules starting from season 2014/2015 (read [2]).

Drafts can be made from 1.5. to 31.1. according to a transferring system. But usually take place after season, so that players might have some time to train together during the summer training. Since one team can purchase or sell the best players its form may change dramatically over seasons.

In every season 14 teams compete, but every season can be replaced up to two teams with two best teams from the First League.

All these factors signalize that every season should be treated independently.

4.1 Description of data

The main source for data was the website `http://www.liga.cz/hokej/cesko/`. There are results of matches and average odds on the home team to win, the draw and guest team to win for team sports – football, basketball, handball, volleyball, baseball and most importantly for ice hockey. Data for Extraliga are provided from the season 1998/1999, but many observations of odds (more than a half in the main part) are missing in the first season. Therefore, data used for the analysis are seasons from 1999/2000 to 2014/2015. We limit our analysis to matches of the main part, because the system might behave differently as there are more matches between two teams in a short period of time for playoffs.

Data that we use for the analysis contain 5824 matches in 16 consecutive seasons, 22 different teams with 52 matches within every season. There were few missing observations of average odds (partly completed from website [3]) and rounds in season 2008/2009 were not provided (completed from website [4]). We ended up with only few missing observations of average odds – 5 in the main part and 1 in playoff.

Data provided are not in a ready-to-use format (see Figure 4.1) and had to be transformed. Software used for downloading the data was Microsoft Excel 2007 and its tool Visual Basic for Applications (VBA). Data were downloaded taking advantage of a html table format. The resulted format can be seen in Figure 4.2. For easier data manipulation, we created SQL database (it can be created with scripts on the attached CD) using SQL Server 2014 Management Studio. For further analysis, we used program R version 3.1.2, especially libraries `car`, `MASS`, `fkf` and `lattice`.

[1] `http://www.cslh.cz/text/119-pravidla-ledniho-hokeje.html`

[2] `http://novy.hokej.cz/prinesou-zmeny-pravidel-atraktivnejsi-hru-a-vice-golu/5001879`

[3] `http://www.oddsportal.com/`

[4] `http://www.hokej.cz/tipsport-extraliga`

52. Kolo		1	0	2	
Karlovy Vary - Kometa Brno	3:0	2.78	4.00	2.13	07.03.2014
Kladno - Chomutov	1:4	1.65	4.33	4.18	07.03.2014
Liberec - Zlín	4:3 Náj.	2.42	3.92	2.43	07.03.2014
Plzeň - Mountfield HK	1:3	1.76	4.11	3.77	07.03.2014
Slavia Praha - Litvínov	3:2 Náj.	1.69	4.29	3.97	07.03.2014
Třinec - Vítkovice	3:1	1.63	4.39	4.18	07.03.2014
Sparta Praha - Pardubice	0:3	1.45	4.81	5.62	06.03.2014
51. Kolo		**1**	**0**	**2**	
Chomutov - Plzeň	1:3	6.04	5.29	1.38	04.03.2014
Kometa Brno - Kladno	4:3 ET	1.32	5.53	7.05	04.03.2014
Mountfield HK - Liberec	3:4 Náj.	1.86	4.06	3.37	04.03.2014
Pardubice - Třinec	7:5	2.51	3.99	2.32	04.03.2014
Slavia Praha - Karlovy Vary	9:3	1.46	4.69	5.48	04.03.2014
Vítkovice - Litvínov	3:1	1.94	4.06	3.11	04.03.2014
Zlín - Sparta Praha	2:1	2.44	3.98	2.39	04.03.2014
50. Kolo		**1**	**0**	**2**	
Karlovy Vary - Pardubice	1:4	2.54	3.95	2.32	02.03.2014
Liberec - Vítkovice	3:1	2.17	4.03	2.70	02.03.2014
Litvínov - Kladno	2:3	1.32	5.32	7.29	02.03.2014
Mountfield HK - Chomutov	2:0	1.19	6.79	10.21	02.03.2014
Plzeň - Zlín	1:2	1.88	4.05	3.33	02.03.2014
Slavia Praha - Sparta Praha	3:2 ET	3.67	4.27	1.75	02.03.2014
Třinec - Kometa Brno	10:2	1.79	4.16	3.60	02.03.2014

Figure 4.1: Screenshot of webpage liga.cz (on 21.1.2015).

	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Home Team	Guest Team	Winner	Goals HT	Goals GT	Overtime	Penalty	Win	Tie	Loss	Date	Year	Type	Round
2	Karlovy Vary	Kometa Brno	HT	3	0	0	0	2.78	4	2.13	07.03.2014	2013	Main	52
3	Kladno	Chomutov	GT	1	4	0	0	1.65	4.33	4.18	07.03.2014	2013	Main	52
4	Liberec	Zlín	HT	4	3	0	1	2.42	3.92	2.43	07.03.2014	2013	Main	52
5	Plzeň	Mountfield HK	GT	1	3	0	0	1.76	4.11	3.77	07.03.2014	2013	Main	52
6	Slavia Praha	Litvínov	HT	3	2	0	1	1.69	4.29	3.97	07.03.2014	2013	Main	52
7	Třinec	Vítkovice	HT	3	1	0	0	1.63	4.39	4.18	07.03.2014	2013	Main	52
8	Sparta Praha	Pardubice	GT	0	3	0	0	1.45	4.81	5.62	06.03.2014	2013	Main	52
9	Chomutov	Plzeň	GT	1	3	0	0	6.04	5.29	1.38	04.03.2014	2013	Main	51
10	Kometa Brno	Kladno	HT	4	3	1	0	1.32	5.53	7.05	04.03.2014	2013	Main	51
11	Mountfield HK	Liberec	GT	3	4	0	1	1.86	4.06	3.37	04.03.2014	2013	Main	51
12	Pardubice	Třinec	HT	7	5	0	0	2.51	3.99	2.32	04.03.2014	2013	Main	51
13	Slavia Praha	Karlovy Vary	HT	9	3	0	0	1.46	4.69	5.48	04.03.2014	2013	Main	51
14	Vítkovice	Litvínov	HT	3	1	0	0	1.94	4.06	3.11	04.03.2014	2013	Main	51
15	Zlín	Sparta Praha	HT	2	1	0	0	2.44	3.98	2.39	04.03.2014	2013	Main	51
16	Karlovy Vary	Pardubice	GT	1	4	0	0	2.54	3.95	2.32	02.03.2014	2013	Main	50
17	Liberec	Vítkovice	HT	3	1	0	0	2.17	4.03	2.7	02.03.2014	2013	Main	50
18	Litvínov	Kladno	GT	2	3	0	0	1.32	5.32	7.29	02.03.2014	2013	Main	50
19	Mountfield HK	Chomutov	HT	2	0	0	0	1.19	6.79	10.21	02.03.2014	2013	Main	50
20	Plzeň	Zlín	GT	1	2	0	0	1.88	4.05	3.33	02.03.2014	2013	Main	50
21	Slavia Praha	Sparta Praha	HT	3	2	1	0	3.67	4.27	1.75	02.03.2014	2013	Main	50
22	Třinec	Kometa Brno	GT	10	2	0	0	1.79	4.16	3.6	02.03.2014	2013	Main	50

Figure 4.2: Screenshot of downloaded data from liga.cz (on 21.1.2015).

The given information for each match is

- home team,
- guest team,
- result – goals of home team, goals of guest team, identifier of overtime and identifier of penalty shootout,
- average odds for a win of home team (home win),
- average odds for a draw,
- average odds for a loss of home team (away win),
- date of the match,
- round,
- season.

Average odds are calculated as average from available odds given by typically more than 20 betting companies. As stated in section 4 there are no matches without a winner since season 2006/2007. To treat the data uniformly, we might want to omit the information about matches after regular time. However, the additional information might be useful. Hence, the goal after regular time is counted as a half goal. This corresponds with the fact that the team shot one extra goal but not in a regular time.

4.2 Determining odds by betting companies

Assume that a betting company sets odds for an event. This event has a certain number of possible outcomes, e.g. In a hockey match the outcome (in regular time) is either home win, draw or away win. The goal is to find appropriate odds such that the betting company would receive expected predetermined margin. There are provided decimal odds (see Figure 4.1) in our data. It means that in case of a successful bet, the pay-off is obtained from the money bet multiplied by the decimal odds. For example if odds on the home win are 2.2 and we bet 100 then we get 220 in case of home win and 0 otherwise. A proposition how to calculate odds using estimates of probabilities and vice versa, will be presented after the introduction of notation.

Notation

- Y denotes a random outcome with possible values $i \in \{1, \ldots, n\}$,
- p_i denotes the probability of outcome i, i.e. $p_i = \mathsf{P}(Y = i)$,
- r_i denotes odds for an outcome i of the event Y, $r_i \in (1, \infty)$,
- B_i denotes absolute amount of money bet on an outcome i, $B_i \in [0, \infty]$, b_i denotes relative amount of money bet on an outcome i, $b_i \in [0, 1]$,
- M_i denotes percentage margin of the betting company for bets i.

By bold symbols we denote the whole vector, e.g. $\boldsymbol{p} = (p_1, \ldots, p_n)^\top$. Using the above notation, it must hold that sum of probabilities p_i and relative amount of money bet b_i are equal to 1, i.e.

$$\sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} b_i = 1.$$

The task of betting companies are to maximize their profit. The profit depends on odds (r_i), amount of money bet (B_i) and result of the event Y. The odds have to be set at first and then people start to bet. It is intuitive that B_i depends strongly on the odds provided. Higher rates will lead to higher amount of money bet.

The maximization problem that maximises the profit of the betting company is as follows:

$$\begin{aligned} &\max_{\boldsymbol{r}} \sum_{i=1}^{n} B_i(\boldsymbol{r}) - \sum_{i=1}^{n} r_i B_i(\boldsymbol{r}) \mathbb{1}_{[Y=i]}, \\ &s.t.\ r_i > 1, \forall i = 1, \ldots, n. \end{aligned} \tag{4.1}$$

We have stressed the dependence of money bet B_i on a vector of odds $\boldsymbol{r}$. We assume that $B_i(\boldsymbol{r})$ is some non-negative and non-decreasing function in every component r_i. The maximization problem (4.1) is a stochastic problem with a random variable $\mathbb{1}_{[Y=i]}$ in the objective function. To solve such a problem needs transformation into a deterministic problem. One possibility is to maximize the expectation of the objective function. Then it simplifies to:

$$\begin{aligned} &\max_{\boldsymbol{r}} \sum_{i=1}^{n} B_i(\boldsymbol{r})(1 - r_i p_i), \\ &s.t.\ r_i > 1, \forall i = 1, \ldots, n. \end{aligned} \tag{4.2}$$

To solve the problem (4.2) we have to find the deterministic function $B_i(\boldsymbol{r})$. It might be done by assuming a suitable functional form (e.g. parametric) and using the historical data for odds $\boldsymbol{r}$ and amount of money bet $\boldsymbol{B}$ to estimate it. Data about $\boldsymbol{B}$ are not accessible, so we do not investigate this problem further. We just remark that the problem might be formulated as a multistage problem because the betting company can change odds every day (but sustain the odds for already given bets).

The preceding approach is complicated and assumes the knowledge (or at least good estimation) of $B_i(\boldsymbol{r})$, therefore, we present another one that will enable us to compute result probabilities for known odds and vice versa. The following approach is based on the idea that the betting company estimates probabilities of an outcome i and then uses some predetermined margin to set r_i. Suppose that the relative expected margin is set to M, i.e.

$$M = 1 - \sum_{i=1}^{n} r_i b_i p_i = \sum_{i=1}^{n} b_i (1 - r_i p_i).$$

The following lemma gives a formula for r_i such that the relative expected margin is fixed for any realizations of bets. We want to find a solution (for the vector $\boldsymbol{r}$, $r_i > 1$) such that the preceding equation holds $\forall\ b_i \geq 0, \sum_{i=1}^{n} b_i = 1$.

Lemma 13. *The following two statements are equivalent*

(i) $\forall\ b_i \geq 0, \sum_{i=1}^{n} b_i = 1 : M = \sum_{i=1}^{n} b_i(1 - r_i p_i)$

(ii) $\forall\ i = 1, \ldots, n : r_i = \frac{1-M}{p_i}$.

Proof. Both implications are straightforward calculations.

$(i) \Rightarrow (ii)$: We set $b_i = 1$ and $b_j = 0$, $i \neq j$ then $M = 1 - r_i p_i$, so $r_i = \dfrac{1-M}{p_i}$.

$(ii) \Rightarrow (i)$: We plug it into the formula for margin: $\displaystyle\sum_{i=1}^{n} b_i \left(1 - \frac{1-M}{p_i} p_i\right) = M.$

□

The preceding lemma gives an answer how to set rates with probabilities estimates to yield predetermined expected margin. If the betting company knew the demand functions $B_i(\boldsymbol{r})$, they could even set the realized relative margin to be fixed. The solution is to solve system of equations $r_i = \frac{1-M_c}{b_i(\boldsymbol{r})}$, $i = 1, \ldots, n$. The relative realized margin would be then $1 - r_i b_i = M_c$. The more realistic situation is reversed; the better might want to have profit without risk (arbitrage). It is only possible when M_c is negative and

$$b_i = \frac{1 - M_c}{r_i}, \quad M_c = 1 - \frac{1}{\sum_{i=1}^{n} \frac{1}{r_i}}.$$

We have calculated M_c based on the fact that $\sum_{i=1}^{n} b_i = 1$. Even though the value M_c has clear interpretation the following formula is used in praxis:

$$M_p = \left(\sum_{i=1}^{n} \frac{1}{r_i}\right) M_c = \sum_{i=1}^{n} \frac{1}{r_i} - 1. \tag{4.3}$$

The value M_p is commonly called *bookmaker's margin*. It holds that negative M_p means arbitrage opportunity and $M_p > M_c$ in case of non-arbitrage opportunity.

From the lemma 13 it can be seen that a fair bet (zero margin) would lead to $r_i = \frac{1}{p_i}$. Therefore, another approach for assessing rates could be that the betting company firstly estimates the probabilities of outcome then calculates fair odds and adjust them according to their value (small odds slightly and higher odds substantially). The resulted odds might be then

$$r_i = \frac{1}{p_i} - K_i. \tag{4.4}$$

The true expected margin for (4.4) is $M = \sum_{i=1}^{n} 1 - p_i(\frac{1}{p_i} - K_i) = \sum_{i=1}^{n} K_i$.

Both suggested formulas for rates are highly dependent on the value p_i. For fixed K_i or M they both go to infinity as $p_i \to 0+$ and are smaller than one as $p_i \to 1-$. Hence, a more robust method might be appropriate. Logical constraints are increasing in p_i and predetermined bounds for the smallest and the largest values. The smallest rate should be one for $p_i = 1$ and largest fixed. An estimate that suits those constraints is

$$r_i = \frac{1 + K_i}{p_i + K_i}. \tag{4.5}$$

Match	Odds					Probabilities		
	r_w	r_t	r_l	M_p	K	p_w	p_t	p_l
Kladno – Chomutov	1.65	4.33	4.18	0.076	0.040	0.59	0.20	0.21
Liberec – Zlín	2.42	3.92	2.43	0.080	0.042	0.39	0.22	0.39
Plzeň – Mountfield HK	1.76	4.11	3.77	0.077	0.040	0.55	0.21	0.24
Slavia Praha – Litvínov	1.69	4.29	3.97	0.077	0.040	0.58	0.20	0.22
Třinec – Vítkovice	1.63	4.39	4.18	0.081	0.042	0.60	0.20	0.21
Sparta Praha – Pardubice	1.45	4.81	5.62	0.075	0.039	0.68	0.18	0.15

Table 4.1: Examples of transformed rates into normalized probabilities according to (4.6) for some matches of 52th round in season 2013/2014 (see Figure 4.1).

The true expect margin for (4.5) is $M = \sum_{i=1}^{n} 1 - p_i(\frac{1+K_i}{p_i+K_i}) = \sum_{i=1}^{n} \frac{K_i(1-p_i)}{p_i+K_i}$. The relative margin is between 0 and 1 and it increases as p_i decreases. To use the formula we need to set the value K_i.

For our purposes we need to estimate probabilities from odds. We use the formula (4.5) and invert it. Further, we assume that K_i is same for every possibility i, so transformed probabilities are:

$$p_i = 1 - \left(1 - \frac{1}{r_i}\right)(1 + K), \quad i \in \{1, \ldots, n\}. \tag{4.6}$$

We can find the value K from the constrain $\sum_{i=1}^{n} p_i = 1$. We simply derive that

$$K = \frac{M_p}{n - 1 - M_p}.$$

We use this relation to estimate probabilities of a plain result in Extraliga. In that case we have $n = 2$, so $K = \frac{M_p}{2-M_p}$. We provide some examples of rates, bookmaker's margin and corresponding probabilities in Table 4.1.

4.3 Datasets for paired comparisons

Before we start identifying significant predictors, we discuss options of working with a dataset of paired comparisons. Basically, there are two options:

- **team perspective** – we want to easily work with results of a particular team. So, we stack matches of all teams together adding identifier of home advantage.
- **match perspective** – we look at matches as events with two teams competing where the first of the ordered pair plays at home field.

Apparently, the first approach contains all matches twice and clearly violates the assumption of independent observations. However, it is easier to work with. The following theorem shows that parameter estimates remain the same in either linear regression or ordinal regression in case of a symmetrical design matrix.

Theorem 14. *Suppose we observe* $(W_i, Y_i, \boldsymbol{x}_i)^\top, i = 1, \ldots, n$ *independent where* W_i *is a continuous variable,* Y_i *is a categorical variable with categories* $j = 0, 1, 2$ *and* $\boldsymbol{x}_i = (1, \widetilde{\boldsymbol{x}}_i)$ *is an ith row of a design matrix* $\mathbb{X}$ *with full rank.*

We define two datasets

(i) (TP) $\left(\begin{pmatrix} \boldsymbol{W} \\ -\boldsymbol{W} \end{pmatrix}, \begin{pmatrix} \mathbb{X} \\ -\mathbb{X} \end{pmatrix}\right)$ *or* $\left(\begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{2} - \boldsymbol{Y} \end{pmatrix}, \begin{pmatrix} \mathbb{X} \\ -\mathbb{X} \end{pmatrix}\right)$,

(i) (MP) $(\boldsymbol{W}, \mathbb{X})$ *or* $(\boldsymbol{Y}, \widetilde{\mathbb{X}})$.

Then using either (MP) or (TP) yields the same OLS estimates in linear regression for a continuous variable and MLE estimates in ordinal regression using the cumulative model (3.13) *in case of unique* MLE *estimates.*

Proof. We start with a continuous variable. We calculate residual sum of squares for both datasets

$$RSS_{MP}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (W_i - \boldsymbol{x}_i\boldsymbol{\beta})^2, \tag{4.7}$$

$$RSS_{TP}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (W_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 + \sum_{i=1}^{n} (-W_i + \boldsymbol{x}_i\boldsymbol{\beta})^2 = 2RSS_{MP}(\boldsymbol{\beta}). \tag{4.8}$$

From the above we see that RSS is minimized in the same point and hence, parameter estimates are the same. Notice that RSS in (TP) is twice bigger.

In case of an ordinal variable, firstly notice that there is one extra parameter in case of (TP). The parameters we denote $(\theta_0, \theta_1, \boldsymbol{\beta}^\top)^\top$ in case of (MP) and $(\gamma_0, \gamma_1, HA, \boldsymbol{\beta}^\top)^\top$ for (TP) with the same meaning as in model (3.13). Then for the log-likelihood of (TP) we have

$$\ell_{TP}(\gamma_0, \gamma_1, HA, \boldsymbol{\beta}^\top) = \ell^1_{TP}(\gamma_0, \gamma_1, HA, \boldsymbol{\beta}^\top) + \ell^2_{TP}(\gamma_0, \gamma_1, HA, \boldsymbol{\beta}^\top).$$

We have divided the log-likelihood into two parts – the first part with data $(\boldsymbol{Y}, \mathbb{X})$ and the second with $(\boldsymbol{2} - \boldsymbol{Y}, -\mathbb{X})$. We denote the estimated coefficients for (MP) as $(\widehat{\theta}_0, \widehat{\theta}_1, \boldsymbol{\beta}^\top)^\top$. The key point is to realize that for both log-likelihoods the maximum is reached for $\widehat{\gamma}_1 = -\widehat{\gamma}_0 = \frac{1}{2}(\widehat{\theta}_1 - \widehat{\theta}_0)$, $\widehat{HA} = \frac{1}{2}(\widehat{\theta}_0 + \widehat{\theta}_1)$ and $\widehat{\boldsymbol{\beta}}$. Hence those estimations maximize $\ell_{TP}(\gamma_0, \gamma_1, HA, \boldsymbol{\beta}^\top)$.

To show the other implication, we use the symmetry of the dataset and the property (3.11) of a distribution function F, which ensures that log-likelihood ℓ_{TP} is maximized when $\widehat{\gamma}_1 = -\widehat{\gamma}_0$. Then for the likelihoods hold

$$\ell_{TP}(-\widehat{\gamma}_1, \widehat{\gamma}_1, \widehat{HA}, \widehat{\boldsymbol{\beta}}^\top) = 2\ell^1_{TP}(-\widehat{\gamma}_1, \widehat{\gamma}_1, \widehat{HA}, \widehat{\boldsymbol{\beta}}^\top) = 2\ell_{MP}(\widehat{HA} - \widehat{\gamma}_1, \widehat{HA} + \widehat{\gamma}_1, \widehat{\boldsymbol{\beta}}^\top).$$

We have used that ℓ^1_{TP} and ℓ_{MP} correspond to the same data but only in different parametrization. The uniqueness of MLE implies that the maximum is reached only in one point. □

Note. Remark that in the proof was shown that $RSS_{TP} = 2RSS_{MP}$ but model TP does not have an intercept. Therefore, the coefficient of determination for MP is smaller as the total sum of square is calculated with an intercept.

In case of TP standard errors for coefficients are underestimated, which leads to smaller p-values and higher (misleading) significance of coefficients.

The uniqueness of MLE in case of cumulative logit and probit model can be found in Burridge (1981).

In the following we refer to a dataset with every match only once as MP and TP for a dataset with every observation twice (once viewed from home team perspective and once from guest team perspective). In the following we use both datasets, MP is preferred for evaluating significance of predictors and for forecasting outcomes of matches. TP is used to evaluate fit of the model because it does account for the explained variability due to home advantage. In MP the home advantage is expressed in the intercept, so the correct estimation of HA with no other regressors yields $R^2 = 0$, which is not intuitive. The same principle we use for ordinal regression – MP is used for fitting the model and Gini coefficient is calculated for TP.

4.4 Suitable transformation of match outcome

The aim of this section is to find suitable one dimensional variable that carries the most information of an outcome of a match. The outcome is represented by goals of home team and goals of guest team. Typically one uses goal differences for the convenience of taking real values and the possibility to use normal distribution as its approximation (see for example Glickman, 1993).

We deal with the problem more analytically and pose a question if the probabilities of outcome would be better estimated knowing goal difference, goal ratio, something in between or just the plain result (Win/Draw or Loss). Every of those possibilities have some reasoning. Goal differences put higher probability of win to matches with higher goal difference but one might argue that result 4:3 was more balanced than 2:1, which would support to use goal ratios. As stated before, the goal in overtime or penalty shooting is taken as a half goal.

To decide what function is better to use, we take advantage of knowing betting odds. We assume that they are relevant estimates of probabilities of outcomes. We use dataset TP and transform odds into terms of probability by using (4.6) for win and loss. We normalize the estimated probabilities and transform them to take real values using logit. The probabilities and the dependent variable are

$$\begin{gathered} p_t^w = 1 - \left(1 - \frac{1}{r_t^w}\right)(1 + K_t), \quad p_t^l = 1 - \left(1 - \frac{1}{r_t^w}\right)(1 + K_t), \\ K_t = \frac{M_p^i}{2 - M_p^t}, \quad M_p^t = \frac{1}{r_t^w} + \frac{1}{r_t^t} + \frac{1}{r_t^l} - 1, \\ Y_t = \operatorname{logit}\left(\frac{p_t^w}{p_t^w + p_t^l}\right). \end{gathered} \tag{4.9}$$

As explanatory variables we consider members of the following parametric family

$$\mathcal{F}(\theta) = \left\{ \frac{g_A + \theta}{g_B + \theta}; \quad \theta \in \mathbb{R}^+ \right\} \tag{4.10}$$

where g_A is a number of goals scored by team A and g_B is a number of goals scored by its opponent.

To decide what is the best explanatory variable we simply run linear regressions and calculate its coefficient of determination $R^2(\theta)$ or residual sum of squares

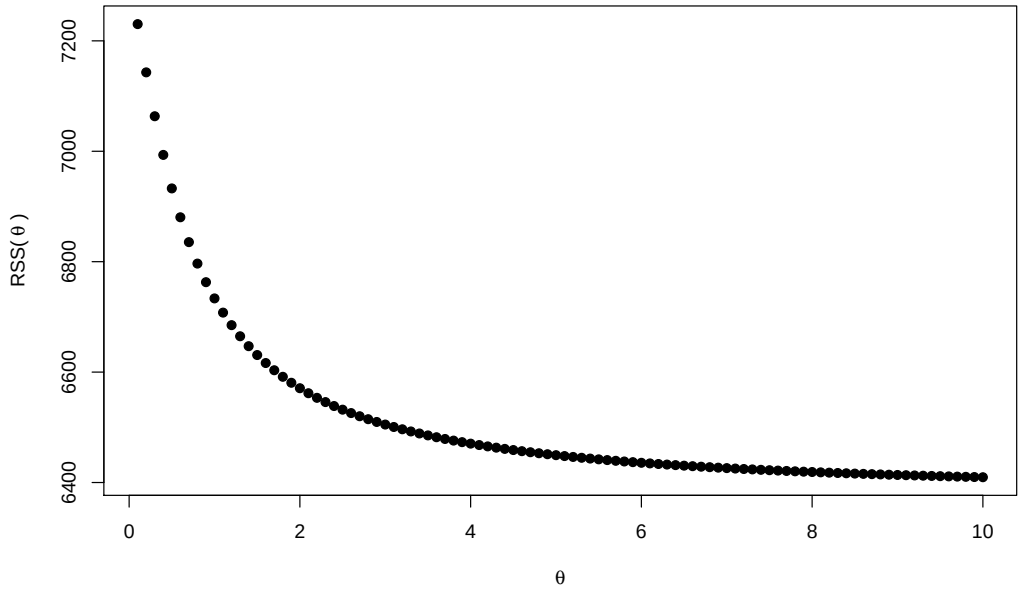


Figure 4.3: Illustration of the minimization problem (4.11) where θ defines an explanatory variable and $RSS(\theta)$ is residual sum of square of linear regression with dependent variable Y_t (transformed probability of win).

$RSS(\theta)$. Our aim si to maximize $R^2(\theta)$ (or equivalently minimize $RSS(\theta)$) on open set $\Theta = (0, D)$ for D going to infinity. The minimization has form:

$$\min_{\theta \in \Theta} \sum_{t=1}^{n} (y_t - \widehat{\alpha} - \widehat{\beta} x_t(\theta))^2 \tag{4.11}$$

where $\widehat{\alpha}$ and $\widehat{\beta}$ are OLS estimates for fixed θ.

The higher values of θ the lower difference in explanatory variables. To see this it is better to rewrite x as follows

$$x = \frac{g_A + \theta}{g_B + \theta} = \frac{g_A - g_B + g_B + \theta}{g_B + \theta} = 1 + \frac{g_A - g_B}{g_B + \theta} = 1 + c(\theta)(g_A - g_B). \tag{4.12}$$

Further, we just realize that changing the explanatory variable by adding a constant or multiplying it by a constant does not effect the quality of fit in terms of R^2 and $c(\theta)$ changes imperceptibly for high values of θ. This also explains why we choose parametric family (4.10). It is now apparent that it contains ratio of goals for limiting case $\theta \to 0_+$ and goal differences for $\theta \to +\infty$.

We solve the minimization problem (4.11) using function `optimize` in programming language R for $D = 10000$ with the result that optimal value is $\theta_{opt} = 166.56$ and $RSS(\theta_{opt}) = 6386.69$. If we run the linear regression for the limiting case using goal differences, we receive that $RSS = 6386.88$, hence the difference is very small. Those results are not surprising if we look at illustrative figure of values of the objective function in Figure 4.3.

Another possibility to use would be a categorical variable with three categories – win, draw or loss. If we calculate RSS in this case we receive $RSS = 6575.00$,

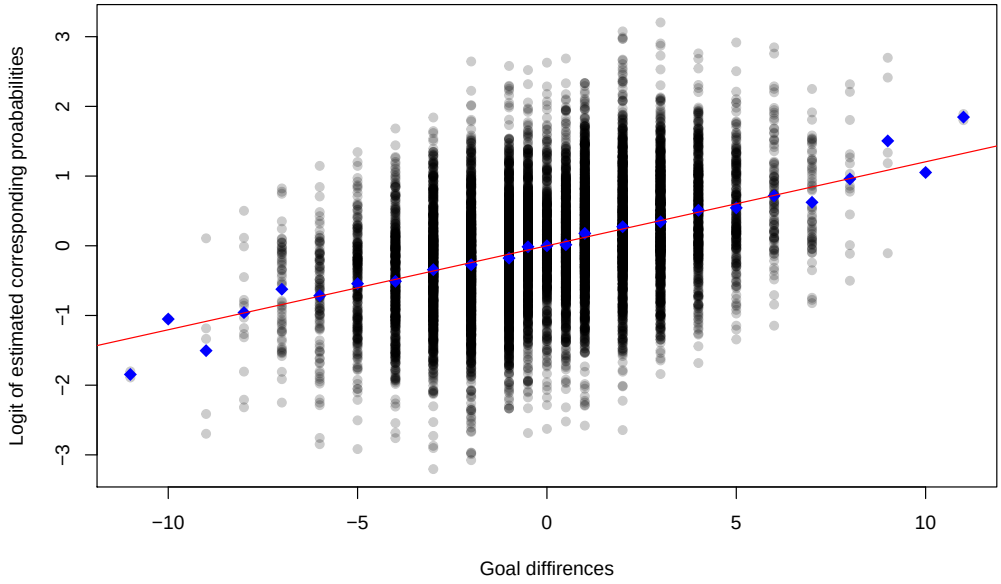


Figure 4.4: Linear regression with dependent variable Y_t (transformed probability of win) and explanatory variable goal difference. Blue points are conditional means and red line is a regression line of the corresponding model.

which is substantially higher than RSS for goal differences. Therefore, in terms of a quadratic loss function the best transformation to use is close to goal difference. In this case the relationship seems to be linear (see Figure 4.4) but yields rather small coefficient of determination $R^2 = 0.148$, which suggests hard predictability of outcomes in Extraliga.

The conclusion of this section is that goal differences are appropriate to carry the information of probabilities of win, draw or loss and will be used in the following section as dependent variable to identify possible predictors.

4.5 Home advantage

The most discussed factor in most sporting events is home advantage (HA). This effect might be a result of four principles: no need to travel, familiarization with the stadium, crowd support or influenced referee decisions due to the crowd support. To decide that home advantage in Extraliga exists is enough to see that 51 % of all matches were won by home team whereas only 28 % by guest team in a regular time period. However, in this section we show that home advantage is not significantly team specific but it is decreasing in time and increasing within a season.

To measure home advantage for a particular team we determine the difference in number of goals scored in a match at home stadium and out (GDInOut). It oscillates between 0 and 3 (see upper Figure 4.5), it was negative only in 5 out of 224 cases. No team seems to play exceptionally better/worse at home stadium than others.

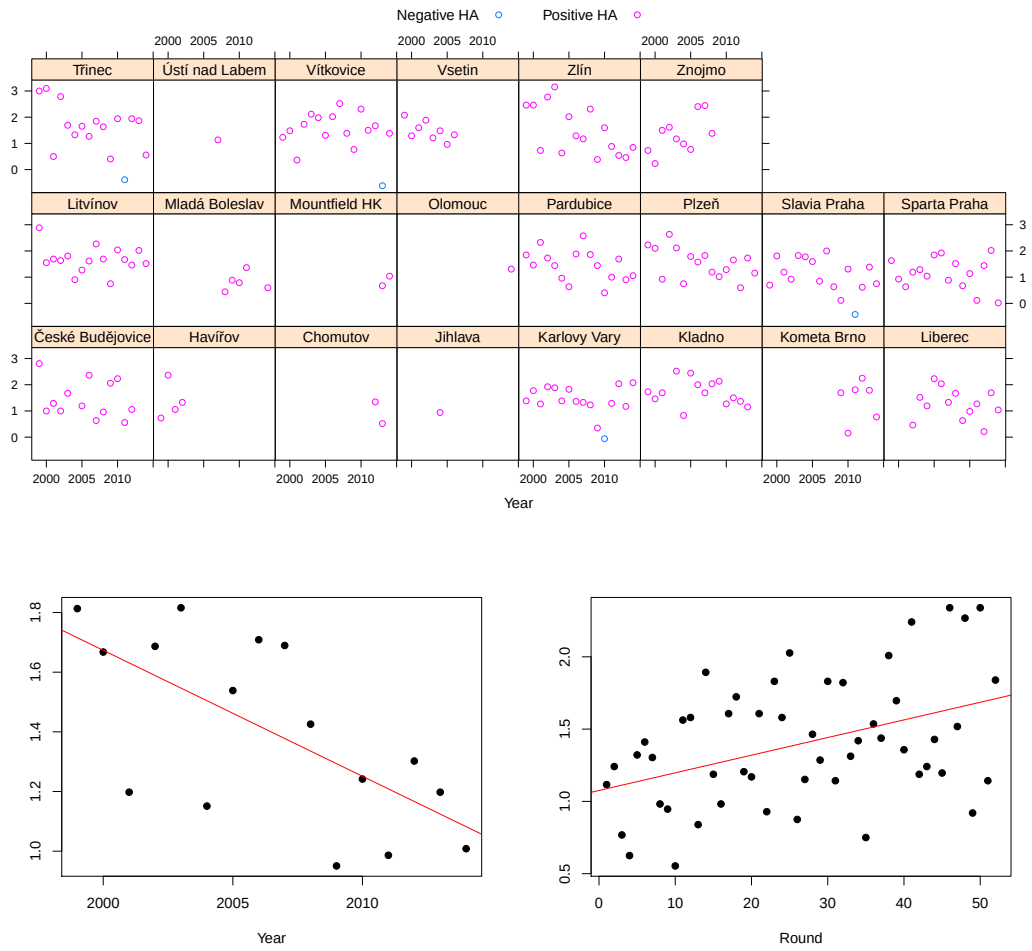


Figure 4.5: Mean goal differences for matches played at home and out. They are depicted for every season and every team in the upper figure and by year and by round with a regression line in the lower figures.

Our data come from 16 consecutive seasons. If we calculate average GDInOut within those seasons we observe significant decrease (see lower left Figure 4.5). This decrease might be a result of several effects – better comfort during travelling, higher number of devoted fans that travel with their team or that players change teams more often and are accustomed to higher number of different stadiums than before. One might ask if this is not a side effect of decreasing number of goals in a match but the average number of goals by one team in a match seems almost constant within years with mean 2.771.

Every season there are 52 rounds that are scheduled before the season starts. Some rounds might be postponed but it happens only exceptionally (in 3.6 % cases). It is obvious that average GDInOut increases within season (see lower right Figure 4.5). Reasons might be that tiredness accumulates during the season and the need of travelling plays higher role or that more fans come to matches during winter months than during autumn. We can separate the effect by calculating average number of goals at home and out in rounds. The increasing effect of HA is due to increasing number of goals scored by home team rather than decreasing number of goals scored by guest team (see Figure A1).

Effect	Sum of Squares	Df	F-value	p-value
Season	46.4	1	8.548	0.003
Round	43.1	1	7.936	0.005
Team HA	124.7	21	1.094	0.346
Forms	3731.5	208	3.305	< 0.001

Table 4.2: Analysis of variance (Type II tests)

To test whether the hypothes is based on exploratory analysis are statistically significant or not, we run linear regression. Every observation is determined by three indexes i team, j its opponent and t identifier of the match between i and j.
Variables that we use throughout this chapter are

- GD_{ijt} - goals of home team i minus goals of guest team j for match t,
- HA_{ijt} - one if team i plays at home, minus one otherwise,
- S_{ijt} - rank of season, starting with 0 for season 1999/2000 and ending with 15 for season 2014/2015,
- R_{ijt} - round for the match minus one (0-51).

We denote $I = 22$ total number of teams within all seasons, n_{ij} total number of matches between teams i and j and m_{ij} total number of matches between teams i and j with home team i. It has to hold that $m_{ij} = m_{ji}$, $\forall i, j \in I$ because there are always four matches between two teams within season with changing home advantage. If we want to use dataset MP, we simply use only data with $HA_{ijt} = 1$.

Regression that we fit contains variables season, round and team (expressed by parameter δ_i). We also add difference of estimated forms (season specific) for team and its opponent to filter the effect of forms. The model has the following form:

$$GD_{ijt} = \beta_0 + \beta_1 S_{ijt} + \beta_2 R_{ijt} + \delta_i + \sum_{r=0}^{15} (\alpha_{ir} - \alpha_{jr}) \mathbb{1}_{[S_{ijt}=r]} + \varepsilon_{ijt}, \\ i, j = 1, \ldots, I,\ t = 1, \ldots, m_{ij}. \tag{4.13}$$

We have to add constraints on parameters δ_i and α_{ir}. For δ_i we set $\delta_1 = 0$ (treatment contrasts) and for performances we use the following

$$\sum_{i=1}^{n} \alpha_{ir} = 0,\ r = 1, \ldots, 15. \tag{4.14}$$

In that model home advantage is expressed in the intercept and therefore the coefficient of season and round can be viewed as changes to the intercept. To determine whether coefficients δ_i, $i = 1, \ldots, I$ are significantly non-zero we run anova F-test. The effect of HA of team has p-value greater than 0.01 (see Table 4.2), which will be our confidence level, hence it yields evidence against its significance. But the effect of round and of season are significant (see Table 4.2).

Effect	Estimate	Std. error	T-value	p-value
HA	0.711	0.078	9.10	< 0.001
Season	-0.021	0.007	-3.19	0.001
Round	0.006	0.002	2.82	0.005

Table 4.3: OLS estimates from model (4.15) of variables of interest.

We can leave out the variable team and build a model without it to assess the estimates. For comparison we provide the same model in MP:

$$GD_{ijt} = \beta_0 + \beta_1 S_{ijt} + \beta_2 R_{ijt} + \sum_{r=0}^{15} (\alpha_{ir} - \alpha_{jr}) \mathbb{1}_{[S_{ijt}=r]} + \varepsilon_{ijt},\ i, j = 1, \ldots, I, t = 1, \ldots, m_{ij}. \tag{4.15}$$

and in TP:

$$GD_{ijt} = \beta_0 HA_{ijt} + \beta_1 HA_{ijt} S_{ijt} + \beta_2 HA_{ijt} R_{ijt} + \sum_{r=0}^{15} (\alpha_{ir} - \alpha_{jr}) \mathbb{1}_{[S_{ijt}=r]} + \varepsilon_{ijt},\ i, j = 1, \ldots, I, t = 1, \ldots, n_{ij}. \tag{4.16}$$

Due to the symmetry of TP both models yield the same coefficient estimates (see theorem 14). HA in the first round and in the season 1999/2000 is estimated as 0.71 goals (or more naturally the estimated goal difference between match on the home field and out against the same opponent is 1.42 goals, see Table 4.3). In every season this effect decreases by -0.02. It means that in the season 2014/2015 in the first round is estimated as only 0.40. But HA increases during the season, at the end of the season, it is higher of 0.298 compared to the beginning (see Table 4.3). Coefficient of determination $R^2 = 0.204$ (for TP) is rather low as expected from the results of the previous section.

We provide some graphical diagnostics of residuals (see Figure 4.6). The fit does not seem to be systematically biased (upper left) but variance seems to be bigger for higher values of fitted values (upper right). Normality of goal differences seems to be realistic assumption, it might have slightly heavier tails. It is not straightforward to assess the information of autocorrelation as in our dataset multiple matches are at the same time and autocorrelation makes sense only for time dependence for a specific team. This is a content of the next section as it represents widely discussed phenomenon (high number of occurrences of win series for certain team would suggest positive autocorrelation in goal differences).

We can conclude that HA is significant effect that must be included in any model forecasting the outcome of a match. We have also seen that this effect should be adjusted for every season and also during the season. On the other hand HA does not seem to be team specific, so one can consider home advantage common for all teams.

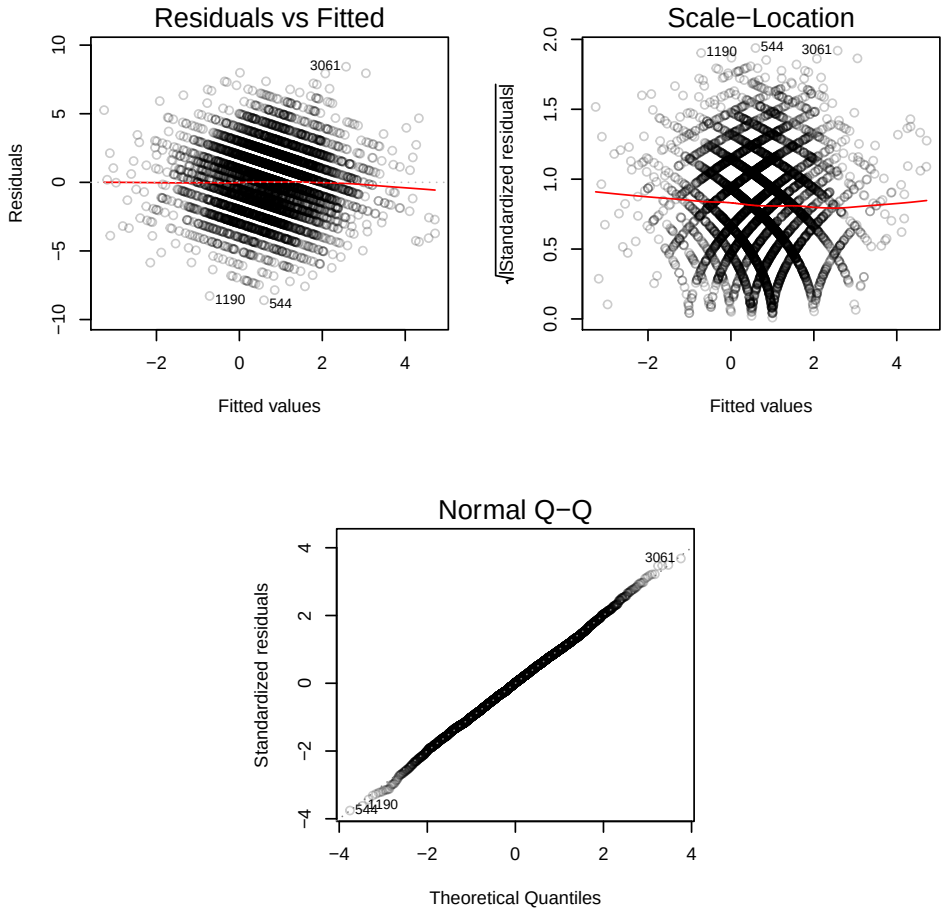


Figure 4.6: Residual diagnostics of regression model (4.15)

4.6 Autocorrelation of results

In this section we want to examine if there is positive or negative effect of result from the previous match. In newspapers and on the internet we can come across headlines stating that series of wins for particular team continues, five wins in a row, etc. The question is if this is randomness, only effect of a good team form or if there is positive effect of previous win, which would mean positive autocorrelation of results and goal differences. Because we need to work with every observation for each team, we use dataset TP.

If we calculate percentage proportion of results after won matches, we get win in 47.9 % and loss in 47.2 % cases, which does not suggest strong evidence for positive autocorrelation. Furthermore, the proportion after loss is 47.5 % wins and 48.4 % losses (reason that it does not sum up to 100 % is that till season 2006/2007 some matches ended as a draw). This suggests rather small positive autocorrelation but by examining the data more carefully, we can conclude that each team typically plays one match on home field and the other out. To graph-

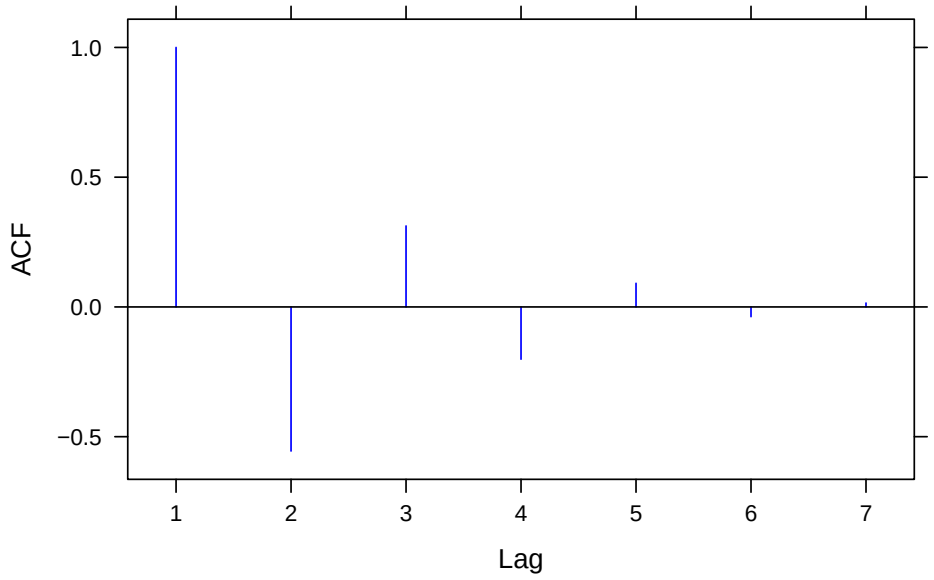


Figure 4.7: Average ACF for home advantage through seasons 1999/2000 till 2014/2015

ically assess this fact, we have calculated ACF of home advantage for each team during every season and computed averages of all estimated correlations (see Figure 4.7). ACF shows clear negative autocorrelation, so we have to filter the effect of home advantage. Therefore we distinguish where the team played at specified time t and count transition probabilities separately (see Table 4.4). We clearly see that if the team played home it is more likely to loose in the next match (effect of negative autocorrelation of HA). We also observe that win in next match is more probable in case that team won (in both cases out and home).

		Result_{t+1}		
HA_t	Result_t	Loss	Draw	Win
Out	Loss	0.424	0.045	0.532
	Draw	0.330	0.167	0.503
	Win	0.390	0.049	0.561
Home	Loss	0.625	0.041	0.334
	Draw	0.560	0.113	0.326
	Win	0.572	0.051	0.377

Table 4.4: Contingency table for plain results at consecutive matches conditionally on home advantage at first match (row percentages). We use results after overtime or penalty shootout if there were some.

Effect	Estimate	Std. error	T-value	p-value
Goal difference of previous match	-0.007	0.009	-0.781	0.435

Table 4.5: OLS estimate from model (4.17) of variable of interest.

It might seem to be enough construct a log-linear model with effect of home advantage and result (denoted as Y) at current time (indexed as 1) and the following (indexed as 2) in form:

$$\log(\mathsf{E}\ n_{ijkl}) = (Y_1 * HA_1)_{ij} + (Y_2 * HA_2)_{kl} + (HA_1 * HA_2)_{jl} + (Y_1 * Y_2)_{ik}$$

where by symbol $(Y_1 * HA_1)_{ij}$ we mean both main effects and interaction (i.e. $(Y_1 * HA_1)_{ij} = \beta_i^{Y_1} + \beta_j^{Y_1} + \beta_{ij}^{HA_1Y_1}$) and test the significance of interaction of results ($\beta_{ik}^{Y_1Y_2}$, $i, k = 0, 1, 2$). Using deviance test yields p-value smaller than 0.01, so we would reject independence of consecutive results. Odds to win in the following match would be estimated as $\exp(\widehat{\beta}_{22}^{Y_1Y_2}) = 1.25$ times higher in case of current victory than in case of current loss.

However, one has to realize that there is another lurking effect and that is the form of the team. It will naturally influence both results current and the following. If we consider categorical variables for each team and season, there are not enough observations for using techniques of log-linear tables. Hence, we have to present another possibility of testing. We expand model (4.16) from the previous section by adding the information of goal difference from the previous match for team i (denoted as $GD_{i.t-1}$, we use dot notation to denote the opponent from previous match) and h is opponent j. Therefore, we use again goal differences instead of direct result to assess the effect of previous goal difference. We do not observe the effect by itself but its mixture of team i and team j, so we use difference of previous matches. We also add difference of HA from the previous match to filter the effect of changing home advantage. It means the model has form (in MP):

$$\begin{aligned} GD_{ijt} = \beta_0 + \beta_1 S_{ijt} + \beta_2 R_{ijt} + \sum_{r=0}^{15} (\alpha_{ir} - \alpha_{jr}) \mathbb{1}_{[S_{ijt}=r]} + \beta_3 (GD_{i.t-1} - GD_{j.t-1}) \\ + \beta_4 (HA_{i.t-1} - HA_{j.t-1}) + \varepsilon_{ijt}, \quad i, j = 1, \ldots, I, t = 1, \ldots, m_{ij}. \end{aligned} \tag{4.17}$$

The estimated coefficient β_3 is not significantly different from zero at confidence level 0.01 (see Table 4.5).

To sum up, the effect of the previous result is not significant for estimating outcome and will not be included in forecasting model. We should also remark that main advantage of using DLM or DGLM is accounting for autocorrelation of dependent variable. Therefore, we provide ACF for goal differences in similar logic as for home advantage earlier. We calculate ACF for goal differences for every team and season and afterwards calculate mean of each estimated correlation with lag up to 7 (see Figure 4.8). The estimated correlation are not large and the first is negative (effect of changing home advantage as discussed before). This result suggests not so strong dynamics in the structure.

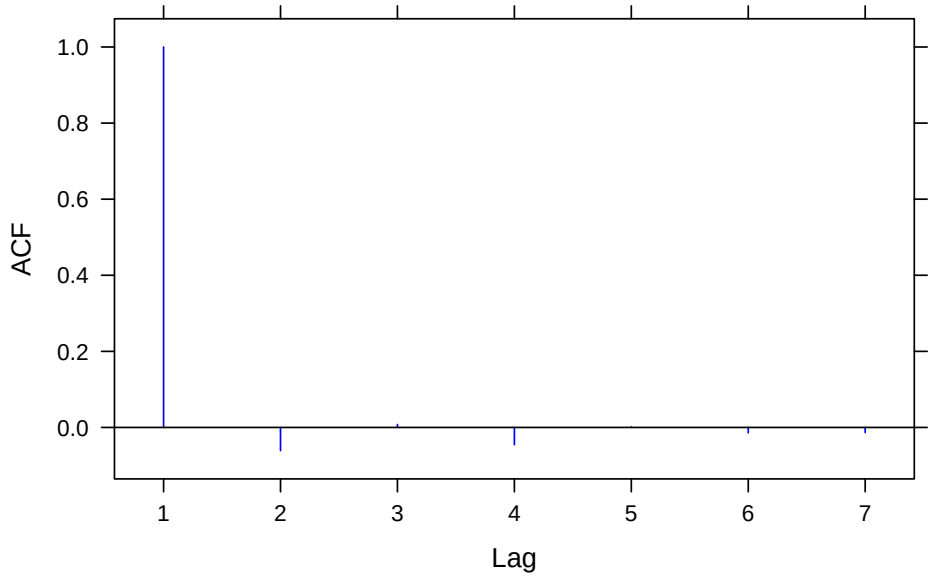


Figure 4.8: Average ACF for goal differences through seasons 1999/2000 till 2014/2015

4.7 Mutual matches

This section is devoted to an analysis that aims to answer a question if some teams play significantly better against others. It is also inspired by newspaper headlines, which often argue that certain team has beaten another one in last ten or even more matches.

To test this hypothesis, we need to make a model with interactions for every pair of teams that have played against one another. For this we use dataset MP and expand model (4.15) of interactions between factors team and opponent, but we need to set a constraint that interaction between two teams is represented only by one interaction term, which is in the model with a plus sign if a team i plays against a team j and with a negative sign otherwise ($i < j$). The model has the following form

$$GD_{ijt} = \beta_0 + \beta_1 S_{ijt} + \beta_2 R_{ijt} + \sum_{r=0}^{15} (\alpha_{ir} - \alpha_{jr}) \mathbb{1}_{[S_{ijt}=r]} + \gamma_{ij} + \varepsilon_{ijt}, \qquad (4.18)$$
$$\gamma_{ji} = -\gamma_{ij},\ t = 1, \dots m_{ij}, 1 \leq i < j \leq I.$$

The factor terms γ_{ij}, $j > i$, are mutually insignificant using F-test (see Table 4.6). However, we have to add that the number of matches between two teams could

Effect	Sum of Squares	Df	F-value	p-value
Team:Opponent	1262	226	1.03	0.37

Table 4.6: Analysis of variance comparing models (4.18) and (4.15).

Effect	Estimate	Std. error	T-value	p-value
Mean goal difference of previous mutual matches	-0.0631	0.03791	-1.67	0.096

Table 4.7: OLS estimate of variable of interest in model (4.15) completed with the variable

be as low as four (in case that they met only in one season). Hence, the efficiency of testing is rather small.

For that reason, we present another type of testing whether the history of matches between two teams plays significant role for the outcome of their next match. We will consider one parameter that contains the information of history between two teams – mean goal differences in previous matches between two teams that play the current match. Again we expand the model (4.15) by adding the variable mean goal differences of previous mutual matches in the whole history of our dataset. The estimated parameter is insignificant at a confidence level 0.01 (see Table 4.7).

Surprisingly, the estimated coefficient is negative. This is not intuitive we would expect positive effect of better performance in the history in mutual matches. The reason is that this variable carries similar information as the difference of estimated forms. If we omitted categorical variables representing forms (α_{ir} from model (4.15)) then the effect of mean goal differences would be strongly positive and significant as it carries the information of different forms. This principle can be graphically assessed by using partial residuals (see Figure 4.9). For definition and properties of partial residuals see Zvára (2008, pages 111-112).

To conclude result of this section, there is not strong evidence that the history of previous mutual matches of two teams would have an influence on the current outcome. Therefore, it will not be included in construction of forecasting model.

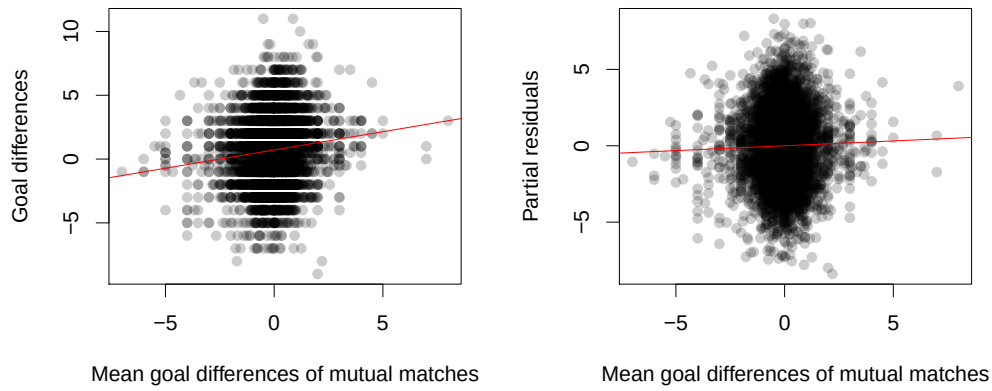


Figure 4.9: Dependence of goal differences on mean goal differences of previous mutual matches (left), partial residuals of the same dependence after adjustment on home advantage and forms of teams (right).

Days from the last match	Mean goal difference	Home advantage	Number of occurences
0	0.163	0.515	295
1	0.005	0.499	5984
2	0.036	0.502	2444
3	-0.008	0.498	428
4	-0.154	0.494	1323
5	0.244	0.504	123
> 5	0.014	0.505	1051

Table 4.8: Mean goal differences for different number of free days between matches.

4.8 Tiredness from the previous match

In Extraliga every team plays 52 matches within a season, which lasted in last 15 seasons from 156 to 181 days. Hence, the schedule is quite busy and players might not have enough time to recover. Teams had one free day between matches in 52 % cases and 2 free days in 21 %. Until season 2006/2007, there were 33 matches in average where teams had two matches in consecutive days. From that season it happens only exceptionally (in season 2014/2015 it occurred only twice, for more detail see Table A1).

To get an intuition how strong the effect of tiredness could be, we calculate mean goal difference conditional on days since the last match. Results are unexpected as the biggest value is for no free days between matches and the relationship between free days and goal differences is not linear (see Table 4.8). There is also percentage of cases that were played on the home field in the table. We see that positive values correspond with more than half of the matches on the home field (with one exception of one free day). Therefore, we should filter the effect of home advantage by using linear regression as in previous sections.

However, at first we need to define the variable of tiredness, which would contain the information of difference of tiredness between both teams. One option could be to use differences of free days but then it would assume the same difference in tiredness for free days 1 and 2 as for 12 and 13. Taking in account that tiredness vanishes after few days we define the following variable

$$TD_{ij} = \begin{cases} 1 & \text{if team } i \text{ had at least 2 free days and team } j \text{ at most 1} \\ -1 & \text{if team } j \text{ had at least 2 free days and team } i \text{ at most 1} \\ 0 & \text{otherwise.} \end{cases}$$

This variable is non-zero only in 10 % of matches. We examine the effect of tiredness through variable TD_{ij} by including it into the model (4.15). We use dataset MP and the model of interest is

$$GD_{ijt} = \beta_0 + \beta_1 S_{ijt} + \beta_2 R_{ijt} + \sum_{r=0}^{15} (\alpha_{ir} - \alpha_{jr}) \mathbb{1}_{[S_{ijt}=r]} + \beta_4 TD_{ijt} + \varepsilon_{ijt}, \quad t = 1, \dots m_{ij}, 1 \leq i < j \leq I. \tag{4.19}$$

Effect	Estimate	Std. error	T-value	p-value
Tiredness	0.179	0.095	1.89	0.059

Table 4.9: OLS estimates from model (4.19) of variables of interest.

Tiredness is insignificant for estimating goal differences at confidence level 0.01 but its coefficient has expected sign – positive in case of more time to recover (see Table 4.9).

4.9 Measures of teams forms

Forms of both teams have strong effect for the outcome (see the results of statistical significance in Table 4.2). Until now we have estimated those effects as categorical variables for every team and every season. However, this approach can be used only for evaluating forms in a certain season when all matches are over. If we wanted to predict the outcomes of next season, there would be no direct approach. We might use the estimated forms from the preceding season, but those estimates might be biased as teams might buy or sell some players. Moreover, we would lack estimates of two teams that might join Extraliga as winners of premier league. Complication of estimation forms is that we never see a result of form of one team but always only result of difference of forms for opposing teams.

In this section we consider several factors that might be used as measures of performance without using more complicated approach (Kalman filter in the following chapter). At first we realize how we computed estimates of forms in the model (4.15). We have used OLS estimates, which means that the estimated forms are linear combinations of goal differences in specific season. Therefore, one estimate might be mean goal differences in previous matches. The question could be how long history should be chosen, the forms may change and results from past might be misleading from some point in time.

In the following we consider measures of performance (denoted as α_i) based on history of goal differences and gained points. Estimate of current form $\widehat{\alpha}_{it}$ is iteratively computed for every team and difference of teams forms comes as explanatory variable to the model. To let the performance measures adjust, we use observations since the 10th match in the season for each team throughout this section. To compare models we use goodness of fit criteria – coefficient of determination and Gini coefficients for dataset TP. All models are fitted using linear and ordinal regression in form:

$$\begin{aligned} GD_{ijt} &= \beta_0 HA_{ijt} + \beta_1 HA_{ijt}S_{ijt} + \beta_2 HA_{ijt}R_{ijt} + \beta_5(\widehat{\alpha}_{it} - \widehat{\alpha}_{jt}) + \varepsilon_{ijt}, \\ \text{logit}(P(Y_{ijt} \leq r) &= \gamma_r + \beta_0 HA_{ijt} + \beta_1 HA_{ijt}S_{ijt} + \beta_2 HA_{ijt}R_{ijt} + \beta_5(\widehat{\alpha}_{it} - \widehat{\alpha}_{jt}), \\ r \in \{0,1\}, &\quad R_{ijt} > 10,\, t = 1,\ldots,n_{ij},\, i,j = 1,\ldots,I. \end{aligned} \tag{4.20}$$

More precisely the dataset contains all matches for team i after its 10th match.

At first we investigate how long history is better to use. As measures of performances we use mean goal differences/points in last x matches with $x = 1,\ldots,10$ and the whole history at given time. To see the improvement of models including

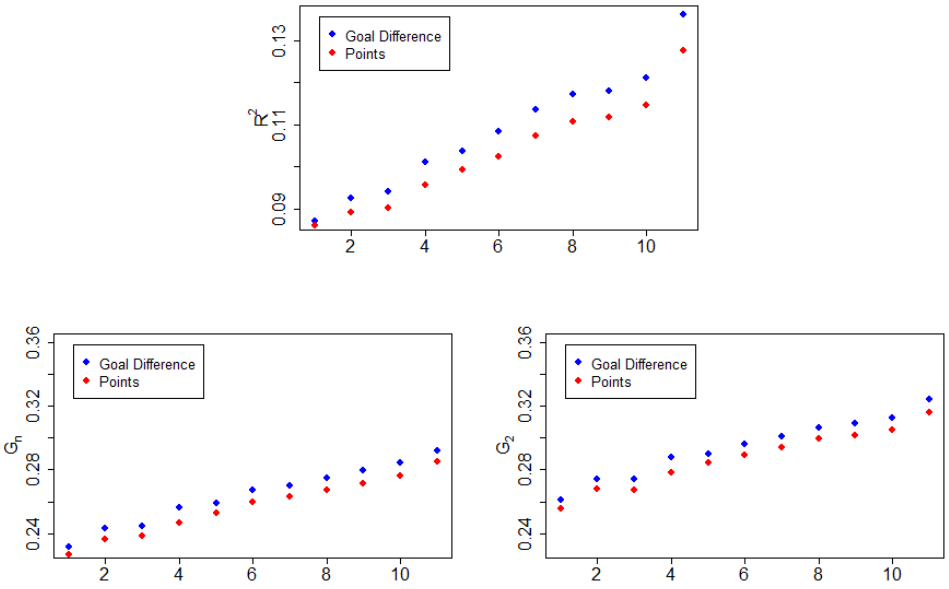


Figure 4.10: Goodness of fit criteria for measures of performance – mean goal differences and mean points in last 1 to 10 matches and for the whole history in season (last points). In the upper picture there is R^2 and in lower part Gini coefficient nominal (left) and based on ranks (right).

forms, we calculate goodness of fit for models (4.20), which are only adjusted for home advantage: $R^2 = 0.083, G_n = 0.215, G_C = 0.241, G_1 = 0.268$ and $G_2 = 0.268$.

This benchmark was exceeded in every measure by each model. Results for different measures of performance are summarized in Table A3. We have also added another measure – position in table, which uses the difference of current rank in the competition. This information might be typically used by bookmakers as it is easily accessible. Surprisingly, it belongs among better ones.

For better clarity, three criteria were depicted for different number of past values used (see Figure 4.10). All criteria suggests that the predictability power increases as we take longer history. The best option is to take the whole current history for all criteria and both variants. Note also that goal differences yield better results than average points.

It seems that long history is relevant for holding information about forms of teams. However, we would expect the new information to be more relevant than older. Therefore, it could be logical to use exponential weighting for evaluating forms. Exponentially smoothed estimates are in the following form:

$$\widehat{\alpha}_t = a\ x_t + (1-a)\widehat{\alpha}_{t-1}, \quad a\ \in (0,1) \tag{4.21}$$

where x_t denotes new observation and we initialize the formula by setting $\widehat{\alpha}_0 = 0$. In our situation it could be either goal difference or number of points. We use this method to assess forms with different values of smoothing parameter a for goal differences and for points.

To get an intuition on the dependence on new observation, we use grid

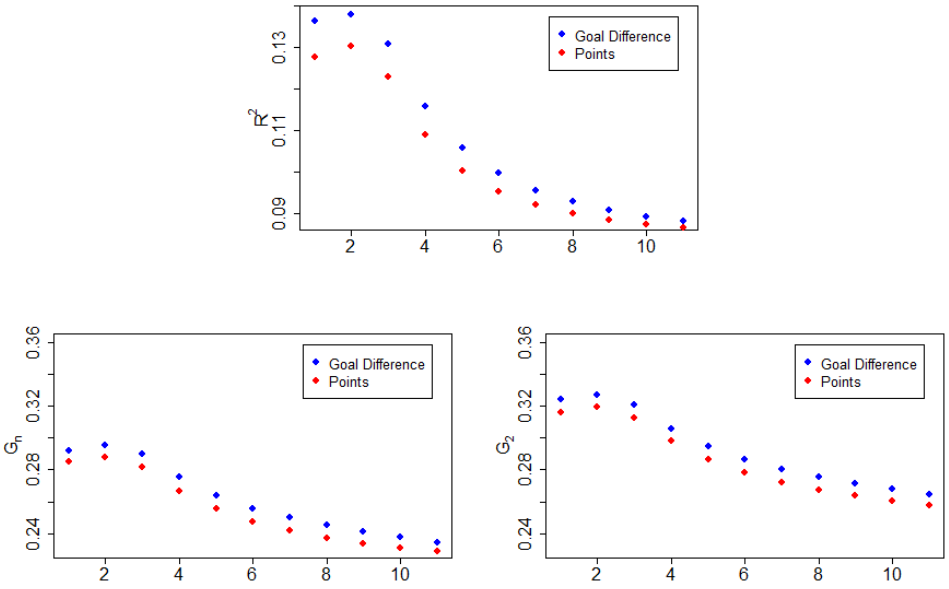


Figure 4.11: Goodness of fit criteria for measures of performance – the first is mean goal differences and points followed by exponentially smoothed forms with smoothing parameter a with values $0.05, 0.1, 0.2, \ldots, 0.9$. In the upper picture there is R^2 and in lower part Gini coefficient nominal (left) and based on ranks (right).

for smoothing parameter a with values $0.5, 0.1, 0.2, \ldots, 0.9$ and calculate R^2 and Gini coefficients as before. For comparison as first measure we include the best option of previous result – mean goal differences/points from the entire history. Results are summarized in Table A4 and some of them depicted in Figure 4.11. All measures give the same result that the best option is the lowest value of smoothing parameter $a = 0.05$.

However, we may want to maximize our criteria to find the best value of smoothing parameter. They all lead to optimal value between 0.028 and 0.036 (see Table 4.10). This result coincides with the previous result that long history remains relevant. In optimal cases it outperforms mean goal differences up to given time.

Disadvantage of using plain goal differences is that it does not account for HA and form of the opponent. We might assume that loss against strong opponent does not indicate such a poor performance as loss against weak opponent. Therefore, we might want to take into account form of the other team and adjust for home advantage. We may use form (4.21) and as observation we take

	R^2	G_n	G_C	G_1	G_2
Optimal smoother	0.028	0.043	0.037	0.036	0.036
Value	0.139	0.295	0.327	0.363	0.363

Table 4.10: Optimal values of smoother a in model (4.21) according to different goodness of fit criteria.

	R^2	G_n	G_C	G_1	G_2
Optimal smoother	0.010	0.017	0.013	0.012	0.012
Value	0.139	0.294	0.326	0.361	0.361

Table 4.11: Optimal values of smoother a in model (4.22) according to different goodness of fit criteria.

residual of regression model (4.20) with estimates of forms at a previous step. The estimation of form for team i at time t for a given year is

$$\begin{aligned} r_{ijt} &= GD_{ijt} - (\widehat{\beta}_t HA_{ijt} + \widehat{\alpha}_{it-1} - \widehat{\alpha}_{jt-1}), \\ \widehat{\alpha}_{it} &= a\ r_{ijt} + (1-a)\widehat{\alpha}_{it-1}. \end{aligned} \tag{4.22}$$

For the initialization we use $\widehat{\alpha}_{i1} = 0$ as before and coefficient $\widehat{\beta}_t$ is calculated from estimates 4.3, so we adjust home advantage for season and round.

For evaluating the model with forms estimated by this procedure, we estimate all forms $\widehat{\alpha}_{it}$, $i = 1, \ldots, I\,, t = 1, \ldots, T$ for every season and then fit linear model (4.20). This method resembles procedure of Kalman filter with $F_t = I$ and $K_t = a\ I$ and suitable design matrix Z_t. The difference is that we diminish the role of previous estimate by multiplying it by $(1-a)$. Another useful property is that if one form is increased by x the opponent's is decreased by x and at the beginning $\widehat{\alpha}_{i1} = 0$. So it holds that at any time forms sum up to zero.

The optimal values of a are even lower than in case of exponential smoothing (see Table 4.11) with comparable fit as for exponential smoothing using goal differences.

Last option that we examine is linear regression with observations up to current time s and forms are estimated as categorical variables. We fix the estimates of home advantage as in previous case, so the dependent variable for a given year r is $Y^r_{ijt} = GD^r_{ijt} - \widehat{\beta}_t HA^r_{ijt}$. To determine which teams played at particular time t we define set of pairs

$$A_t = \{(i,j) : \text{in time } t \text{ there is a match between home team } i \text{ and guest team } j\}.$$

We fit linear regression with all measurements that are available at current time s since tenth match for a given team as follows

$$Y^r_{ijt} = \alpha^r_{it} - \alpha^r_{jt} + \varepsilon^r_{ijt}, \quad (i,j) \in A_t, \quad t = 1, \ldots, s, \quad r = 0, \ldots 15 \tag{4.23}$$

with constraints for forms (4.14). The estimates of α^r_{it} are estimates of forms. For estimation we use weighted least squares with higher weights for more current observations in exponential form. Weighted residual sum of squares is in form:

$$RSS(w) = \sum_{r=0}^{15} \sum_{s=1}^{t} w^s \sum_{(i,j)\in A_t} (Y^r_{ijs} - \alpha^r_{is} + \alpha^r_{js})^2.$$

For a given weight w we estimate forms and plug them into model 4.20 and calculate its coefficient of determination R^2. To find the most suitable w we maximize

	R^2	G_n	G_C	G_1	G_2
Value	0.137	0.290	0.323	0.358	0.358

Table 4.12: Goodness of fit criteria for weighted least squares with optimal exponential weights.

R^2 on interval $[1, 2]$. The optimal weight is $w = 1.011$ and goodness of fit criteria are slightly smaller than for optimal cases of preceding measures (see Table 4.12).

We can conclude that goal differences are more appropriate for assessing forms of teams than points and every observation remains relevant for long period. Suggested measures were based on exponential smoothing of goal differences, goal differences adjusted for the effect of HA and rolling regression with increasing number of observations. All measures yield similar quality of fit. The first two measures are dependent on initial estimate for long time, which diminishes their quality as it is set to zero. From that reason rolling regression would be better to use. However, in the next section we implement Kalman filter and compare given results.

4.10 Kalman filter for estimating forms

In this section we consider Extraliga matches in particular season as dynamic linear model. *Measurements* are goal differences as it was shown that it is more suitable for estimating outcome probabilities than plain result (see section 4.4). Unobserved *state vector* are forms of teams and possibly coefficients of other predictors.

Since we assume particular season, number of teams equals $I = 14$ and time points t correspond to different days when at least one match was played (total is denoted by T). Minimum value is $T = 52$ if no match is postponed and every round is within one day. Number of matches within time point t (denoted as N_t) is between one and seven.

In preceding sections we have identified predictors that influence goal differences and the conclusion was that the only significant effect has home advantage, which increases during a season and is not significantly team specific. This information is essential to build suitable state space model. Home advantage might be taken as a fixed parameter or as a random process. We choose the option as fixed because time dependence is given rather by seasonal effect or increasing importance of matches than as a consequence of previous realizations. To include the effect of home advantage in the model we use the formulation (2.15) which enables to separate random and fixed effects.

For this purpose we use two parameters. One for home advantage at the beginning of the season and the other for its increasing trend. The state vector and fixed parameters are denoted

$$\boldsymbol{\alpha}_t = (\alpha_{1t}, \ldots, \alpha_{It})^\top, \quad \boldsymbol{\beta} = (\beta_0, \beta_1)^\top.$$

The measurement equation (2.1) for Extraliga matches is given by following

assumptions

$$GD_{ijt} = \boldsymbol{x}_{ijt}\boldsymbol{\beta} + \boldsymbol{z}_{ijt}\boldsymbol{\alpha}_t + v_{ijt} = \beta_0 + \beta_1 R_{ijt} + \alpha_{it} - \alpha_{jt} + v_{ijt},$$

$$\boldsymbol{x}_{ijt} = (1, R_{ijt}), \quad \boldsymbol{z}_{ijt} = \boldsymbol{c}_{ij}^\top, \quad \boldsymbol{c}_{ij} = (0, \ldots, 1, \ldots, -1, \ldots, 0)^\top,$$

$$v_{ijt} \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma_g^2), \quad (i,j) \in A_t,\, t = 1, \ldots, T.$$

The vector $\boldsymbol{c}_{ij}$ is an indicator of home and guest team, component i is 1, component j is -1 and all others are 0. Variable R_{ijt} expresses time in the season, it might be either round or number of matches played before by home team. We assume that matches that are played at the same time t are independent. However, we have to specify the whole vector of measurements such that the dimension is not changed for different times. For this reason we stack all possible goal differences into a vector of dimension $I^2 \times 1$ as follows

$$\boldsymbol{GD}_t = (GD_{11t}, \ldots, GD_{1It}, GD_{2It}, \ldots, GD_{IIt}). \tag{4.24}$$

and similarly for design matrices

$$Z_t = \begin{pmatrix} \boldsymbol{d}_{11} \\ \vdots \\ \boldsymbol{d}_{1I} \\ \boldsymbol{d}_{21} \\ \vdots \\ \boldsymbol{d}_{II} \end{pmatrix}, \quad \boldsymbol{d}_{ij} = \begin{cases} \boldsymbol{c}_{ij}, & \text{if } (i,j) \in A_t, \\ \boldsymbol{0}, & \text{otherwise.} \end{cases} \quad \mathbb{X}_t = \begin{pmatrix} 1 & R_{11t} \\ \vdots & \vdots \\ 1 & R_{1It} \\ 1 & R_{21t} \\ \vdots & \vdots \\ 1 & R_{IIt} \end{pmatrix}. \tag{4.25}$$

It is evident that vector $\boldsymbol{GD}_t$ contains large number of systematically missing values but it does not matter as we can easily define matrix W_t that determines which measurements of vector $\boldsymbol{GD}_t$ are missing as in section 2.3. Its dimension is $N_t \times I^2$ and it equals identity matrix with omitted rows where components of $\boldsymbol{GD}_t$ are missing. Now the measurement equation can be written in matrix form:

$$W_t\boldsymbol{GD}_t = W_t Z_t \boldsymbol{\alpha}_t + W_t \mathbb{X}_t \boldsymbol{\beta} + W_t \boldsymbol{v}_t,$$

$$\boldsymbol{v}_t \overset{\text{iid}}{\sim} \mathsf{N}(\boldsymbol{0}, \sigma_g^2 \mathbb{I}), \quad t = 1, \ldots, T.$$

We have determined the form of measurement equation. In the following, we specify equation for state vector $\boldsymbol{\alpha}_t$. We assume that team forms are mutually independent. For each form we assume that they are Gaussian random walks, the same assumption was taken in Farhmeir and Tutz (1994) and Knorr-Held (2000). For form of one team we assume:

$$\alpha_{it} | \alpha_{it-1} \sim \mathsf{N}(\alpha_{it-1}, \sigma_a^2), \quad \alpha_{i0} \sim \mathsf{N}(0, \sigma_0^2).$$

However, to assure identifiability we must impose some constraint on forms $\boldsymbol{\alpha}_t$, $t = 1, \ldots, T$. The most natural is to set their sum to zero as in (4.14) but for every time $t = 1, \ldots, T$. This means that vector $\boldsymbol{\alpha}_t$ follows multivariate Gaussian random walk with variance matrix Q that ensures $\boldsymbol{1}^\top \boldsymbol{\alpha}_t = 0$ and the same covariances between all forms.

Such matrix is given by

$$Q = \frac{I}{I-1}\left(\mathbb{I}_I - \frac{1}{I}\mathbf{1}\mathbf{1}^\top\right).$$

In the following we investigate distribution of forms according to definition

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + \boldsymbol{w}_t, \quad \boldsymbol{w}_t \sim \mathsf{N}(\mathbf{0}, \sigma_a^2 Q), \quad \boldsymbol{\alpha}_0 \sim \mathsf{N}(\mathbf{0}, \sigma_0^2 Q). \tag{4.26}$$

For variances of components of disturbances it holds that $\mathrm{var}(w_{it}) = \sigma_a^2$ and for covariances $\mathrm{cov}(w_{it}, w_{jt}) = -\frac{1}{I-1}\sigma_a^2$, $i, j = 1, \ldots, I, i \neq j$, $t = 1, \ldots, T$. It means that forms are negatively correlated. Moreover, it implies that $\mathsf{E}\ \mathbf{1}^\top \boldsymbol{\alpha}_t = \mathbf{0}$ and for variance

$$\mathrm{var}(\mathbf{1}^\top \boldsymbol{\alpha}_t) = \sigma_a^2 \mathbf{1}^\top Q \mathbf{1} = 0.$$

Hence, the choice of matrix Q ensures that condition (4.14) holds almost surely.

As discussed in Knorr-Held (2000) it is computationally convenient to transform $\boldsymbol{\alpha}_t$ such that its first $I-1$ components follow regular random walk and the last is equal to zero almost surely. Such transformation might be linear with matrix L defined as

$$L = \begin{pmatrix} \mathbb{I}_{I-1} & -\mathbf{1} \\ \mathbf{1}^\top & 1 \end{pmatrix}.$$

It holds that $L\boldsymbol{\alpha}_t = (\alpha_1 - \alpha_I, \alpha_2 - \alpha_I, \ldots, \alpha_{I-1} - \alpha_I, 0)^\top = (\widetilde{\boldsymbol{\alpha}}_t^\top, 0)^\top$. The innovation of random walk for first $I-1$ components has the following variance matrix

$$\widetilde{Q} = \mathrm{var}(\widetilde{\boldsymbol{w}_t}) = (\mathbb{I}_{I-1}, -\mathbf{1})\, \sigma_a^2 Q\, (\mathbb{I}_{I-1}, -\mathbf{1})^\top = \frac{I}{I-1}\sigma_a^2(\mathbb{I}_{I-1} + \mathbf{1}\mathbf{1}^\top).$$

We may now perform the analysis using random walk of $\widetilde{\boldsymbol{\alpha}}_t$ with slightly changed design matrix Z_t. It holds that $\alpha_{it} - \alpha_{jt} = \alpha_{it} - \alpha_{It} - \alpha_{jt} + \alpha_{It} = \widetilde{\alpha}_{it} - \widetilde{\alpha}_{jt}$, so in matrix Z_t we only omit the last column. After we get the estimated forms of $\widetilde{\boldsymbol{\alpha}}_t$, we transform them back using linear transformation

$$\alpha_{jt} = \widetilde{\alpha}_{jt} - \frac{1}{I-1}\sum_{i=1}^{I-1}\widetilde{\alpha}_{it}, \quad \alpha_{It} = -\frac{1}{I-1}\sum_{i=1}^{I-1}\widetilde{\alpha}_{it}, \quad j = 1, \ldots, I-1, t = 1, \ldots, T.$$

Unknown hyperparameters are $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma_g^2, \sigma_a^2)^\top$ and σ_0^2.

The dynamic linear model for one season in ice hockey is defined by following equations and distributional conditions

$$\boldsymbol{G}\boldsymbol{D}_t = Z_t\widetilde{\boldsymbol{\alpha}}_t + \mathbb{X}_t\boldsymbol{\beta} + \boldsymbol{v}_t, \tag{4.27}$$

$$\widetilde{\boldsymbol{\alpha}}_t = \widetilde{\boldsymbol{\alpha}}_{t-1} + \boldsymbol{w}_t, \tag{4.28}$$

$$\widetilde{\boldsymbol{\alpha}}_0 \sim \mathsf{N}(\mathbf{0}, \sigma_0^2\widetilde{Q}), \quad \boldsymbol{v}_t \overset{\text{iid}}{\sim} \mathsf{N}(\mathbf{0}, \sigma_g^2\mathbb{I}_I), \quad \boldsymbol{w}_t \overset{\text{iid}}{\sim} \mathsf{N}(\mathbf{0}, \sigma_a^2\widetilde{Q}), \quad t = 1, \ldots, T.$$

Using this system we may filter forms $\widetilde{\boldsymbol{\alpha}}_t$ and then transform them into $\boldsymbol{\alpha}_t$ and also predict goal differences in the next match. To do this we need to estimate hyperparameters $\boldsymbol{\theta}$. This could be done separately for every season using maximum likelihood estimation or EM algorithm presented in section 2.2. However, we want to use the whole data at once for estimating hyperparameters $\boldsymbol{\theta}$.

σ_0^2	R^2	G_n	G_C	G_1	G_2
100	0.137	0.290	0.293	0.324	0.324
10	0.136	0.289	0.294	0.326	0.326
1	0.138	0.291	0.317	0.351	0.351
0.1	0.138	0.293	0.321	0.356	0.356
$8.207 \cdot 10^{-3}$	0.134	0.288	0.319	0.353	0.353

Table 4.13: Goodness of fit criteria for different values of initial state variance σ_0^2. The last option is fitted with the constraint $\sigma_0^2 = \sigma_a^2$ in model (4.20) for estimated forms $\boldsymbol{\alpha}_t$ using Kalman filter.

This would assume that all parameters are the same for every season. In section 4.5, we have discovered that home advantage is decreasing in time. For that reason we add one additional fixed effect S_{ijt} that is identifier of the season with parameter β_3. This choice guarantees the same conditions as for measures of performance in previous section and enables their comparison.

For estimation we use direct approach based on integrated log-likelihood. For a given season s the log-likelihood is $\ell_s(\boldsymbol{\theta})$ as in section 2.2. We assume that results of matches over seasons are independent, hence, the log-likelihood of the whole data is

$$\ell(\boldsymbol{\theta}) = \sum_{s=1}^{15} \ell_s(\boldsymbol{\theta}).$$

We maximize $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. However, we have to choose initial value of σ_0^2. In Knorr-Held (2000) they set $\sigma_0^2 = 1$ claiming that it is weakly informative but avoids numerical problems with more diffuse priors. We investigate the results using different values of σ_0^2.

Kalman filter might be used to estimate goal differences in the following match. For the moment, we want to compare estimated forms by Kalman filter and methods developed in the previous section. For that reason, we use Kalman filter only to get estimates $\boldsymbol{\alpha}_{t|t-1}$ in every season and then use models (4.20) with observations since 10th match in every season to calculate R^2 and Gini coefficients. For values $\sigma_0^2 = 0.1$ and then $\sigma_0^2 = 1$ is the fit better than for others (see Table 4.13). In comparison with measures of performance in previous section the results are similar (see Tables 4.10 and 4.11).

By investigating the fitted forms for different values of initial state σ_0^2 it holds that the higher the value the more volatile fitted forms (see Figure 4.12). This corresponds with the intuition since bigger initial variance means lower validity of the initial value and depends more on first observations. There is no significant difference for values 100 and 10 and on the other hand for 0.1 and 0.008. We also focus on residuals of models given by $v_{ijt} = GD_{ijt} - \widehat{\alpha}_{it} + \widehat{\alpha}_{jt} - \boldsymbol{x}_t\widehat{\boldsymbol{\beta}}$ with $\sigma_0^2 = 1$ and $\sigma_0^2 = 0.1$. Residuals and fitted values should be independent, so the regression line should have no trend and go through zero. In both cases residuals have mean close to zero but for $\sigma_0^2 = 1$ there is significant decreasing trend resulting in overestimation for high values of fitted values (or underestimation otherwise, see Figure A2). For $\sigma_0^2 = 0.1$ we observe reverse situation but in this case the slope is not significant. For that reason and having higher values of goodness of fit cri-

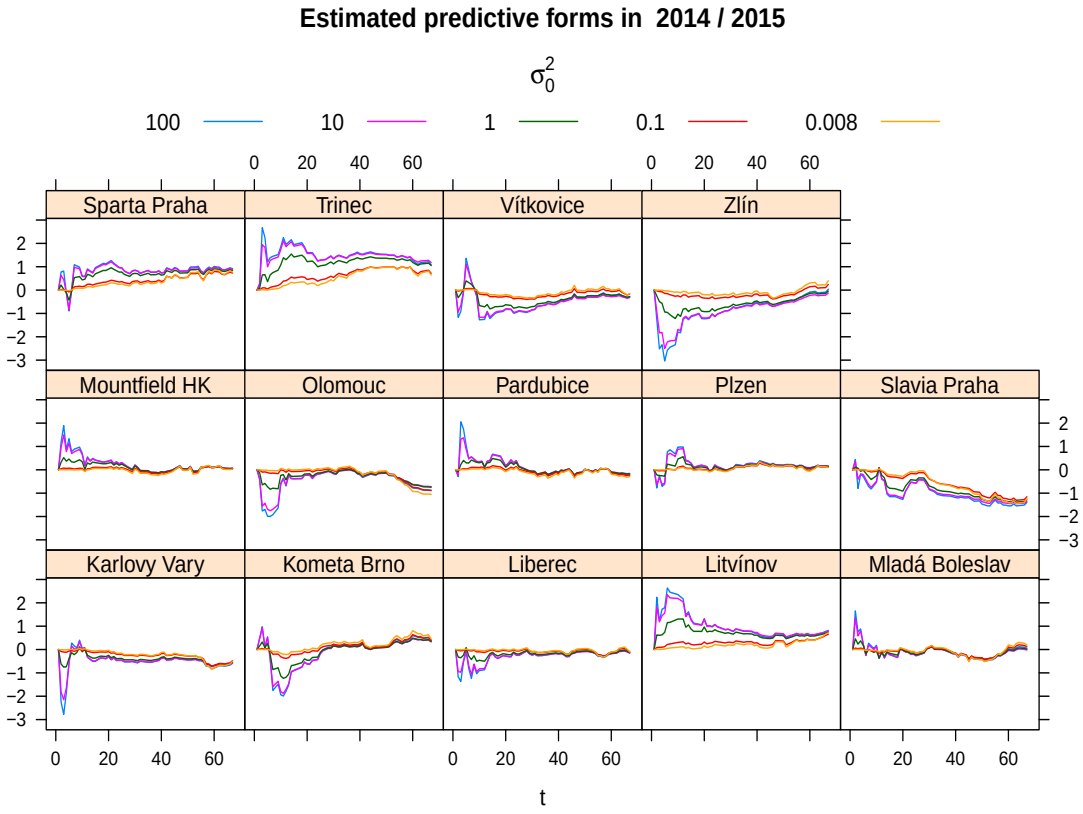


Figure 4.12: Estimated predictive forms using Kalman filter in season 2014/2015 for different values of variance of initial state σ_0^2 (the last option is fitted with $\sigma_0^2 = \sigma_b^2$).

teria following statistics refer to this option.

The estimated effects using MLE give similar estimates as OLS estimates in model with forms as dummy variables (compare Tables 4.3 and 4.14). The estimated variance of forms σ_a^2 is rather low and corresponds with earlier results suggesting that forms of teams do not vary dramatically over the season. This result assures that the assumption of constant forms over the season is not misplaced, which is an assumption of linear regression that we used for identifying significant predictors.

Effect	**Estimate**
HA	0.729
Season	−0.022
Round	0.005
σ_g^2	5.354
σ_a^2	0.005

Table 4.14: MLE estimates from Kalman filter with $\sigma_0^2 = 0.1$.

Chapter 5

Search for profitable strategy

In this chapter we firstly focus on average odds offered by betting companies and analyse, which predictors bookmakers use to set these odds and if it corresponds with the effect they have on goal differences. As in the previous chapter we investigate the influence of home advantage, effect of goal difference from the previous match, results of previous mutual matches, tiredness and team forms. Secondly, we compare goodness of fit criteria for odds and models from the previous chapter. At last we consider the possibility of using odds for a prediction of a match outcome. To use transformation of odds as explanatory variable is possible because betting companies provide them before the match.

5.1 Analysis of betting odds

To transform odds into a variable with real values we turn them into probabilities and then use logit function (as in equations (4.9)). Useful property of this transformation is symmetry because

$$\operatorname{logit}\left(\frac{p^w}{p^w+p^l}\right) = -\operatorname{logit}\left(\frac{p^l}{p^w+p^l}\right).$$

This assures the symmetry of dataset TP where instead of goal differences we use transformed odds. In the following, we call them logits and for ith match are $L_i = \operatorname{logit}\left(\frac{p_i^w}{p_i^w+p_i^l}\right)$. This continuous variable is defined such that zero is reached in case that probabilities to win for both teams equal and expected goal difference is zero and is positive in case that expected goal difference is positive.

We look closely on the effect of home advantage (HA) because it is the only predictor that was significant for goal differences except team forms. The average odds for the home win is 2.02 and for the away win is 3.19. Therefore, it is obvious that betting companies take HA into account. In the previous chapter we have seen that HA is decreasing in years and increasing in rounds. In that way logits follow the same pattern (see Figures 5.1). Difference of logits at home and out for every team and every year is also depicted in the same figure.

To determine whether those trends are significant and if difference of logits is independent on teams we use the model (4.13) with response logits instead of goal differences.

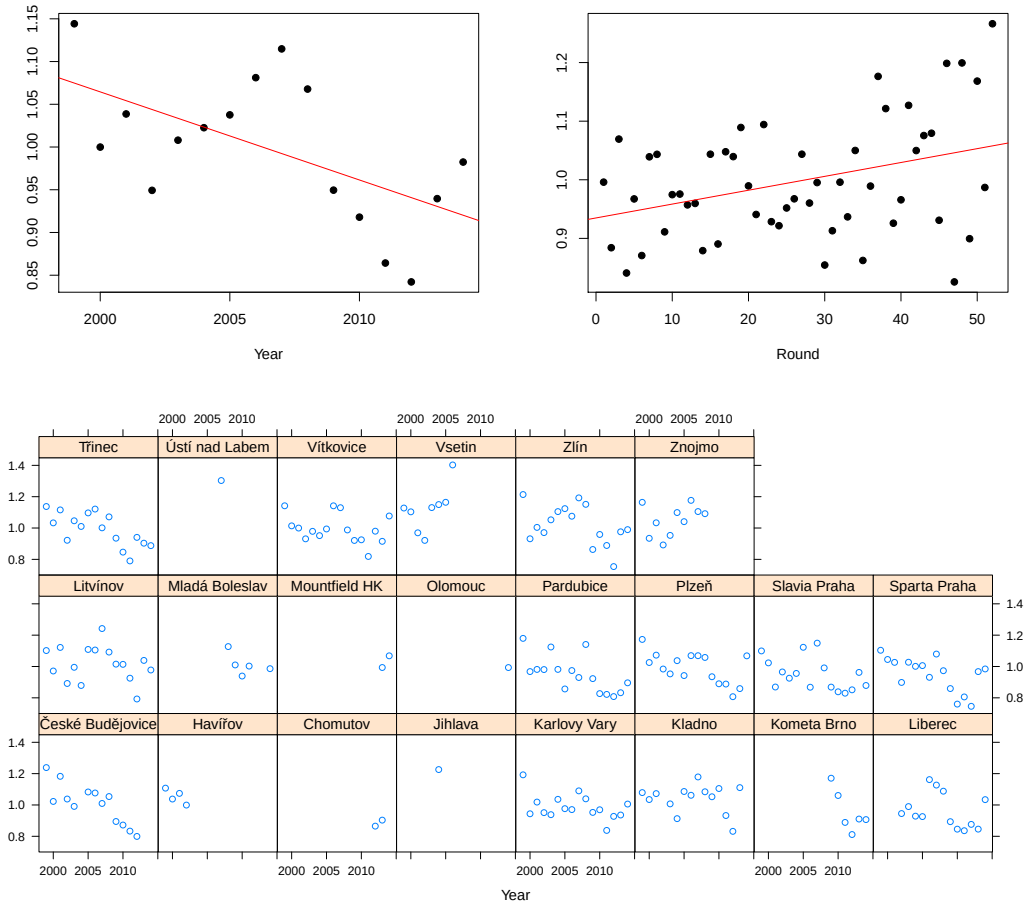


Figure 5.1: Mean logit differences for matches played at home and out. They are depicted for every season and every team in the upper picture and by year and by round in the lower figures with a regression line.

The model is

$$L_{ijt} = \beta_0 + \beta_1 S_{ijt} + \beta_2 R_{ijt} + \delta_i + \sum_{r=0}^{15} (\alpha_{ir} - \alpha_{jr}) \mathbb{1}_{[S_{ijt}=r]} + \varepsilon_{ijt}, \qquad (5.1)$$
$$i, j = 1, \ldots, I,\ t = 1, \ldots, m_{ij}$$

with the same constraints on δ_i and α_{ir} as for model (4.13).

We test significance of all effects using anova F-test. All p-values are smaller than 0.01, so there is evidence that logits are decreasing in season, increasing in rounds (see Table 5.1). The effect of team is also significant; for certain teams the difference between logits home and out is different. The estimated coefficients $\widehat{\delta}_i$ are significantly higher for Ústí nad Labem and Jihlava. Both of those teams played only in one season and performed poorly. If we also look that the highest difference (see Figure 5.1) it was observed for Vsetín in its last season in Extraliga, we might assume that greater differences in logits at home and out might be for teams with bad forms.

For that reason, we have calculated average logits for every team and every season and difference between average logits at home and out. There is

Effect	Sum of Squares	Df	F-value	p-value
Season	2.15	1	31.201	< 0.001
Round	1.50	1	21.883	< 0.001
Team	5.37	21	3.716	< 0.001
Forms	1398.60	208	97.784	< 0.001

Table 5.1: Analysis of variance (Type II tests)

a significant quadratic relationship between those variables (for illustration see Figure A3). It means that the estimated odds to win home are higher for the non-average teams.

To quantify the effect of other variables we would use the analogous steps as in previous chapter with response logits instead of goal differences. Not to repeat the same steps we restrict the commentary to stating results. Firstly, we discuss the effect of result from the previous match. We fit analogous model as (4.15). The estimated coefficient $\widehat{\beta}_3 = 0.013$ and is highly significant. This means that goal difference from the last match has a positive effect on the following logits. For goal differences the effect was insignificant after filtering the average effect of forms.

Secondly, mutual matches are significant in both terms as a categorical variable and also as mean goal difference of previous mutual matches. The estimated coefficient for the second mentioned is positive, which is the opposite effect than for goal differences.

At last, the covariate tiredness is insignificant with p-value = 0.429. The same result was observed for goal differences. These results are useful for identifying covariates that are used to set odds. However, the comparison with goal differences is questionable as it contains categorical variables of team forms. It means that team forms are either computed based on goal differences or logits depending on the response that we use. For that reason, we construct two models with the same measure of performance. We use estimated coefficients by rolling regression fitted in model (4.23) with $w = 1.011$. We use data since tenth match for each team and every year to have more reliant estimates of forms. Models for particular year are:

$$\begin{aligned} Y_{ijt} = \beta_0 &+ \beta_1 S_{ijt} + \beta_2 R_{ijt} + \beta_3 (GD_{i.t-1} - GD_{j.t-1}) \\ &+ \beta_4 GDMM_{ijt} + \beta_5 TD_{ijt} + \beta_6 (\widehat{\alpha}_{it} - \widehat{\alpha}_{jt}) + \varepsilon_{ijt}, \\ &R_{ijt} > 10,\ t = 1, \ldots, n_{ij},\ i, j = 1, \ldots, I \end{aligned}$$

where Y_{ijt} is either GD_{ijt} or L_{ijt} and variable $GDMM_{ijt}$ is mean goal difference in previous matches between teams i and j.

Estimated effects and their significances are summarized in Table 5.2. Form of teams is the most significant predictor for both responses, which corresponds with intuition. Surprisingly, the estimated coefficient for response goal differences is not equal to one as is fitted in model for estimation of forms but only 0.565. This means that filtered forms are more volatile than they should be.

The second most important factor is home advantage, which decreases in seasons and increases in rounds. The estimated coefficients for goal differences are

	Goal Differences			*Logits*		
Effect	**Estimate**	**T-val.**	**p-value**	**Estimate**	**T-val.**	**p-value**
HA	0.705	6.23	< 0.001	0.511	31.23	< 0.001
Season	-0.021	-2.79	0.005	-0.006	-5.82	< 0.001
Round	0.006	1.96	0.049	0.001	2.88	0.004
Form	0.565	14.07	< 0.001	0.509	87.66	< 0.001
Previous match[a]	-0.002	-0.17	0.869	0.003	1.94	0.052
Mutual matches[b]	0.094	2.61	0.009	0.081	15.59	< 0.001
Tiredness	-0.220	-2.09	0.037	0.014	0.91	0.361

Note: [a] Goal difference of the previous match.

[b] Mean goal difference of previous mutual matches.

Table 5.2: Comparison of different effects on goal differences and logits with estimates of forms calculated by rolling regression.

almost identical as for model with forms as categorical variable (compare with the Table 4.3). The estimated coefficient for round is no longer significant.

Mean goal difference of previous mutual matches has positive effect for both responses. In case of goal differences this contradicts results given in model with forms as categorical variables where the estimated coefficient was negative but insignificant. Change of sign could be caused by using filtered instead of smooth estimates of forms.

Variables goal difference of the previous match and tiredness are insignificant on confidence level 0.01 and in the following analysis are omitted. The fit is much better for logits with coefficient of determination $R^2 = 0.823$ than for goal differences where $R^2 = 0.139$ for TP. This indicates that betting companies use substantially quantitative information that is observed and measured.

5.2 Comparison of analytical model and betting odds

In this short section we compare analytical model and odds for estimating outcome of a match. For an outcome in terms of plain result (Win/Draw/Loss), we compare models based on G_n. We do need to know the exact probabilities estimates of betting companies to access its diversification power. We use the fact that estimated probabilities are monotonous transformation of odds and G_n is invariant on monotonous transformations. We only need to change sign for G_n because the transformation is decreasing.

We take goal differences as response and logit transformation of odds as the only predictor and calculate coefficient of determination R^2. We have calculated the value on the whole sample in section 4.4 because R^2 equals squared correlation coefficient between variables logits and goal differences (see Zvára, 2008, page 36). However, for comparison we use dataset TP with matches since tenth match for every team in every season. As analytical model we use model with forms estimated by rolling regression (see equation 4.23).

	R^2	G_n	G_w	G_t	G_l
Analytical model	0.137	0.290	0.358	0.033	0.359
Betting companies	0.154	0.312	0.378	0.063	0.378

Table 5.3: Goodness of fit criteria for analytical model with odds.

In both criteria the odds of betting companies outperforms the analytical model (see Table 5.3). From that point, we can conclude that betting companies use efficiently some extra information as current absence of key players. In that table we provide partial Gini coefficients, from which is G_n calculated. We see that the value of G_w and G_l is almost identical, it is consequence of using TP; they are not identical because not all matches are there twice since we use dataset with restriction that the team has played more than ten matches (not necessarily his opponent). Value G_t is close to zero, which suggests that correct estimation of probability of draw is more demanding than win.

5.3 Prediction of ice hockey match result

In this section we develop a model that could be used for estimation of probabilities of outcomes win, draw and loss. We have constructed such models using cumulative logit link for calculating Gini coefficients in the previous chapter. We have seen that odds by betting companies have greater diversification power than the analytical model. However, some predictors might be underestimated or overestimated and their combination with logits might result in a better model. The combined model is used to detect bets that are profitable in mean.

Firstly, we detect predictors, which are significant along with logits by using linear regression. In the second step we use the same predictors to fit ordinal regression. Theoretical background for this method is that goal differences can be viewed as a latent variable for the plain result (for detail see Cipra, 2008). The principle is expressed by the following equations:

$$\mathsf{P}(Y=0) = \mathsf{P}(GD \le \theta_0) = \mathsf{P}(\boldsymbol{x}^\top\boldsymbol{\beta} + \varepsilon \le \theta_0) = \mathsf{P}(\varepsilon \le \theta_0 - \boldsymbol{x}^\top\boldsymbol{\beta}) = F(\theta_0 - \boldsymbol{x}^\top\boldsymbol{\beta})$$
$$\mathsf{P}(Y\le 1) = \mathsf{P}(GD \le \theta_1) = \mathsf{P}(\boldsymbol{x}^\top\boldsymbol{\beta} + \varepsilon \le \theta_1) = \mathsf{P}(\varepsilon \le \theta_1 - \boldsymbol{x}^\top\boldsymbol{\beta}) = F(\theta_1 - \boldsymbol{x}^\top\boldsymbol{\beta})$$

where F is a distribution function of ε. In reality, the threshold parameter θ_0 could be anything between -1 and 0 and θ_1 between 0 and 1 but we assume goal differences to be continuous and hence, the parameters are to be estimated.

We fit linear regression with logits, forms and possibly significant predictors. We use dataset MP with data from tenth match to have estimates of forms via rolling regression. The estimates and their significances are summarized in Table 5.4. The measure of performance is insignificant along with logits as well as mean goal differences in mutual matches. Estimated coefficients of HA and round are positive, which suggests that home advantage is underestimated when odds are set up. This motivates us to compare strategies betting on a win of home team and on a win of guest team. It means in every match we bet on home (or guest) team and then calculate the average margin. If we bet on a win of home team,

Effect	Estimate	Std. error	T-val.	p-value
HA	0.175	0.123	1.43	0.15
Season	−0.015	0.008	−1.96	0.050
Round	0.004	0.003	1.54	0.125
Form	0.025	0.064	0.40	0.695
Logit	1.05	0.101	10.36	< 0.001
Mutual matches[a]	0.009	0.037	0.239	0.811

Note: [a] Mean goal difference of previous mutual matches.

Table 5.4: OLS estimates of linear regression with response goal differences.

we have average loss 2.6 %, for a win of guest team 20.8 % and for a draw 8.4 %. It means that random betting on a win of home team is more favourable than on a win of guest team. In some years, it leads even to a positive profit (see Table A2).

Since the predictor form is insignificant we omit it and fit another model with only HA, season, round and logits on the whole sample. In that model, fitted values seem to be overestimated for negative values and values higher than one (see Figure 5.2). This graph along with the Figure A3 motivates us to use some

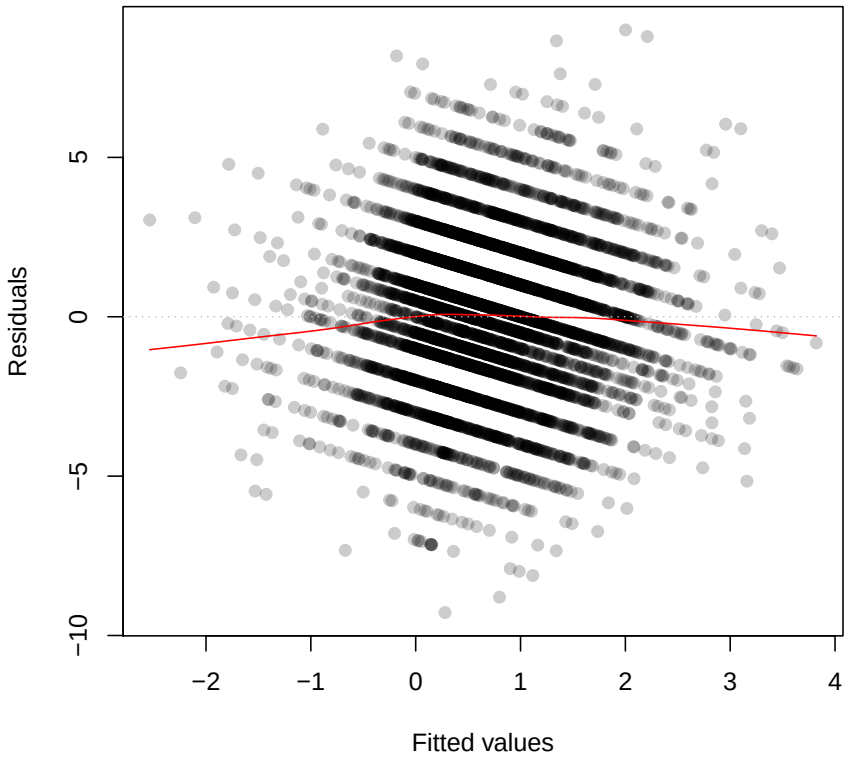


Figure 5.2: Residuals of linear model with response goal differences and predictors logits, HA, season and round.

Effect	**Estimate**	**Std. error**	**T-val.**	**p-value**
Threshold θ_0	-0.577	0.070	-8.29	< 0.001
Threshold θ_1	0.382	0.070	5.51	< 0.001
Season	-0.009	0.006	-1.60	0.110
Round	0.003	0.002	1.65	0.099
Logit	0.607	0.199	3.01	0.002
arctan(Logit)	0.301	0.278	1.08	0.279

Table 5.5: ML estimates of cumulative logit model with response result.

transformation of logits. We would like to diminish the effect of logits for its higher absolute values. We want also the function to be odd to maintain symmetry. Suitable transformation seems to be arcus tangens. We use it along with its linear form to fit the cumulative model. Hence, the model used for predicting probabilities of outcomes has the following form:

$$\begin{aligned} \text{logit}(\mathsf{P}(Y_{ijt} = 0)) &= \theta_0 - \beta_1 S_{ijt} - \beta_2 R_{ijt} - \beta_3 L_{ijt} - \beta_4 \arctan(L_{ijt}), \\ \text{logit}(\mathsf{P}(Y_{ijt} \leq 1)) &= \theta_1 - \beta_1 S_{ijt} - \beta_2 R_{ijt} - \beta_3 L_{ijt} - \beta_4 \arctan(L_{ijt}), \\ t &= 1, \ldots, n_{ij},\ i, j = 1, \ldots, I. \end{aligned} \tag{5.2}$$

To access probabilities, we use cumulative logit model. Estimated variables have expected signs (see Table 5.5). T-values are not high because the effect of logits is split between its linear and arcus tangens transformation. The odds to win in a season 1999/2000 at the first round after accounting for logits are $\exp(-\widehat{\theta}_1 - \widehat{\theta}_0) = 1.21$ higher home than out.

To compare this model with the others in terms of goodness of fit, we use dataset with matches for teams after tenth match. Goodness of fit criteria for this model are higher than for any previous model (see Table 5.6). Compared with odds of betting companies the biggest increase in a partial Gini coefficient was for a draw.

To see if this model would yield a positive margin in the history, we place a bet in case that the product of provided rate r and the estimated probability of outcome q is greater than one. It means we place a bet if the expected margin is greater than zero.

	R^2	G_n	G_w	G_t	G_l
Prediction model	0.158	0.323	0.386	0.084	0.386
Analytical model	0.137	0.290	0.358	0.033	0.359
Betting companies	0.154	0.312	0.378	0.063	0.378

Table 5.6: Goodness of fit criteria for prediction model and other models for comparison.

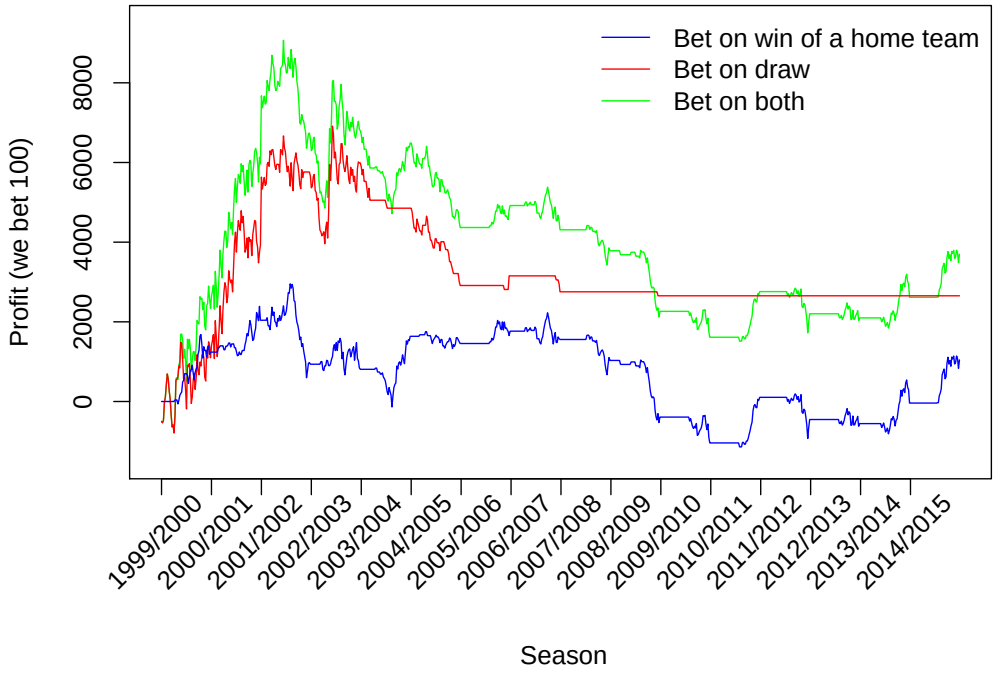


Figure 5.3: Profit of a betting strategy based on the prediction model (5.2). Lines show resulted cumulative profit if we bet 100 at every occasion when the expected profit is greater than zero.

To calculate the average margin of one bet according to our betting strategy we use the formula:

$$M = \frac{\sum_{i=0}^{N} \sum_{j=0}^{2} (r_i^j - 1) \mathbb{1}_{[r_i^j q_i^j > 1, Y_i = j]}}{\sum_{i=0}^{N} \sum_{j=0}^{2} \mathbb{1}_{[r_i^j q_i^j > 1]}}.$$

where $N = 5817$ is the total number of matches without missing values of odds. The resulted margin is $M = 1.78$ %. If we use the same formula but separately for win, draw and loss of home team, we find out that there wasn't a single bet on a loss, the average margin for a draw was 3.64 % and for a win 0.77 %.

We plot a figure simulating our strategy of betting. We start with 0 and every time when the expected margin is greater than one we bet 100. We plot the profit of our strategy for betting on draw and on win separately and together (see Figure 5.3). Most of the bets were placed at first seasons, specifically we have not placed a single bet on draw since season 2009/2010 and in four preceding seasons we bet only five times (see Table 5.7). Betting on a win of home team was primarily done in higher rounds of a season. This corresponds with increasing HA according to the prediction model.

The final cumulative profit of the proposed strategy is 3697 for 2081 bets, which suggests reasonable profitability. However, the profit since season 2004/2005 was mostly negative (see Table 5.7). It became profitable again in the last two seasons. The main disadvantage of this model is that it is highly dependent on

			Number of bets		
Season	Profit	Cumulative Profit	Draw	Win	Both
1999/2000	2339	2339	192	132	324
2000/2001	3626	5965	185	69	254
2001/2002	633	6598	116	188	304
2002/2003	201	6799	168	129	297
2003/2004	−309	6490	20	112	132
2004/2005	−2122	4368	41	50	91
2005/2006	552	4920	2	53	55
2006/2007	−609	4311	4	79	83
2007/2008	−526	3785	0	59	59
2008/2009	−1522	2263	1	65	66
2009/2010	−649	1614	0	63	63
2010/2011	1145	2759	0	49	49
2011/2012	−557	2202	0	62	62
2012/2013	−103	2099	0	57	57
2013/2014	516	2615	0	104	104
2014/2015	1082	3697	0	81	81

Table 5.7: Profit of the betting strategy in every season if we bet 100 at every occasion when the expected profit is greater than zero.

provided odds of betting companies and we lack the information how they create them.

The strategy might be changed and we may bet only in case that the expected margin is greater than some positive value. This would lower number of bets and should be more profitable on average. If we bet only in case when margin is greater than 1 % the realized margin is $M = 4.23$ %, for win it is 2.1 % and for draw 6.7 %. The total number of bets is 1340 and the total profit is 5670. The cumulative profit is in the figure A4. The reason that this strategy was more successful than the first one was that it stopped betting on draw earlier and that betting on home win was more successful in later seasons except the last one.

Conclusion

At the beginning of the thesis we presented main theoretical background that was needed for better understanding of derivation of the Kalman filter and for evaluation of models created in the practical part. For derivation of the Kalman filter we formulated and derived theorem for calculating the current state estimate in an original way. We have outlined two methods of hyperparameters estimation – maximum likelihood estimation and EM algorithm. In the last section we described how to handle missing observations in the Kalman filter.

Further, we have defined multivariate generalized linear models and showed that an ordinal categorical variable as a response with unspecified covariates satisfies the conditions. It was outlined how the estimation of a state vector could be derived in case that measurements come from a multivariate generalized linear model. This approach was not described in better detail as the usage of goal differences proved to be more informative than a plain result.

In the practical part we have analysed 16 seasons of Czech league ice hockey competition Extraliga matches in the main part. Based on specifics of Extraliga we have concluded that forms of teams are better to be estimated for every season independently. We discussed two different types of datasets that can be used for an analysis of paired comparisons and proved that estimates coincide in both linear and cumulative link models for symmetrical regressors.

We have examined several factors that may have an impact on the team form. The only significant predictor proved to be home advantage. Its positive effect is increasing in rounds and decreasing in seasons. We have examined the autocorrelation of results. It was necessary to filter the effect of home advantage because it typically changes after every match and the average form of the team in that season. Taking those factors into account, we found out that there is not a significant effect of a result from the previous match. Similarly, tiredness from the previous match and history of mutual matches were insignificant.

Several measures of team forms were proposed and discussed. Firstly, we observed that long history of results remains valid. For example we get better estimates of team forms using the knowledge of a current position in the league than by knowing five last results. Further, we saw that using goal differences leads to better team form estimates than using the plain result. Based on this conclusion we did not implement the discrete form of the Kalman filter but its linear form with goal differences as measurements and team forms as a state vector. However, the goodness of fit criteria did not indicate better prediction power than using weighted goal differences of the history. A disadvantage of the Kalman filter is the necessity of setting several hyperparameters, which influences the resulted estimates.

In the last chapter we compared the analytical model based on information

of previous results with odds of betting companies. For that purpose, we used the invariance of Gini coefficient on monotone transformations, which enabled to calculate its value directly from odds by betting companies. The odds of betting companies ($G_n = 0.312$) outperformed our model ($G_n = 0.290$). However, we have found out that betting companies underestimate the home advantage and that blind betting on a home team is not so much loss making as betting on a guest team.

At the end we have used the fact that odds are known before the actual match and can be used as a predictor along with other variables. We have combined a transformation of odds with the effect of home advantage to build a model ($G_n = 0.323$) and used it for betting. Our strategy for betting was to place a bet when the expected margin is positive. The strategy proved to yield a positive margin but it was highly profitable only at first seasons. Therefore, further analysis would be beneficial before using the strategy for the next season.

Bibliography

A. Agresti. Analysis of Ordinal Paired Comparison Data. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, 41(2):287–297, 1992.

B. D. O. Anderson and J. B. Moore. *Optimal filtering*. Prentice-Hall, 1979. ISBN 0-13-638122-7.

J. Anděl. *Základy matematické statistiky*. 2. edition. Matfyzpress, Praha, 2011. ISBN 80-7378-001-1.

J. Burridge. A note on maximum likelihood estimation for regression models using grouped data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(1):pp. 41–45, 1981. ISSN 00359246.

T. Cipra. *Finanční Ekonometrie*. 1. edition. Ekopress, Praha, 2008. ISBN 978-80-86929-43-9.

J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press Inc., 2001. ISBN 0-19-852354-8.

L. Fahrmeir. Posterior Mode Estimation by Extended Kalman Filtering for Multivariate Dynamic Generalized Linear Models. *Journal of the Amarican Statistical Association*, 87(418):501–509, 1992.

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. 1. edition. Springer, Praha, 1994. ISBN 0-387-94233-5.

L. Farhmeir and H. Kaufmann. On Kalman Filtering, Posterior Mode Estimation and Fisher Scoring in Dynamic Exponential Family. *Metrika*, 38:37–60, 1991.

L. Farhmeir and G. Tutz. Dynamic Stochastic Models for Time-Dependent Ordered Paired Comparison Systems. *Journal of the Amarican Statistical Association*, 89:1438–1449, 1994.

M. E. Glickman. *Paired Comparison Models with Time-Varying Parameters*. PhD thesis, Harvard University, 1993.

M. Hušková. *Bayesovské metody*. 1. edition. Matfyzpress, Praha, 1985.

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

W J H. Laird, N. M. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

C. P. Robert. *The Bayesian choice. From decision-theoretic foundations to computational implementation. 2nd ed.* New York, NY: Springer, 2nd ed. edition, 2007. ISBN 978-0-387-71598-8/pbk.

F. Šimsa. Kritéria těsnosti regrese dle typu vysvětlovaně proměnné. Bachelor's Thesis, Charles University in Prague, 2012.

A. Šulcová. Úrazovost v ledním hokeji v rámci mužských profesionálních soutěží v České republice. Master's thesis, Charles University in Prague, 2011.

O. Welch and G. Bishop. An introduction to the kalman filter. 2006. URL `http://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf`.

K. Zvára. *Regrese*. 1. edition. Matfyzpress, Praha, 2008. ISBN 978-80-7378-041-8.

List of Figures

List of Tables

List of Scripts

(On the attached CD)

VBA Excel

Download Data	downloads data from www.liga.cz

SQL

01 Tables	creates tables in schema Extraliga
02 Champions	determines winners of each season
03 TP with most predictors	creates data used in most analysis
Basic betting strategies	calculates margin of basic betting strategies
GD previous match	creates data for section 4.6
Measures of performance	creates data for section 4.9

R

Datasets	creates datasets TP and MP
Function of goals	script for section 4.4
Home advantage	script for section 4.5
Home advantage ACF	script for section 4.5
Autocorrelation of result	script for section 4.6
Mutual matches	script for section 4.7
Tiredness	script for section 4.8
MoP exponential smoothing	script for section 4.9
MoP simple KF	script for section 4.9
MoP part history	script for section 4.9
MoP rolling regression	script for section 4.9
KF parameter estimation	script for section 4.10
KF	script for section 4.10
Logits	script for sections 5.1
Comparison analytical model and odds	script for sections 5.2
Prediction	script for section 5.3

Appendix

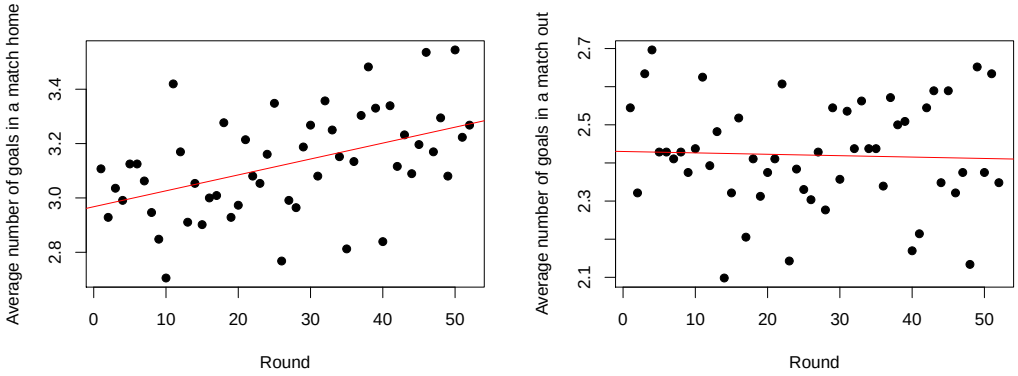


Figure A1: Figures show average number of goals scored by home team (left) and guest team (right) in rounds.

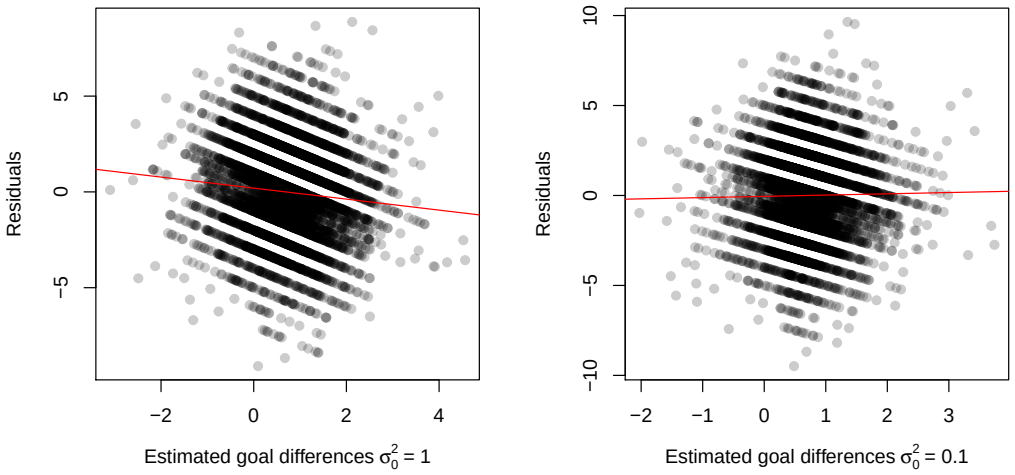


Figure A2: Residuals of Kalman filter defined by equations (4.27) and (4.28) with different value of initial variance σ_0^2 and red regression line.

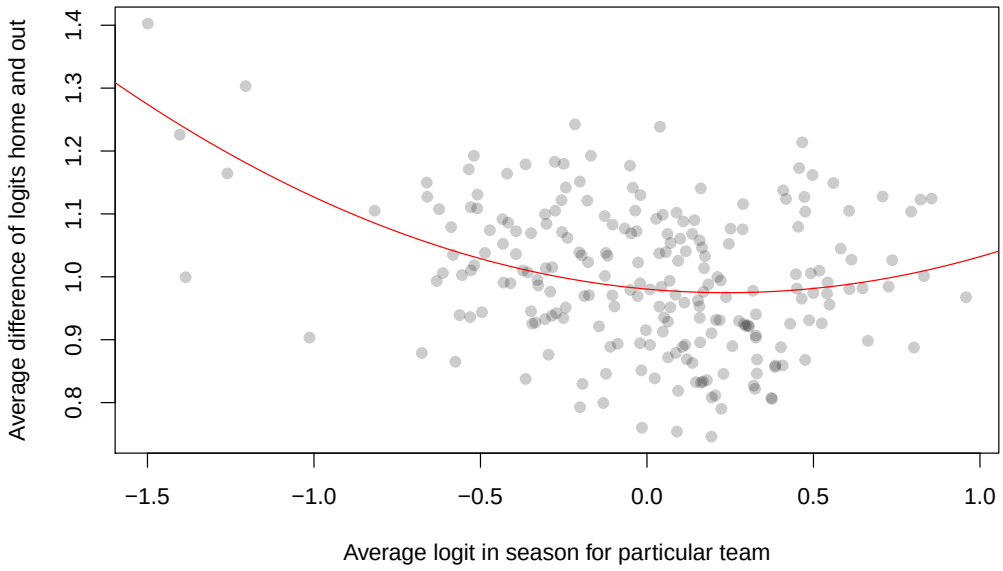


Figure A3: Dependence of difference in logits home and out on average logits.

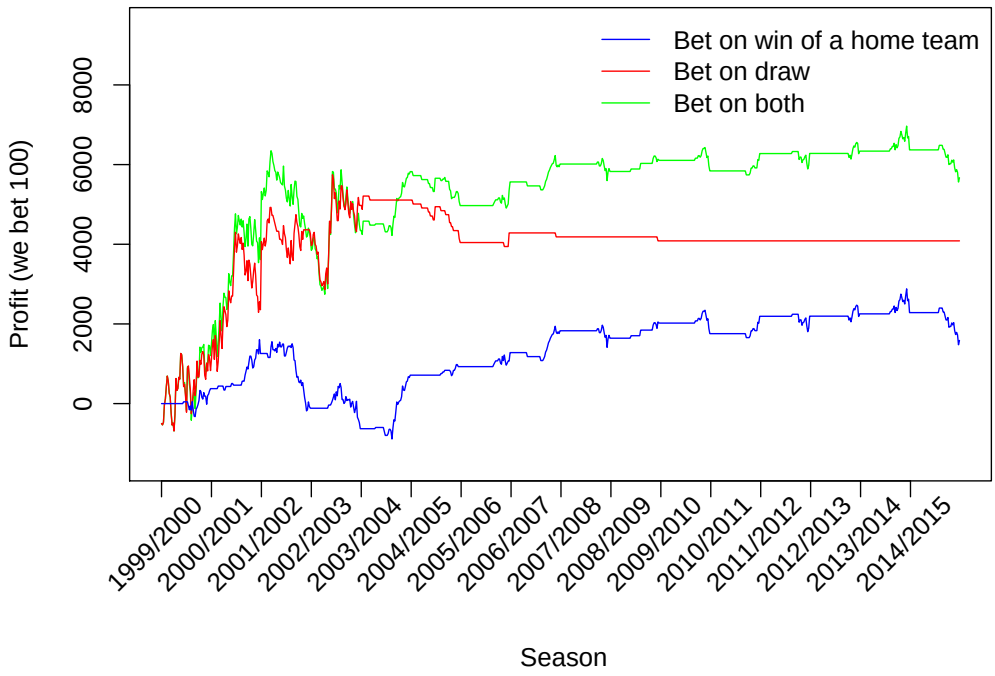


Figure A4: Profit of a betting strategy based on the prediction model (5.2). Lines show resulted cumulative profit if we bet 100 at every occasion when the expected profit is greater than 1 %.

	Days between matches								
Season	0	1	2	3	4	5	6	7	> 7
1999/2000	40	329	122	17	121	21	18	0	60
2000/2001	30	361	136	10	110	10	12	4	55
2001/2002	33	356	190	18	58	5	12	0	56
2002/2003	32	361	164	14	89	7	4	1	56
2003/2004	33	360	171	21	62	11	14	0	56
2004/2005	30	354	180	19	69	9	7	0	60
2005/2006	35	333	157	54	55	23	13	3	55
2006/2007	29	370	164	32	56	10	9	2	56
2007/2008	10	388	159	39	75	0	1	1	55
2008/2009	7	372	167	38	85	1	2	2	54
2009/2010	0	396	146	30	90	3	6	0	57
2010/2011	4	399	172	39	56	1	0	0	57
2011/2012	6	399	162	25	47	15	12	11	51
2012/2013	4	393	133	34	98	3	3	5	55
2013/2014	0	405	116	25	117	4	3	2	56
2014/2015	2	408	105	13	135	0	8	1	56

Table A1: Historical data of number of days between matches for all teams in main part of all season.

Season	Bet on home team	Bet on draw	Bet on guest team
1999/2000	-0.007	-0.040	-0.323
2000/2001	-0.014	-0.020	-0.280
2001/2002	-0.096	-0.033	-0.152
2002/2003	-0.001	-0.036	-0.296
2003/2004	0.035	-0.110	-0.340
2004/2005	-0.074	-0.256	-0.157
2005/2006	-0.020	-0.071	-0.240
2006/2007	0.036	-0.204	-0.216
2007/2008	-0.054	0.120	-0.300
2008/2009	-0.017	0.004	-0.222
2009/2010	-0.111	-0.003	-0.121
2010/2011	-0.024	-0.144	-0.150
2011/2012	-0.055	0.019	-0.198
2012/2013	0.005	-0.081	-0.163
2013/2014	-0.052	-0.162	-0.106
2014/2015	0.040	-0.320	-0.065

Table A2: Realized margin in different seasons and betting at each match on either home win, draw or away win.

Method	Last matches	Goodness of fit criteria				
		R^2	G_n	G_C	G_1	G_2
Position in table		0.120	0.278	0.310	0.344	0.344
Goal differences		0.136	0.292	0.324	0.358	0.358
	1	0.087	0.232	0.261	0.290	0.290
	2	0.092	0.243	0.274	0.305	0.305
	3	0.094	0.245	0.274	0.305	0.305
	4	0.101	0.256	0.288	0.319	0.319
	5	0.104	0.259	0.290	0.322	0.322
	6	0.108	0.267	0.296	0.328	0.328
	7	0.113	0.270	0.301	0.334	0.334
	8	0.117	0.275	0.306	0.340	0.340
	9	0.118	0.279	0.309	0.342	0.342
	10	0.121	0.284	0.312	0.346	0.346
Points		0.128	0.285	0.316	0.350	0.350
	1	0.086	0.227	0.255	0.284	0.284
	2	0.089	0.236	0.268	0.298	0.298
	3	0.090	0.239	0.267	0.297	0.297
	4	0.096	0.247	0.278	0.309	0.309
	5	0.099	0.253	0.284	0.316	0.316
	6	0.102	0.260	0.289	0.321	0.321
	7	0.107	0.263	0.294	0.327	0.327
	8	0.111	0.267	0.300	0.333	0.333
	9	0.112	0.271	0.302	0.335	0.335
	10	0.115	0.276	0.305	0.338	0.338

Table A3: Goodness of fit criteria based on model 4.20 with measures of performances based on mean differences calculated from specified number of last matches. Methods without specified number use all history up to a given time.

		Goodness of fit criteria				
Method	a	R^2	G_n	G_C	G_1	G_2
Goal differences		0.136	0.292	0.324	0.358	0.358
	0.05	0.138	0.295	0.327	0.362	0.362
	0.1	0.131	0.290	0.321	0.355	0.355
	0.2	0.116	0.275	0.306	0.339	0.339
	0.3	0.106	0.264	0.295	0.327	0.327
	0.4	0.100	0.256	0.286	0.318	0.318
	0.5	0.096	0.250	0.281	0.312	0.312
	0.6	0.093	0.245	0.276	0.306	0.306
	0.7	0.091	0.241	0.271	0.302	0.302
	0.8	0.089	0.237	0.268	0.297	0.297
	0.9	0.088	0.234	0.264	0.294	0.294
Points		0.128	0.285	0.316	0.350	0.350
	0.05	0.130	0.288	0.319	0.353	0.353
	0.1	0.123	0.282	0.312	0.346	0.346
	0.2	0.109	0.267	0.298	0.331	0.331
	0.3	0.100	0.256	0.286	0.318	0.318
	0.4	0.095	0.248	0.278	0.309	0.309
	0.5	0.092	0.242	0.272	0.302	0.302
	0.6	0.090	0.237	0.267	0.297	0.297
	0.7	0.088	0.234	0.264	0.293	0.293
	0.8	0.087	0.231	0.260	0.290	0.290
	0.9	0.087	0.229	0.258	0.287	0.287

Table A4: Different measures of performance defined according to 4.21 with smoothing parameter a. Methods without specified a are mean differences of those measures.