MAS8391: MMathStat Project

School of Mathematics & Statistics

Bayesian Analysis of Paired Comparison Data

Author:
Maciej Misiura

Supervisor:
Dr. Daniel Henderson

April 30, 2015

Bayesian Analysis of Paired Comparison Data

MAS8391: MMathStat Project

Maciej Misiura

Abstract

This project aims to provide an extensive review of models for paired comparison data. Two key models for paired comparison data, the Bradley–Terry model and the Thurstone model, are thoroughly examined to identify any shared characteristics. A Bayesian approach to inference is adopted. Fundamental concepts, such as Markov chain Monte Carlo (MCMC) methods are introduced. A methodology is presented for fitting Bayesian versions of models for paired comparisons to sports data with a binary outcome. A National Collegiate Athletic Association (NCAA) basketball dataset is analysed to predict the outcomes of the 2015 "March Madness" tournament to determine the best men's collegiate team in the United States.

Contents

Chapter 1

Introduction

The roots of probability theory originate as far back as the 17th Century, when prominent mathematicians such as Pierre de Fermat and Blaise Pascal tried to analyse games of chance including roulette and poker. Nowadays, many statisticians across the world attempt to analyse vast quantities of sports data in order to model uncertainty and thus forecast results of a given match or event. A very important question arises naturally. Is it possible to derive a statistical model which improves the possibility of correctly predicting the outcome of an event other than just guessing? One of the most famous 20th Century statisticians, George E. P. Box, once quoted "*all models are wrong, but some are useful*". The focus of this paper is to review some of the most fundamental models of predicting probabilities for paired comparison data and to fit some Bayesian versions of these models to National Collegiate Athletic Association (NCAA) basketball data.

1.1 Paired comparison data

There are numerous areas in which data resulting from paired comparisons (between individuals, items, teams, etc) can arise. A published bibliography on this subject includes several hundred entries (Hunter 2004). Analysis of paired comparison data can be traced back to the early 20th Century in the field of psychometrics with Louis Leon Thurstone at the forefront of such research. It was his law of comparative judgment (Thurstone 1927) that completely revolutionised mathematical analysis of such data. Rather than letting test subjects in his experiments rank the items in order of their preferences, Thurstone asked these individuals to select the favoured option between two possible outcomes. He then applied a type of binomial regression to the information collected. In general, paired comparison data can be considered to be the result of a series of Bernoulli trials with success and failure defined depending upon the setting. Paired comparisons considered in this report occur in sports data, and in particular NCAA basketball data, and an example of the nature of questions to be asked is: "What is the probability of Connecticut Huskies beating Kentucky Wildcats?"

The rest of this report is structured as follows. Chapter 2 outlines key models for paired comparison data. Chapter 3 considers a Bayesian approach to inference via Markov chain Monte Carlo (MCMC) methods, as well as the need for specific computational software. Chapter 4 describes an application of the methods to data from the NCAA basketball tournament. Chapter 5 offers conclusions as well as further discussion.

Chapter 2

Models for paired comparisons

2.1 The Bradley–Terry model

The Bradley–Terry model is arguably the most important and well studied model for paired comparison data. In this chapter, the Bradley–Terry model is presented as follows: firstly, it will be introduced in its original, most intuitive form, then some of its applications as well as extensions will be presented.

2.1.1 Definition

The Bradley–Terry model (Bradley & Terry 1952) assumes that in a contest between two teams, say team i and team j ($i, j \in \{1, \ldots, K\}$), the probability that i beats j is

$$Pr(i \text{ beats } j) = \frac{\lambda_i}{\lambda_i + \lambda_j}, \tag{2.1}$$

where $\lambda_z > 0$ is a parameter associated to element $z \in \{1, \ldots, K\}$ representing the skill rating or strength of a team. For example, as a quick illustration, suppose there are $K = 3$ teams with strength parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3) = (1, 2, 3)$. The probability that team 2 beats team 3 can be calculated as follows

$$Pr(2 \text{ beats } 3) = \frac{\lambda_2}{\lambda_2 + \lambda_3} = \frac{2}{2+3} = 0.4.$$

It is very interesting to point out that the model described in Equation (2.1) was also derived separately and independently by Zermelo (1929) and Ford (1957). Note that the probability that i beats j is invariant to scalar multiplication of the strength parameters; when the λ parameters are multiplied by $a \neq 0$, then

$$Pr(i \text{ beats } j) = \frac{a\lambda_i}{a\lambda_i + a\lambda_j} = \frac{a\lambda_i}{a(\lambda_i + \lambda_j)} = \frac{\lambda_i}{\lambda_i + \lambda_j}.$$

Therefore a constraint is needed in order to be able to identify the parameters in this model. Various different constraints can be imposed, for example a sum constraint,

$\sum_{i=1}^{K} \lambda_i = 1$, or fixing one of the skill parameters at a particular value. In this report, the latter type of constraint is used and we set $\lambda_1 = 1$. Independence between the skill ratings is also assumed.

2.1.2 Applications

There are numerous potential scenarios to which the Bradley–Terry model can be applied. The majority of research that implements this model is related to either the fields of medicine, psychology or sport.

For instance, Matthews & Morris (1995) adapted this model to the measurement of pain in patients undergoing long–term haemodialysis. Each patient was given two different treatments and then was asked to specify which treatment was less painful. The efficiency of each treatment was then analysed using the Bradley–Terry model with ties, which is discussed in Section 2.1.3 of this report.

Paired comparison data arise frequently in psychology, because of the nature of the research conducted. In many studies, participants are asked to choose between two scenarios or rank their preferences. For example, facial attractiveness and the preference for either upright or upside–down faces has been discussed by Bäuml (1994). On the other hand, Kissler & Bäuml (2000) derived an attractiveness scale by fitting a Plackett–Luce model (an extension of the Bradley-Terry model to experiments with more than two possible outcomes) and Duinevald et al. (2000) analysed data on customers' preference between eight carbonated soft drinks.

Several authors employ Bradley-Terry type models to forecast various sporting tournaments. McHale & Morton (2011) use it to predict tennis match results whilst Baker & McHale (2014) apply it to determine who is the greatest tennis player of the Open–Era. Recently, dynamic extensions to the Bradley–Terry model have been proposed. Cattelan et al. (2013) introduce such a model to analyse data from the National Basketball Association (NBA), whereas Tutz & Schauberger (2014) fit it to football data from the German Bundesliga. However, the implementations of the Bradley–Terry model are not limited to the above areas. Perhaps one of the most interesting and unusual applications of this model can be found in Stuart-Fox et al. (2006), in which the Bradley–Terry model was used to determine animal dominance and fighting ability between 36 male Cape dwarf chameleons.

2.1.3 Extensions of the original model

Model with ties

The original Bradley–Terry model has its limitations. One of its biggest drawbacks is the fact that it cannot be used to model events with more than two outcomes. In the measurement of pain example presented by Matthews & Morris (1995), a patient could have no preference between the two treatments. In this instance, the Bradley–Terry model

becomes obsolete. Rao & Kupper (1967) suggested the following model as a solution to this problem,

$$Pr(i \text{ beats } j) = \frac{\lambda_i}{\lambda_i + \tau\lambda_j}$$
$$Pr(i \text{ ties } j) = \frac{(\tau^2 - 1)\lambda_i\lambda_j}{(\lambda_i + \tau\lambda_j)(\tau\lambda_i + \lambda_j)},$$

where τ, $\tau > 1$ is the ties parameter.

Home advantage

In sports such as basketball, there tends to be a so–called "home advantage", whereby the chance a team wins is higher when it plays at home compared to it playing at a neutral venue. According to Snyder & Purdy (1985), home teams in collegiate basketball win 66% of their games, indicating very strong evidence of this phenomenon. One may speculate that this is due to spectator booing: an away team might get intimidated by the opposition's fans and thus lose their self belief and confidence (Greer 1983). Often the travelling team employs conservative tactics, setting up to defend, and this could also be a plausible factor. However, contrary to most research, several authors, see for example Baumeister & Steinhilber (1984), indicate that there might exist "home disadvantage", when a host team cannot cope with additional pressure and, subsequently, is more likely to lose a match. Empirically, this counter hypothesis has not been confirmed (Jones 2014).

The most accepted way of accounting for home advantage in the Bradley–Terry model was proposed by Agresti (1990). He suggested the following model

$$Pr(i \text{ beats } j) = \begin{cases} \dfrac{\rho\lambda_i}{\rho\lambda_i + \lambda_j} & \text{if } i \text{ is home} \\ \dfrac{\lambda_i}{\lambda_i + \rho\lambda_j} & \text{if } j \text{ is home.} \end{cases} \tag{2.2}$$

The parameter $\rho, \rho > 0$, measures the strength of the home advantage ($\rho > 1$) or disadvantage ($\rho < 1$). Again, for illustrative purposes, assume $K = 3$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3) = (1, 2, 3)$. Suppose team 2 plays against team 3 in front of its own supporters and that it rarely loses a home match. It was elicited that $\rho = 2$, indicating a strong home advantage. Then

$$Pr(2 \text{ beats } 3 \mid 2 \text{ is at home}) = \frac{2 \times 2}{2 \times 2 + 3} = \frac{4}{7} = 0.571.$$

This probability of team 2 winning this match with team 3 changed from 0.4 to 0.571 when team 2 is at home, showing that home advantage can play a major factor.

2.2 The Thurstone model

A closely related model to the Bradley–Terry model is the Thurstone model (Thurstone 1927). In the literature, this model is frequently referred to as the Thurstone–Mosteller

model. Frederick Mosteller largely contributed to the development and analysis of this model by calculating the least squares estimate as well as deriving a test of goodness of fit (Mosteller 1951).

2.2.1 Definition

The Thurstone model assumes that the performance of team i, $i \in \{1, \ldots, K\}$ follows a normal distribution $X_i \sim N(\mu_i, \sigma_i^2)$. In Thurstone's original paper, five different variations of this model were proposed. However, due to its simplicity and elegance, only the Case V model is considered here. The two underlying assumptions in the Case V model are such that team performances are uncorrelated and variances are equal, that is $\sigma_i^2 = \sigma^2 = 1/2$, for all i. In this setting, without loss of generality, the difference between the skill rating of team i and team j follows a normal distribution $X_i - X_j \sim N(\mu_i - \mu_j, 1)$ and thus

$$\begin{aligned} Pr(i \text{ beats } j) &= P(X_i > X_j) \\ &= P(X_i - X_j > 0) \\ &= P\left(X_j - X_i - \{\mu_i - \mu_j\} > -\{\mu_i - \mu_j\}\right) \\ &= P\left(Z > -\{\mu_i - \mu_j\}\right), \qquad \text{where } Z \sim N(0,1) \\ &= P\left(Z < \mu_i - \mu_j\right) \\ &= \Phi\left(\mu_i - \mu_j\right), \end{aligned} \tag{2.3}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF).

2.2.2 Applications

Perhaps the most well known application of the Thurstone model is to the FIDE World Rankings in chess. First proposed by Elo (1978), this subject has been of interest to many statisticians and has been thoroughly documented in numerous research journals. Alternative modifications of the Elo rating system can be found for example in Henery (1992) and Glickman (1999), whilst Joe (1990) extends such a model to make use of various covariates including the age of a player.

The Thurstone model has found numerous applications in the field of medicine. Maydeu-Olivares & Böckenholt (2008) provide a list of ten reasons why this particular model should be used to make inferences for subjective health outcomes. Amongst them, they list the ease of incorporating prior information into the model as well as being able to test the validity of the results by using methods such as goodness of fit. Health indicators for a particular individual, population, region, etc can be measured using the Thurstone model, as seen in Kind (1982). Quantifying subjective outcomes, such as severity of side effects or amount of pain can also be accomplished using this model, as proposed by Krabbe (2008).

2.3 Generalised linear models

In the normal linear model the ith observation Y_i has a systematic component μ_i and a random component ϵ_i with

$$Y_i = \mu_i + \epsilon_i, \tag{2.4}$$

where the ϵ_i have a normal distribution with constant variance and are independent. The systematic component μ_i can be represented as

$$\mu_i = \sum_{j=1}^{p} \beta_j x_{ij},$$

where $\beta_1, \ldots, \beta_p$ are parameters and x_{ij} is the value of covariate j for observation i. The model specified in Equation (2.4) is often too restrictive and cannot be applied in a wide range of scenarios. For example, when the outcome is binary, that is $Y_i = 0$ or $Y_i = 1$, or strictly positive ($Y_i > 0$), the assumption of normal variation is not feasible. Nelder & Wedderburn (1972) proposed a unified approach called the generalised linear model (GLM). This model can be defined in three stages, as defined by Gelman et al. (2004):

1. The *linear predictor*, $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$,
2. The *link function* $g(\cdot)$ that relates the mean μ_i and the linear predictor: $\eta_i = g(\mu_i)$ or equivalently $\mu_i = g^{-1}(\eta_i)$. The link function must be monotonic and differentiable.
3. The error distribution which then specifies the distribution of the outcome variable Y_i. It can be chosen from an exponential family of distributions, which includes (amongst many others) normal, Poisson, binomial and gamma.

In binomial type problems, that is where $Y_i \sim \text{Bin}(n_i, \mu_i)$ with n_i known, the most common link function is a sigmoid based on the distribution function of the logistic distribution. It is most commonly known as the logistic transformation, $g(\mu_i) = \log(\mu_i/(1-\mu_i))$. Other link functions can also be used, for example, the probit link, $g(\mu_i) = \Phi^{-1}(\mu_i)$, where $\Phi(\cdot)$ is the standard normal CDF and $\Phi^{-1}(\cdot)$ is its inverse. These two link functions give rise to the logistic and probit regression models, respectively.

2.3.1 Paired comparison models and binary regression

There exists a clear relationship between the Bradley–Terry model as defined in Equation (2.1) and the logistic regression model. By letting $\theta_i = \log \lambda_i$, it is possible to express the Bradley-Terry model in the logit linear form

$$\begin{aligned} \text{logit}\{Pr(i \text{ beats } j)\} &= \log\left(\frac{\lambda_i/(\lambda_i - \lambda_j)}{\lambda_j(\lambda_i - \lambda_j)}\right) \\ &= \log \lambda_i - \log \lambda_j \\ &= \theta_i - \theta_j. \end{aligned} \tag{2.5}$$

In this representation, home advantage can be accounted for easily as follows

$$\text{logit}\left\{Pr(i \text{ beats } j)\right\} = \delta + (\theta_i - \theta_j), \tag{2.6}$$

where $\delta = \log(\rho)$ represents the extent of home advantage ($\delta > 0$) or disadvantage ($\delta < 0$).

One of the biggest advantages of a GLM (logistic regression) formulation of the Bradley–Terry model is the fact that it can be readily modified to include various covariates. For example, by letting

$$\eta_k = \text{logit}\left\{Pr(i \text{ beats } j)\right\} = (\theta_i - \theta_j) + \sum_{r=1}^{p} \beta_r x_{kr},$$

in which the probability that team i beats team j in match k is related to the explanatory variables $x_{k1}, \ldots, x_{kp}$ through a linear predictor with coefficients $\beta_1, \ldots, \beta_p$. Hence, if there is a strong belief that distance travelled, points scored in the previous game, or a whole multitude of other factors affect team performance, then this information may be incorporated into the original Bradley–Terry model.

Similarly, the Thurstone model defined in Equation (2.3) is related to probit regression. For example, it can be represented, by letting $\theta_i = \mu_i$, in the probit linear form

$$\Phi^{-1}\left\{Pr(i \text{ beats } j)\right\} = \theta_i - \theta_j. \tag{2.7}$$

Both probit and logistic regression yield very similar results when fitted to data. Chambers & Cox (1967) proved that it is almost impossible to distinguish between these two models. Indeed, the only tangible difference between these binary regression models lies in the error distribution. In the logistic model, errors are assumed to follow a standard logistic distribution, whilst errors in the probit model are normally distributed. Figure 2.1 illustrates the similarity between the two density and distribution functions.

2.4 Equivalence of paired comparison models

Bradley explored the logistic regression formulation of the Bradley-Terry model in his two papers (Bradley 1953, 1965). He noted that by letting $\theta_i = \log \lambda_i$, $\theta_i - \theta_j$ could be regarded as the difference between the skill rating of team i and team j, which follows a squared hyperbolic secant density (the density of a logistic random variable). Thus the probability that i beats j can be expressed as follows

$$Pr(i \text{ beats } j) = Pr(X_i > X_j) = \int_{-(\theta_i - \theta_j)}^{\infty} \frac{1}{4}\text{sech}^2\ \frac{x}{2} dx \equiv \frac{\lambda_j}{\lambda_i + \lambda_j}.$$

As illustrated in Figure 2.1, the probability density function (pdf) of a standard logistic distribution is symmetric, centered around a mean of zero and has unit variance. Several other authors, such as Kuk (1995) consider the Thurstone and the Bradley–Terry model

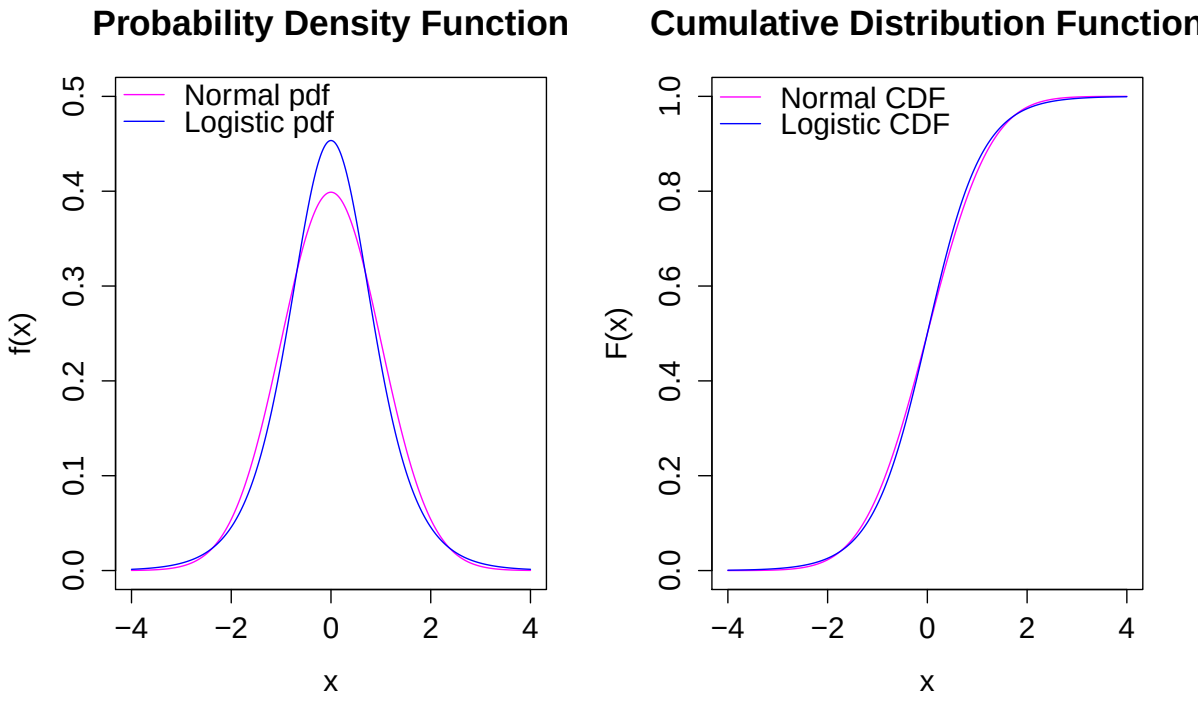


Figure 2.1: Probability density functions (left) and cumulative distribution functions (right) of the normal and logistic distributions with mean 0 and variance 1

to be special cases of the linear paired comparison model (David 1988). The probability that i beats j can then be written as

$$Pr(i \text{ beats } j) = F(\theta_i - \theta_j) \tag{2.8}$$

where F is a distribution function that is symmetric about 0 and $\theta_1, \ldots, \theta_K$ correspond to the strength parameters. Therefore, it can be shown that when $F = \Phi$, the standard normal CDF, this is the Thurstone model. Similarly, when F corresponds to the logistic CDF, the linear paired comparison model takes the form of the Bradley–Terry model presented in Equation (2.1) with $\lambda_i = \exp \theta_i$.

On the other hand, Stern (1990) recommends to treat the Thurstone model and the Bradley–Terry model as special cases of a gamma paired comparison model, which he defined as follows. Suppose that the number of points team i scores follows a Poisson process with rate λ_i. Then the time X_i until team i scores r points follows a Gamma(r, λ_i) distribution. Team i beats team j if and only if they score r points before their opposition. As the number of points scored by each team is an independent process, the probability that i beats j can be regarded as the probability that $X_i \sim \text{Gamma}(r, \lambda_i)$ is smaller than $X_j \sim \text{Gamma}(r, \lambda_j)$. Thus

$$Pr(i \text{ beats } j) = Pr(X_i < X_j) = \int_0^\infty \int_0^{x_j} \frac{\lambda_i^r x_i^{r-1} \exp(-\lambda_i x_i)}{\Gamma(r)} \times \frac{\lambda_j^r x_j^{r-1} \exp(-\lambda_j x_j)}{\Gamma(r)} dx_i dx_j.$$

When the shape parameter $r = 1$, Stern recognises this as the simple Bradley–Terry

model defined in Equation (2.1). When r is a large positive integer, then the gamma model resembles the Thurstone model from Equation (2.3). As r increases (from $r = 1$), Stern's model provides a bridge between the Bradley–Terry model and the Thurstone model.

Baker & McHale (2014) showed that a closed form expression for the probability that team i beats team j can be derived under Stern's gamma paired comparison model. An alternative derivation of this result is shown below, where the probability that i beats j can be expressed as

$$\begin{aligned} Pr(i \text{ beats } j) &= Pr(X_i < X_j) \\ &= Pr(\lambda_i X_i < \lambda_i X_j) \\ &= Pr\left(\lambda_i X_i < \frac{\lambda_i}{\lambda_j}\lambda_j X_j\right) \\ &= Pr\left(Z < \frac{\lambda_i}{\lambda_j} S\right), \end{aligned}$$

where $Z = \lambda_i X_i \sim \text{Gamma}(r, 1)$ and, independently, $S = \lambda_j X_j \sim \text{Gamma}(r, 1)$, therefore

$$\begin{aligned} Pr(i \text{ beats } j) &= Pr\left(\frac{Z}{S} < \frac{\lambda_i}{\lambda_j}\right) \\ &= Pr\left(\frac{\frac{Z}{S}}{\frac{Z}{S}+1} < \frac{\frac{\lambda_i}{\lambda_j}}{\frac{\lambda_i}{\lambda_j}+1}\right) \\ &= Pr\left(\frac{Z}{Z+S} < \frac{\lambda_i}{\lambda_i+\lambda_j}\right) \\ &= Pr\left(Y < \frac{\lambda_i}{\lambda_i+\lambda_j}\right), \end{aligned}$$

where $Y = Z/(Z + S) \sim \text{Beta}(r, r)$ since Z and S are independent $\text{Gamma}(r, 1)$ random variables. In other words, the probability that i beats j is the distribution function of a $\text{Beta}(r, r)$ random variable, evaluated at the Bradley-Terry type ratio $\lambda_i/(\lambda_i + \lambda_j)$,

$$\begin{aligned} Pr(i \text{ beats } j) &= P\left(Y < \frac{\lambda_i}{\lambda_i+\lambda_j}\right) \\ &= \frac{\Gamma(r+r)}{\Gamma(r)\Gamma(r)} \int_0^{\frac{\lambda_i}{\lambda_i+\lambda_j}} y^{r-1}(1-y)^{r-1}\mathrm{d}y. \end{aligned}$$

This Beta CDF can be readily computed in software such as **R**. When $r = 1$, $Y \sim \text{Beta}(1, 1) \equiv U(0, 1)$ and so $Pr(i \text{ beats } j) = \lambda_i/(\lambda_i + \lambda_j)$, the probability associated with the Bradley–Terry model in Equation (2.1). This closed-form representation makes it easy to fit Stern's model, although it is debatable, given the similarity between the Bradley-Terry model and the Thurstone model, whether this additional flexibility is worthwhile.

Chapter 3

Bayesian inference for paired comparisons

3.1 Introduction

Most of the statistical analysis of paired comparison data has been performed from the frequentist point of view. Recently however, several authors have proposed to perform Bayesian inference for Bradley–Terry type models. Perhaps one of the earliest entries related to this subject was published by Davidson & Solomon (1973). It provides estimators of the skill parameters in the Bradley–Terry model. This topic is further discussed in Leonard (1977) and then inferences are made regarding '*the distribution of genital display in a colony of six squirrel monkeys*'. More recently Yao & Böckenholt (1999) suggested an efficient Bayesian procedure for inferring the parameters of Thurstonian models based on the Gibbs sampler, whilst Caron & Doucet (2012) proposed efficient Gibbs samplers for generalised Bradley–Terry models. In the next section, a brief overview of the Bayesian approach to inference will be given, followed by an overview of Markov chain Monte Carlo methods.

3.2 Overview of Bayesian methods

Consider a continuous (vector) parameter $\boldsymbol{\Theta}$ and observed data $\boldsymbol{X}$. Both data and parameters are random variables. Prior beliefs about the parameters define the prior distribution for $\boldsymbol{\Theta}$, specified by the density $\pi(\boldsymbol{\theta})$. A model is then formulated that defines the distribution of $\boldsymbol{X}$ given the parameters, in other words a density $f_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}|\boldsymbol{\theta})$ is specified. This can be regarded as a function of $\boldsymbol{\theta}$ when there exists some fixed observed data $\boldsymbol{x}$, called the likelihood

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = f_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}|\boldsymbol{\theta}).$$

The prior and likelihood determine the full joint density over data and parameters

$$f_{\boldsymbol{\Theta},\boldsymbol{X}}(\boldsymbol{\theta},\boldsymbol{x}) = \pi(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{x}).$$

Given the joint density, it is then possible to compute its marginals as well as conditionals

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} f_{\boldsymbol{\Theta},\boldsymbol{X}}(\boldsymbol{\theta},\boldsymbol{x})d\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}$$

and

$$f_{\boldsymbol{\Theta},\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f_{\boldsymbol{\Theta},\boldsymbol{X}}(\boldsymbol{\theta},\boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})} = \frac{\pi(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{x})}{\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}},$$

where $f_{\boldsymbol{\Theta},\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x})$ is most commonly known as the posterior density, and is usually denoted $\pi(\boldsymbol{\theta}|\boldsymbol{x})$. This leads to the continuous version of Bayes' theorem

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\pi(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{x})}{\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}}.$$

The denominator is not a function of $\boldsymbol{\theta}$, so we can in fact write this as

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \propto \pi(\boldsymbol{\theta})L(\boldsymbol{\theta}|\boldsymbol{x}),$$

where the constant of proportionality is chosen to ensure that the density integrates to one. Hence, the posterior is proportional to the prior times the likelihood. For most models, the posterior is not available in closed form. One way to make inferences is to sample values from this posterior distribution. This is most often done using Markov chain Monte Carlo (MCMC) methods, which rely on sampling from a Markov chain whose stationary distribution is the posterior distribution of interest. Before describing MCMC in more detail, a brief introduction to Markov chains is given.

3.2.1 A quick note on Markov chains

The theory behind Markov chains can be traced back to the early 20th Century. Two mathematicians simultaneously developed the theory behind this subject: Henri Poincare and Andrei Markov. Markov, Pafnuty Chebyshev's protégé, was interested in the occurrence of vowels and consonants in the famous novel written by Alexander Pushkin, *Eugene Onegin*. For ease of notation, consider a discrete–time stochastic process $\{X_t\}$, $t = 0, 1, 2, \ldots$, where X_t is a discrete random variable defined on a finite or countably infinite state space $\mathcal{S}$. Then the random variables $X_0, X_1, X_2, \ldots$ form a homogeneous Markov chain with state space $\mathcal{S}$ if

$$Pr(X_{t+1} = j|X_t = i, X_{t-1}, \ldots, X_0) = Pr(X_{t+1} = j|X_t = i)$$

for all t and for $i, j \in \mathcal{S}$. In other words, a Markov chain is a stochastic process whose future state is only dependent on the current state and is independent of the past. This lack of dependence on the past, also known as the Markov property, allows for significant simplification of complex problems and forms the basis of Markov chain Monte Carlo methods.

3.3 Markov chain Monte Carlo methods

Simulating Markov chains enables us to obtain realisations from virtually any posterior distribution, regardless of how complicated it is. Perhaps the most common way of simulating these chains are Metropolis–Hastings algorithms. Metropolis et al. (1953) came up with a computer algorithm that enabled study of the properties of chemical substances based on collisions between individual particles. Hastings (1970) on the other hand, modified this algorithm and discussed how it could be used in the field of statistics. The draws from the posterior distribution can be obtained by considering reversible Markov chains (a Markov chain can be called reversible, if the reverse order sequence of states also forms a Markov chain). The following algorithm can then be defined (Gammerman & Lopes 2006):

1. Initialise the iteration counter to $j = 1$ and set an arbitrary initial value $\boldsymbol{\theta}^{(0)}$.
2. Propose a new value $\boldsymbol{\phi}$ generated from a proposal density $q(\boldsymbol{\theta}^{(j-1)}, \cdot)$.
3. Evaluate the acceptance probability of the move, $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\phi})$, defined by
$$\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) = \min\left\{1, \frac{\pi(\boldsymbol{\phi})q(\boldsymbol{\phi}, \boldsymbol{\theta})}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}, \boldsymbol{\phi})}\right\}.$$
4. If the move is accepted, set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\phi}$. If it is not accepted, set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$ and the chain does not move.
5. Change the iteration counter from j to $j + 1$ and return to step 2 until convergence is reached.

A special case of the Metropolis–Hastings algorithm, where the proposed value is always accepted, is called the Gibbs sampler. The Gibbs sampler was first derived by Geman & Geman (1984), but at that time it failed to achieve any major recognition in the world of statistics. Following the work of Tanner & Wong (1987) and Gelfand & Smith (1990) this iterative procedure of sampling from the posterior distribution became one the most well researched topics in Bayesian statistics. Gammerman & Lopes (2006) describe the following algorithm defining the Gibbs sampler:

1. Initialise the iteration counter to $j = 1$ and set an arbitrary initial values $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_d^{(0)}\right)^T$.
2. Obtain a new value $\boldsymbol{\theta}^{(j)} = \left(\theta_1^{(j)}, \ldots, \theta_d^{(j)}\right)^T$ from $\boldsymbol{\theta}^{(j-1)}$ through successive generation

of values

$$\begin{aligned}
\theta_1^{(j)} &\sim \pi\left(\theta_1|\theta_2^{(j-1)},\ldots,\theta_d^{(j-1)}\right),\\
\theta_2^{(j)} &\sim \pi\left(\theta_2|\theta_1^{(j)},\theta_3^{(j-1)},\ldots,\theta_d^{(j-1)}\right),\\
&\vdots\\
\theta_d^{(j)} &\sim \pi\left(\theta_d|\theta_1^{(j)},\ldots,\theta_{(d-1)}^{(j)}\right).
\end{aligned}$$

3. Change the iteration counter from j to $j+1$ and return to step 2 until convergence is reached.

3.3.1 Markov chain Monte Carlo methods in practice

Suppose a Markov chain with a stationary distribution $\pi(\cdot|\boldsymbol{x})$ equal to the posterior distribution has been constructed using MCMC methods. Ideally, each random variable $\boldsymbol{\theta}^{(j)}$ should have the desired, target posterior distribution $\pi(\cdot|\boldsymbol{x})$. However, there is no guarantee of this, as direct simulation from $\pi(\cdot|\boldsymbol{x})$ is theoretically impossible. For large enough j and independent of the starting distribution $\pi^{(0)}$, it is possible to assume that the chain will converge towards the target distribution. The best scenario would be to make the number of iterations of the chain approach infinity, but this is not attainable in practice. Instead, the simulated chain is run for a certain number of iterations, say N. These N iterations are then discarded in a process known as *burn–in* and are not included in the further inference. The subsequent iterations can be regarded as (dependent) samples from $\pi(\cdot|\boldsymbol{x})$. Various diagnostic methods can be performed to further ensure that the chain has reached equilibrium. The study of convergence can be split into two main approaches: theoretical and empirical. Many authors have proposed different solutions to this problem, see for example Meyn & Tweedie (1994). Using theoretical results is often cumbersome. There is no proof of geometric convergence for the generic sampling algorithms, such as the Metropolis–Hastings algorithm for an arbitrary target, which makes theoretical analysis for such cases virtually impossible. Alternatively, however, convergence of the chain can be studied based on the empirical data. The analysis of the chain output is performed to fully assess whether the target posterior distribution can be approximated by this particular chain. Three such methods are used throughout this report.

1. Time series (trace) plots: by inspection of the trace plots, it is possible to determine whether the chain has not reached stationarity and to check whether it explores the parameter space efficiently. This is done by the 'thick pen test' (Gelfand et al. 1990). If the lines in the trace are further apart than the width of a thick pen, then it is impossible to make an assumption of stationarity. Trace plots can also be a useful tool in determining the size of a burn–in period. An increasing or decreasing trend in trace plots would indicate that the burn–in is not over.

2. Thinning: since draws from Markov chains are not independent, it is sensible to look at their autocorrelation plots (Markov chains can often be approximated by the AR(1) process). If the autocorrelation is high, thinning, in which only every k-th sample from a chain is kept and the rest is disregarded, may be applied. The samples obtained this way are less autocorrelated than for the full chain. For chains that mix well, autocorrelations are expected to decline rapidly beyond lag 0. Thinning also reduces the computational overheads of a long time series.

3. Multiple chains: a number of parallel chains starting from different initial values is run and then the output from each is then compared. While an individual chain may show no particular evidence for lack of convergence, the time series plots for a particular variable might show that different chains have not converged to the same marginal distribution for that variable.

3.4 Implementation of MCMC in rjags

There exists vast amounts of legacy code and additional packages in **R**, such as the package **BradleyTerry2** (Turner & Firth 2012, Firth 2005), which help to specify and fit Bradley–Terry type models. However, these packages are deemed ineffective from a Bayesian point of view and original code has to be adapted in order to analyse paired comparison data. An extension to **R**, called rjags, is used throughout this report.

3.4.1 Overview of rjags

The **R** package rjags provides an **R** interface to the JAGS software (Plummer 2013). The name of this software is an acronym of Just Another Gibbs Sampler, although it has the ability to deal with far more complex scenarios than just those which use a Gibbs sampler. JAGS uses MCMC methods to sample from a posterior distribution. This happens in five steps:

1. Definition of the model
2. Compilation
3. Intialisation
4. Adaptation and burn–in
5. Monitoring

This information is then fed into **R** and various convergence diagnostics and model performance measures can be performed.

3.4.2 An illustrative example

To demonstrate how rjags works, the following example is considered. Suppose a tournament has been simulated, with $n = 500$ games played between $K = 4$ teams. Let Y_i be the result of the ith game, which is played between teams t_{i1} and t_{i2}, $t_{i1}, t_{i2} \in \{1, 2, 3, 4\}$ and let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (1, 2, 3, 4)$ be the vector of skill parameters for the 4 teams. The following Bradley–Terry model is assumed

$$Y_i|\boldsymbol{\lambda}, \boldsymbol{t} \sim \text{Bern}(p_i), \quad \text{independently for } i = 1, \ldots, 500$$

where

$$p_i = \frac{\lambda_{t_{i1}}}{\lambda_{t_{i1}} + \lambda_{t_{i2}}}.$$

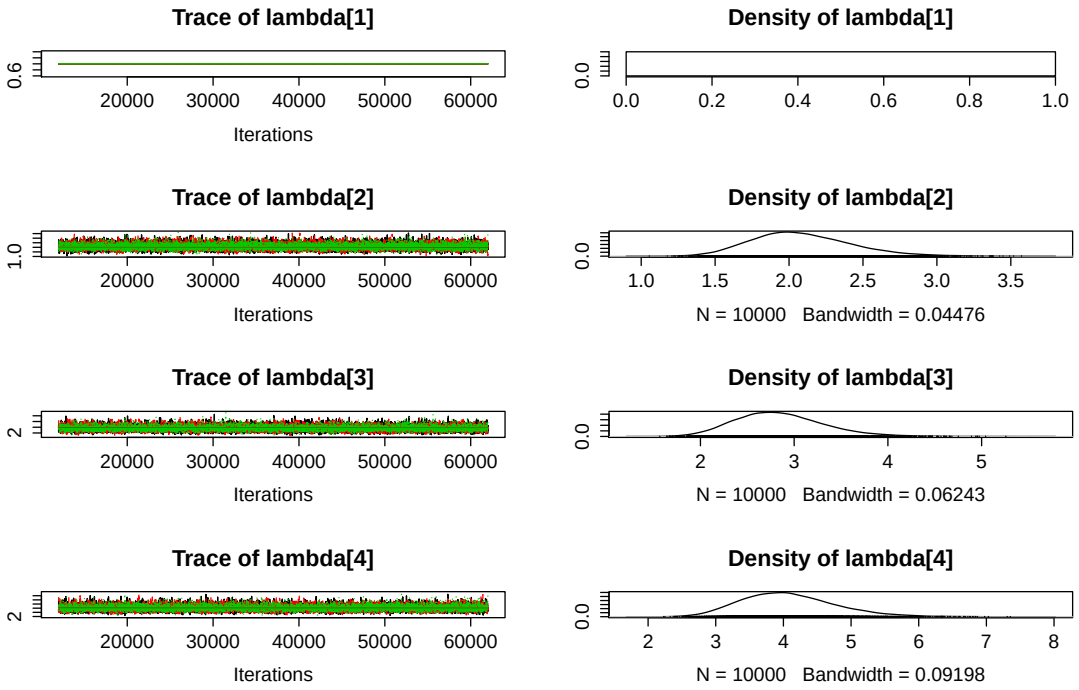
Note that inference on the parameter λ_1 need not be performed since all 4 parameters are not identifiable; see the discussion in Section 2.1.1. To overcome this lack of identifiability the constraint that $\lambda_1 = 1$ is imposed. The following prior distribution is assigned

$$\lambda_j \sim \text{Gamma}(1, 1) \equiv \text{Exp}(1), \quad \text{independently for } j \in \{2, 3, 4\}.$$

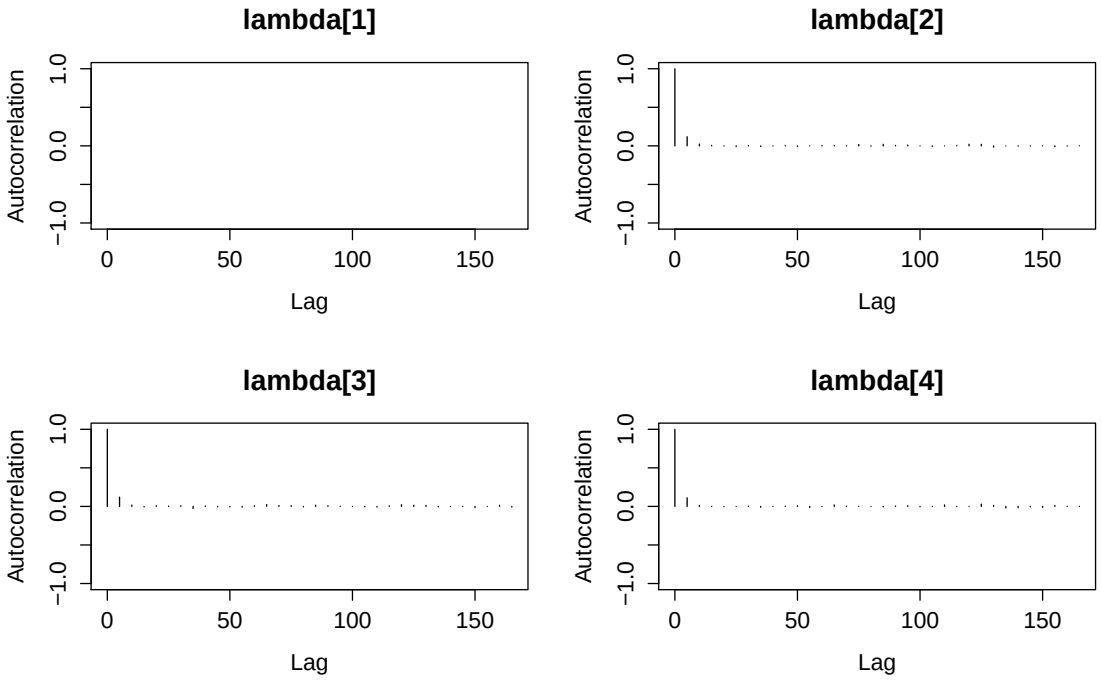
This prior distribution aims to provide a level playing field in which no particular team is considered to be superior to the others.

The aim of this analysis is to make inferences on the skill parameter vector $\boldsymbol{\lambda}$. According to Bayes' theorem, it is possible to combine the data with the prior distribution in order to update the beliefs about these parameters. This is done in rjags by sampling from the posterior distribution of the Bradley–Terry model parameters $\pi(\boldsymbol{\lambda}|D)$, where $D = \{\boldsymbol{Y}, \boldsymbol{t}\}$ denotes the observed results. Three parallel MCMC chains with different initial conditions were run. Each chain ran for 1,000 iterations as burn-in and then a further 50,000 iterations with a thin of 5. Convergence can be checked by looking at the trace and autocorrelation plots in Figure 3.1. From Figure 3.1, it is clear that these three chains explore the parameter space efficiently. They are virtually indistinguishable from each other and by applying the 'thick pen test', stationarity can be safely assumed. No significant autocorrelation beyond lag 0 (which, by definition always equals to one) can be seen. Marginal posterior densities are fairly symmetric and roughly centered around the true values for each skill parameter. From the rjags output, it is also possible to determine values for posterior means of the parameters (with standard errors to indicate their accuracy) and posterior standard deviations. These values can be found in Table 3.1. The posterior means are very close to the true values indicating that the data have been reasonably informative about the parameters.

The Bradley–Terry model can also be used to work out probabilities of each team winning (or losing), with $p_i = \lambda_{t_{i1}}/(\lambda_{t_{i1}} + \lambda_{t_{i2}})$, the probability that team t_{i1} wins game i. It is then possible to plot histograms of the posterior samples of these probabilities, which can be found in Figure 3.2. Comparison between the actual probabilities, calculated from the true values of the skill parameters and those obtained from the Bradley–Terry model can be made. The histograms are roughly centered around the true probabilities, indicated by



(a) Trace and marginal density plots



(b) Autocorrelation plots

Figure 3.1: Convergence diagnostic plots

Skill parameter	Posterior Mean	S.d	Naive SE	Time-series SE
λ_1	1.0000	0.0000	0.0000	0.0000
λ_2	2.0860	0.3348	0.0019	0.0022
λ_3	2.8440	0.4699	0.0027	0.0030
λ_4	4.0760	0.6877	0.0040	0.0044

Table 3.1: Posterior means for the skill parameters

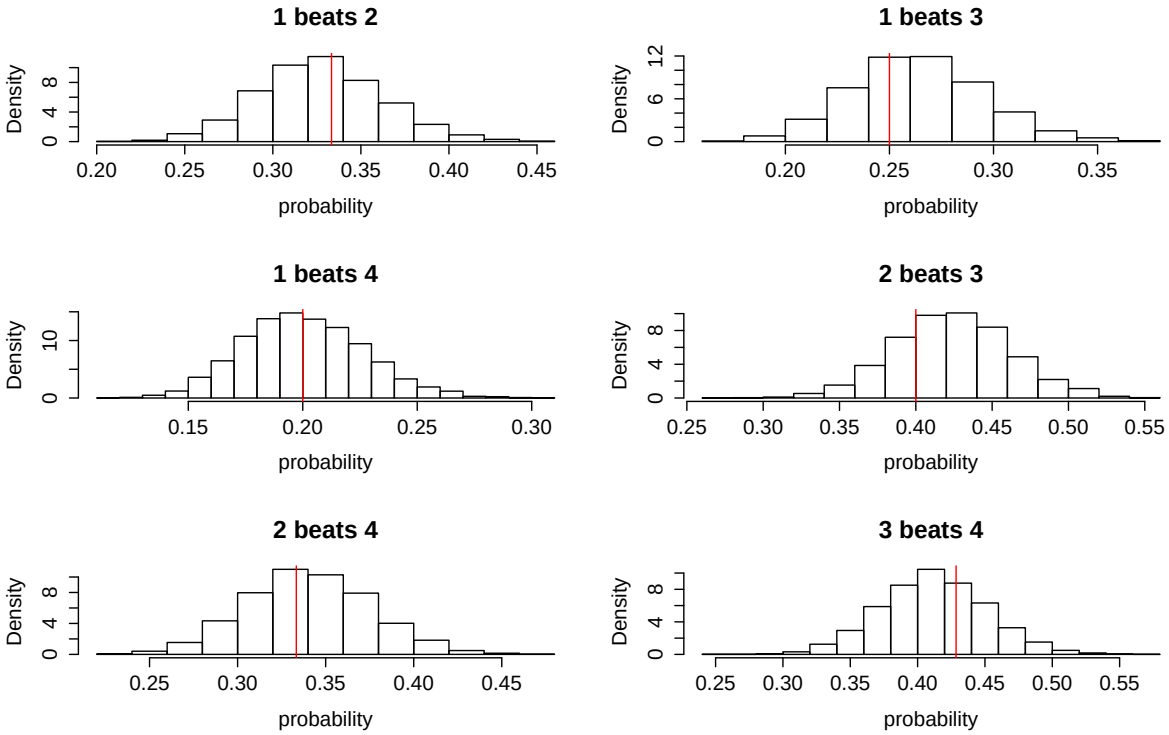


Figure 3.2: Histogram of probabilities

the red vertical lines in Figure 3.2. This suggests that this paired comparison model can be a useful tool for any potential forecasts. This analysis also demonstrates the usefulness of rjags as a tool for performing MCMC for Bayesian inference.

3.5 A note on a semi–conjugate analysis

For the example above, it is also possible to perform Bayesian inference using the Gibbs sampler of Caron & Doucet (2012). The likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\lambda}) &= \prod_{1\leq i\neq j\leq K} \left(\frac{\lambda_i}{\lambda_i+\lambda_j}\right)^{w_{ij}} \\ &= \prod_{i=1}^{K} \lambda_i^{w_i} \sum_{1\leq i\leq j\leq K} (\lambda_i+\lambda_j)^{-n_{ij}}, \end{aligned}$$

where w_{ij} is the number of comparisons in which i beats j, $n_{ij} = w_{ij} + w_{ji}$ is the total number of comparisons between i and j, w_i is the total number of wins for team i and K is the number of teams. By letting $\lambda_k \sim \text{Gamma}(a_k, b)$, (independent for $k = 1, 2 \ldots, K$) be the prior distribution, it is possible to work out the posterior distribution using Bayes' Theorem

$$\begin{aligned} \pi(\boldsymbol{\lambda}|\boldsymbol{x}) &\propto \pi(\boldsymbol{\lambda})L(\boldsymbol{\lambda}) \\ &= \prod_{k=1}^{K} \frac{b^{a_k}\lambda_k^{a_k}-1}{\Gamma(a_k)} \prod_{k=1}^{K} \lambda_k^{w_k} \prod_{1\leq i<j\leq K} (\lambda_i+\lambda_j)^{-n_{ij}}. \end{aligned}$$

Unfortunately, this posterior is not conjugate to its prior, thus making the analysis very difficult. It is possible to make the inference semi–conjugate by introducing latent variables $z_{ij} \sim \text{Gamma}(n_{ij}, \lambda_i + \lambda_j)$ for $1 \leq i < j \leq K$ and $n_{ij} > 0$. A Gibbs sampler can be constructed for sampling from the joint density of parameters, latent variables and data as described in Caron & Doucet (2012). In this report, however, this approach will not be further pursued as rjags provides a more efficient way of obtaining samples from the posterior distribution for the more complex models that will be introduced in the next chapter.

Chapter 4

Analysis of US college basketball results

4.1 "March Madness" NCAA basketball tournament

The annual "March Madness" basketball tournament, played at the end of March each year, is one of the most popular sporting events in the United States. It was founded in 1939 by the National Collegiate Athletic Association (NCAA) and is designated to determine the best men's collegiate basketball team in the country. The format of this competition has changed since its inception. At present, it is the culmination of a regular season, where 354 colleges are split into 31 conferences. During the very short period of time of 21 days in March/April, 67 games are played involving 68 teams, including 32 Divisional I conference champions; this hectic schedule is where the tournament gets its name "March Madness". The remaining 36 teams are chosen by the "Selection Committee" (a body of 10 NCAA athletic directors and conference commissioners) at the end of each regular season. These 68 teams are then split into 4 regions depending on their geographical location. In 2015, these regions were:

1. Midwest (Host: Cleveland State University, Cleveland, Ohio)
2. West (Host: Pepperdine University, Los Angeles, California)
3. East (Host: Syracuse University, Syracuse, New York)
4. South (Host: Rice University and University of Houston, Houston, Texas)

Teams are then seeded from 1 to 68, with the "best" team receiving a number 1 rank. Currently, there are 7 rounds to the tournament:

1. The First Four, where the four lowest seeded teams chosen by the Selection Committee battle it out against the four lowest seeded conference champions.

2. Round of 64, where the highest seeded team in the region plays against the lowest seeded team in that region, that is #1 challenges #16, #2 contests #15, and so on.
3. Round of 32, where the remaining 32 teams are reduced to 16.
4. Regional semifinals, otherwise known as "Sweet Sixteen" round.
5. Regional finals, otherwise known as "Elite Eight", where regional champions are determined.
6. National semifinals, also called Final Four containing four regional champions.
7. National Championship, where two winners of the National semifinals battle it out to find out the winner of the competition and hence the best US college team.

The tournament is single–elimination, where the loser of a match is immediately eliminated from the competition. Although slightly more complicated at first, the format of the "March Madness" competition resembles closely that of the Wimbledon Championship in tennis. The so-called "bracket" in Figure 4.1[1] helps to visualise the structure of the tournament and contains all match results from 2014. The main aim of this project is to apply some of the paired comparison models described previously to form predictions for the 2015 "March Madness" competition, such as the probability that the Duke Blue Devils will beat the Kentucky Wildcats, and thus in turn to "fill out the bracket".

4.2 Kaggle competition and dataset

Kaggle[2] is an online platform for predictive modelling. Individuals, companies and research institutions can post their data, which is then accessible for download from Kaggle's website. Numerous competitions are advertised, some with monetary prizes awarded for the best model. One of them, the *March Machine Learning Mania 2015*[3] competition, sponsored by Hewlett-Packard and with a $15,000 jackpot is intended to find the best prediction of the 2015 "March Madness" tournament. Throughout the remainder of the report the data and structure of this Kaggle competition will be used to guide our predictions for the 2015 tournament.

The dataset available from Kaggle contains a large amount of historical results. It includes the game by game results for the last 31 regular seasons, from 1984 to 2015. A total of 139,920 matches were played during this time. A wide variety of additional information, such as the number of points scored by each team, assists, rebounds and interceptions can also be accessed. The Kaggle competition itself was split into two stages. A non-compulsory first stage offered an incentive to build and test statistical models against the

[1] `http://i.turner.ncaa.com/dr/ncaa/ncaa7/release/sites/default/files/external/gametool/brackets/basketball-men_d1_2013.pdf`

[2] `https://www.kaggle.com/`

[3] `https://www.kaggle.com/c/march-machine-learning-mania-2015`

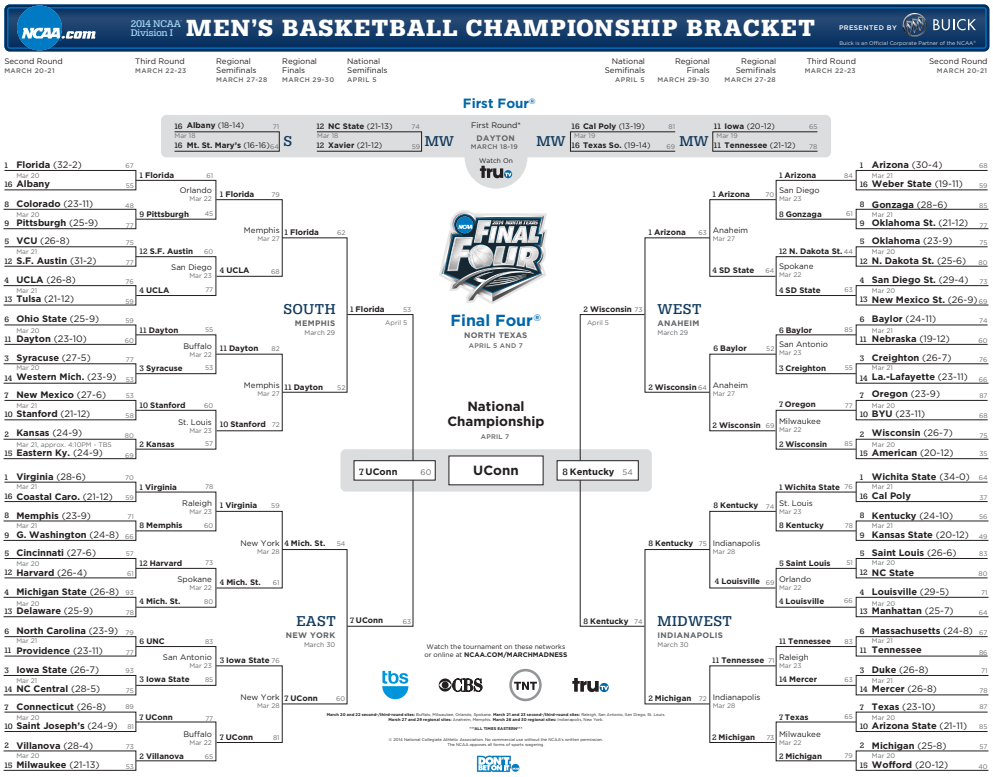


Figure 4.1: NCAA "March Madness" 2014 Results

legacy data available. During the main stage, 2015 "March Madness" forecasts were made immediately prior to the start of the tournament on March 19th 2015, based on the data available up to that point. This was done by predicting the probability that one team would beat another team for all $68 \times 67 = 2287$ possible match–ups of the 68 teams in the tournament.

4.2.1 Assessing the predictions

The criteria used by Kaggle to assess the quality of predictions is the log loss, also known as the predictive binomial deviance. The log loss can be expressed in the following way

$$\log \text{ loss} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)],$$

where n is the number of games played in the prediction sample, $\hat{p}_i$ is the predicted probability of team t_{i1} beating team t_{i2} and y_i is the outcome of each game, taking a value of 1 when team t_{i1} wins and 0 otherwise. The smaller the log loss, the better. Models with good fits and therefore accurate predictions are expected to achieve log loss values near zero. Another criteria that can be used to assess predictions is computing the proportion of games correctly predicted using a model which compares the results of a simulated tournament to the actual results. Note that the probability of correctly

predicting all the tournament games is even smaller than the chances of winning the National Lottery jackpot. The odds of filling out the perfect bracket are 2^{67} to one (67 games, either win or loss, so 2^{67} possible outcomes), thus making it virtually impossible.

Adopting a Bayesian approach to prediction, a prediction of the outcome of a match between two teams i and j is provided by the posterior predictive probability

$$Pr(i \text{ beats } j|D) = \int Pr(i \text{ beats } j|\boldsymbol{\lambda}, D)\pi(\boldsymbol{\lambda}|D)d\boldsymbol{\lambda} = E_{\boldsymbol{\lambda}|D}\left[\frac{\lambda_i}{\lambda_i + \lambda_j}\right],$$

where $D = \{\boldsymbol{Y}, \boldsymbol{t}\}$ denotes the observed results. This probability can be estimated from the MCMC output by

$$\hat{Pr}(i \text{ beats } j|D) = \frac{1}{M}\sum_{m=1}^{M}\frac{\lambda_i^{(m)}}{\lambda_i^{(m)} + \lambda_j^{(m)}},$$

where $\lambda_i^{(m)}, \lambda_j^{(m)}$ for $m = 1, \ldots, M$ are sampled values from the posterior distribution $\pi(\boldsymbol{\lambda}|D)$.

4.3 Fitting the models

The various models will be first tested against the regular season data before actually forecasting the 2015 "March Madness" tournament. It became evident that it was impractical to compile rjags models using the entire dataset of 139,920 games due to the prolonged period of time required to do this. Some of the more sophisticated models would take days to run thus making them computationally expensive and inefficient. Only a subset of this vast dataset was used, namely the match results of the 2015 regular season. There are 354 teams that played 5354 matches in total that took place during 120 "match days". The number of games varies every match day and ranges from 4 to 151, with an average of 45 games. Depending on the size of the conference, each team plays either 14, 16, 18 or 20 matches, half of them in front of their own supporters and the other half away from home (with the exception of New Jersey Institute of Technology, which had to play an unusually high number of 33 matches, due to not being affiliated with any of the conferences). Let Y_i be result of a contest between teams t_{i1} and t_{i2}, $t_{i1}, t_{i2} \in \{1, \ldots, 354\}$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$ be the skill parameter corresponding to each team and $n = 5354$ be the number of matches played. The random variable Y_i is a binary outcome, and $Y_i = 1$ corresponds to team t_{i1} winning a match and $Y_i = 0$ to team t_{i1} losing. This dataset will be analysed sequentially and in chronological order. For match day 1, the predictions for all the games played on that day will be calculated using the prior distribution. For match day 2, the forecasts will be made using the data available for match day 1 and the desired paired comparison model. For match day 3, this will be done by incorporating the data from match days 1 and 2 and so on, so that predictions for matches on day t are based on the data up to day $t - 1$. At each step, the log loss function will be calculated

for each candidate model to see how it evolves through time and the model with the smallest overall cumulative log loss will be chosen and then used to predict the results of the "March Madness" tournament. In the results presented shortly, all the MCMC chains ran for 1,000 iterations as burn–in and then used a further 10,000 iterations to obtain the posterior sample. No thinning was required. The constraint $\lambda_1 = 1$ was also imposed in order to solve the identifiability issue; this corresponds to the strength parameter for Abeline Christian University.

4.3.1 Basic models

The basic Bradley–Terry model as defined in Equation (2.1) and the Thurstone model described in Equation (2.3) were fitted to the data; JAGS code for all the models considered is listed in the Appendix. As expected, the results were very similar and it confirms the hypothesis that these two models are virtually indistinguishable. In both cases, the prior distribution $\lambda_j \sim \text{Gamma}(1,1) \equiv \text{Exp}(1)$ independently for $j \in \{2, 3, \dots 354\}$ was fitted, assuming a level playing field. Because of the striking resemblance between the two models, it seems reasonable to focus on only one of them and then try to modify it.

4.3.2 Home advantage

The phenomenon of home advantage was discussed in Section 2.1.3 of this report and it will be applied to the Bradley–Terry model. A global version of the home advantage will be considered, where there is no assumption of this factor playing a more important role for some teams than for others. The following model was then considered

$$Y_i|\boldsymbol{\theta}, \boldsymbol{t}, \delta \sim \text{Bern}(p_i), \quad \text{independently for } i = \{1, 2, \dots, 5354\}$$

where

$$\text{logit } p_i = \delta h_i + \theta_{t_{i1}} - \theta_{t_{i2}}$$

and

$$\theta_{t_{i1}} = \log \lambda_{t_{i1}} \quad \text{for } i = \{1, \dots, 354\}.$$

The following prior distributions were assigned

$$\begin{aligned} \lambda_j &\sim \text{Gamma}(1,1) \equiv \text{Exp}(1) \quad \text{independently for } j = \{2, \dots, 354\} \\ \delta &\sim N(0,1). \end{aligned}$$

The additional parameter δ represents the home effect and h_i is a binary indicator ($h_i = 1$ if team t_{i1} is at home and $h_i = 0$ otherwise). The choice of $\delta \sim N(0,1)$ reflects a belief that home advantage is just as likely as home disadvantage.

4.3.3 Recent form

Whilst it is reasonably straightforward to extend the Bradley–Terry model to include home advantage, it is much harder to factor in recent form. A sudden change of fortune, alteration in personnel, injuries to key players and so forth, can all shift the momentum, no matter how much silverware any given team has won in the past. In the literature, many authors have proposed their own solutions to circumvent this issue. First proposed by Glickman (1993), the most common way is to ensure that skill parameters vary in time. This subject has been pursued further by other researchers, for example Knorr-Held (2000) applied a dynamic paired comparison model to the 1996–1997 German football league data and more recently Cattelan et al. (2013) applied a slightly different model to the 2008–2009 Italian Serie A and 2009–2010 NBA league. In this report two methods of accounting for recent form are considered. Firstly an indicator parameter $\beta \in \mathbb{R}$ of whether the last game was won is added to the model. A normal distribution with mean 0 and unit precision is assigned as the prior to this parameter. Secondly, and perhaps more importantly, more weight is given to recent results. This can be done by introducing another parameter $\alpha \in [0, 1]$, which then can be either fixed at a particular value or assigned a Beta prior distribution. This in turn gives rise to three possible candidate models. For example, the Bradley–Terry model with home advantage, and parameters α and β as described above can be specified as

$$Y_i|\boldsymbol{\theta}, \boldsymbol{t}, \delta, \alpha, \beta \sim \text{Bern}(p_i), \quad \text{independently for } i = \{1, 2, \ldots, 5354\}$$

where

$$\text{logit } p_i = \delta h_i + \beta(I_{t_i} - I_{t_j}) + \alpha^{d_i}(\theta_{t_{i1}} - \theta_{t_{i2}})$$

and d_i is the number of days in the past that the game was played. I_{t_i} is an indicator of whether the last game for team t_{i1} was won ($I_{t_i} = 1$) or lost ($I_{t_i} = -1$). The following prior distributions were assigned

$$\begin{aligned}
\lambda_j &\sim \text{Gamma}(1, 1) \equiv \text{Exp}(1) \quad \text{independently for } j = \{1, \ldots, 354\} \\
\delta &\sim N(0, 1) \\
\alpha &\sim \text{Beta}(99, 1) \\
\beta &\sim N(0, 1).
\end{aligned}$$

Note that the prior mean of α is 0.99 indicating that a match played yesterday is expected to be 99% as important as a match played today. As a quick illustration of how this model works, suppose a match took place today and for the time being assume α is fixed at 0.99. In this setting, $d_i = 0$ and thus $\alpha^{d_i} = 0.99^0 = 1$. Now consider a game played 14 days ago, hence $d_i = 14$ and $\alpha^{d_i} = 0.99^{14} = 0.869$ (3 d.p). Clearly more weight is attached to the recent results. With $\alpha < 1$, $\alpha^{d_i} \to 0$ as $d_i \to \infty$ and thus $\text{logit}(p_i) \to \delta h_i + \beta(I_{t_i} - I_{t_j})$ indicating that effectively the difference in team strengths has no impact on the probability of a win.

4.4 Model choice

During the sequential analysis, the log loss function was calculated for each of the 5354 games using the five models described above. The simple Bradley–Terry model is referred to as model 1. The Bradley–Terry with home advantage as model 2. The Bradley–Terry model with home advantage and time weighting with an assigned prior distribution as model 3. The Bradley–Terry model, where the time weighting is fixed at a particular value ($\alpha = 0.99$) is referred to as model 4. Finally, model 5 is the Bradley–Terry model with home advantage, time weighting and an indicator whether the last game was won or lost.

The results of fitting these five models to the 2015 regular season results are summarised in Table 4.1 and Figure 4.2.

Model	Cumulative log loss	Avg. log loss per game
1	3213.6300	0.6002
2	3076.0190	0.5745
3	3077.6900	0.5748
4	3100.0530	0.5790
5	3080.3480	0.5753

Table 4.1: Log loss for the five candidate models

From Table 4.1 it can be seen that Model 1 has the highest overall cumulative log loss and hence the highest log loss value per game. Therefore, according to the theory presented in the previous section, this model performs the worst. The difference between the other four remaining models is negligible, but the log loss for model 2 is marginally the smallest. This highlights the importance of the home advantage as a factor in this type of analysis. Figure 4.2 contains a time series plot illustrating how the log loss values for each model relative to model 2 change over time. Model 1 does not achieve a satisfactory log loss value and is not recommended for the future forecasts. Models 3, 4 and 5 all provide very similar results approximately up to halfway through the season. After around match day 58, the log loss values for model 4 increase significantly indicating poorer predictions. On the other hand, models 3 and 5 provide similar predictions, with model 3 having slightly smaller log loss values. Several conclusions spring to mind. Firstly, an indicator of whether the last game was won or lost does not appear to provide a significant improvement in the forecasts. Secondly, time weighting appears to have the second highest impact on the forecasts, after the inclusion of home advantage. It is better to assign a prior distribution to this parameter, rather than fixing it. Overall, model 2 provides the most accurate predictions. Model 3 also generates very reasonable results, however, compared to model 2, it is more complex and therefore less computationally efficient. Therefore model 2, the Bradley–Terry model with home advantage, is chosen to produce forecasts for the 2015 March Madness tournament and the related Kaggle contest. This section not only

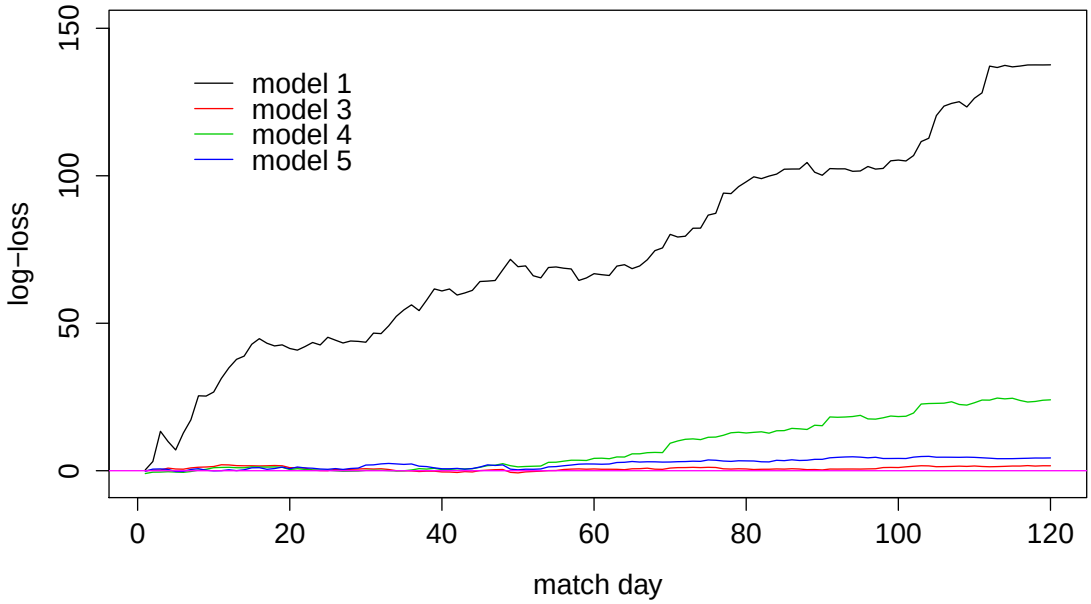


Figure 4.2: Log loss over time for each model, relative to model 2; values above 0 indicate a worse log loss than model 2

provides a means of selecting the optimal model, but also a way of checking the validity and appropriateness of these models. Simulating match predictions using the posterior distribution and comparing it to the actual results using a quantitative measure, such as log loss gives a chance to test the models against any potential major discrepancies. Goodness of fit can also be checked using more standard diagnostic procedures, such as plotting residuals. It is possible to calculate residuals for binary data, transform them in such a way that they are approximately normally distributed, and then analyse these using classical least squares theory. This approach however is not recommended, as it does not yield optimal results in a Bayesian setting. Johnson & Albert (1999) advise to use Bayesian residuals instead. This can be done by examining the probability distribution of the difference between the observed and the fitted observations. Such Bayesian residuals may be defined as

$$r_{i,b} = y_i/n_i - \hat{p}_i$$

where y_i takes a value of 1 if team t_{i1} wins the game or 0 otherwise, n_i is the match number for team t_{i1} (which here is 1) and $\hat{p}_i$ is the model-based probability that team t_{i1} wins. Several other types of residuals have also been proposed, such as posterior–predictive, cross–validation and Bayesian latent residuals. These, however, are not discussed in this report.

4.5 Forecasting the 2015 March Madness results

4.5.1 Tournament picture

The Kentucky Wildcats, runners–up in the 2014 March Madness tournament, entered the 2015 tournament unbeaten and were widely backed to be the favourites after recording one of the longest winning streaks in the history of the tournament (34-0). Last year's winners, the Connecticut Huskies did not qualify after finishing fifth in their conference. Three teams, Buffalo Bulls, UC Irvine Anteaters and North Florida Ospreys qualified to the tournament for the first time in their history. New Mexico State Aggies, Stephen F. Austin Lumberjacks and Georgia State Panthers were lower seeded teams identified by the experts as "Cinderellas", that is, teams with a potential to have a deep play–off run. The eight best teams according to three different measures: Associated Press (AP)[4] Ranking, ESPN Power Index[5] and official NCAA seeding information[6] are presented in the Table 4.2.

AP Ranking	ESPN Power Index	NCAA seeds
Kentucky	Kentucky	Kentucky
Arizona	Wisconsin	Villanova
Wisconsin	Arizona	Duke
Duke	UVA	Wisconsin
Kansas	Villanova	Virginia
North California	Duke	Arizona
Florida	Gonzaga	Gonzaga
Louisville	Kansas	Kansas

Table 4.2: Top 8 best teams according to three different sources

4.5.2 Application of the chosen model for prediction

As mentioned previously, the Bradley–Terry model with home advantage was chosen as the model with which to make predictions. As also mentioned previously, the whole dataset is not used in making the final predictions. This is due to two reasons. Firstly, the algorithm incorporating all 139,920 games would be computationally inefficient and very slow to compile. Secondly, and perhaps most importantly, the teams' strength parameters may have changed substantially over a long time period, for example decades. This would therefore lead to less accurate predictions, even with the time weighting factor included in the model. For instance, the Kentucky Wildcats suffered a severe decline in form during

[4] http://espn.go.com/mens-college-basketball/rankings/_/year/2015/week/1/seasontype/2

[5] http://espn.go.com/mens-college-basketball/bpi

[6] http://www.sbnation.com/college-basketball/2015/3/15/8220261/ncaa-tournament-2015-full-seed-list-dayton

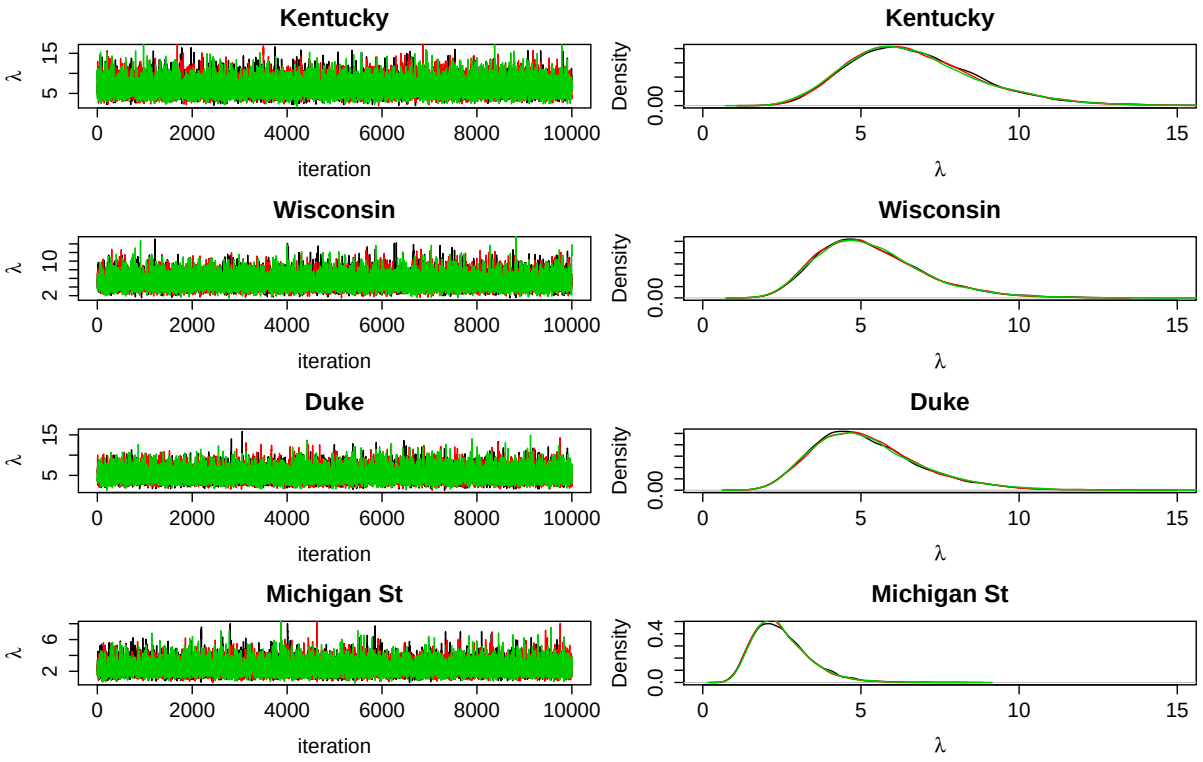
the Final Four drought from 1999–2011. Thus including information from these years would result in a much lower estimated skill parameter than would be anticipated for this team. A compromise needs to be reached, whereby enough data is used to ensure that it improves the forecasts, but not the whole dataset, as this would slow the algorithm too much and impact the predictions in a negative way. There is no quantitative way of deciding this and subjective judgment is the only potential option. By trial and error, it was decided that using the dataset from the 2015 regular season is the optimal size for forecasting the 2015 March Madness tournament.

Each of three MCMC chains ran for 1,000 iterations as burn–in and then took further 10,000 iterations with no thinning. The trace and autocorrelation plots were produced for each of the 68 teams that qualified for the tournament. In this report, however, convergence diagnostic plots are only presented for the 2015 Final Four teams: Kentucky Wildcats, Duke Blue Devils, Wisconsin Badgers and Michigan State Spartans. These findings are summarised in Figure 4.3. The MCMC chains explore the state space efficiently and therefore mix well. The 'thick pen' test indicates that the stationarity can be safely assumed. There is no significant autocorrelations beyond lag 0 and these indeed decrease rapidly, thus showing that that no thinning was required. It can be seen that the marginal posterior density for Kentucky Wildcats' strength parameter is centered around the highest value of $\lambda \approx 6.7$. This indicates that, according to the model fitted, Kentucky Wildcats can be regarded as the strongest team in the Final Four of the competition. Similarly, as the marginal posterior density is centered around the smallest value of λ ($\lambda \approx 2.5$), this implies that Michigan State Spartans can be considered to be the weakest team in the Final Four. As the strength parameter for this particular team is considerably lower than for the remaining three Final Four teams, one might perhaps refer to the Michigan State Spartans as this year's "Cinderella" team, an underdog who exceeded expectations. Some more detailed results are provided in the following section.

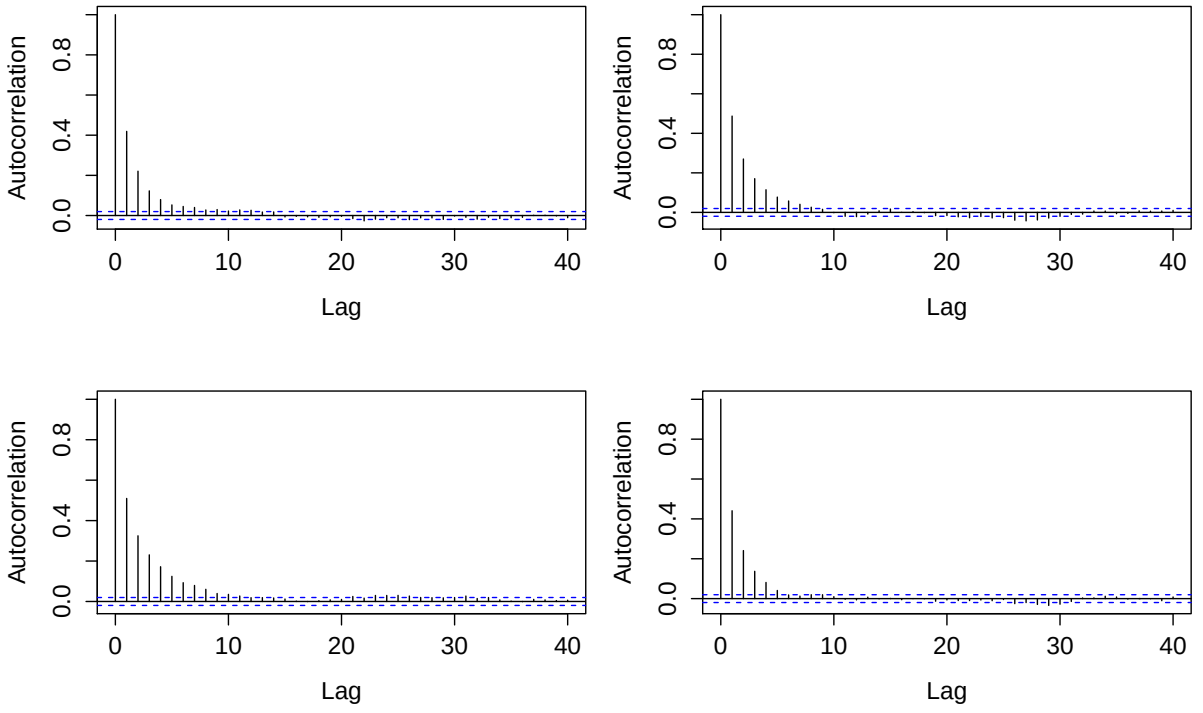
4.5.3 Results

It is possible to rank the teams according to the posterior mean of their strength parameter λ. These findings are summarised in Table 4.3 for the teams reaching the final 64 of the tournament. According to the model fitted, Kentucky is clearly the best team and was expected to win the whole tournament. This agrees with the experts' beliefs presented in Table 4.2. There are no major discrepancies between the top 8 best teams according to the model fitted and those in Table 4.2, with the exception of Northern Iowa featuring as the 8th best team according to the Bradley–Terry model with home advantage. The lowest ranked teams agree with the official seeding information.

Based on the posterior distribution of the strength parameters, predictive probabilities for each of the 63 games in the final 6 rounds of the tournament can be obtained using the method outlined in Section 4.2.1. These predictive probabilities can be compared to the actual results from the 2015 tournament and a log loss can be calculated. The log loss value for this particular model is 0.570 (3 d.p). The proportion of games correctly



(a) Trace and marginal density plots



(b) Autocorrelation plots

Figure 4.3: Convergence diagnostics plots

Ranking	Team	Posterior mean λ	Posterior s.d. λ
1	Kentucky	6.686	2.056
2	Villanova	5.731	1.856
3	Wisconsin	5.376	1.746
4	Virginia	5.271	1.705
5	Duke	5.223	1.651
6	Arizona	5.113	1.656
7	Gonzaga	5.087	1.720
8	Northern Iowa	4.119	1.253
⋮	⋮	⋮	⋮
29	Michigan St	2.472	0.895
⋮	⋮	⋮	⋮
57	Belmont	1.257	0.551
58	Coastal Car.	1.239	0.579
59	New Mexico St.	1.160	0.518
60	Lafayette.	1.150	0.480
61	Texas Southern	1.110	0.505
62	UAB	0.885	0.350
63	Robert Morris	0.820	0.347
64	Hampton	0.402	0.180

Table 4.3: Top 8 highest ranked and bottom 8 lowest ranked teams according to the model fitted

predicted by the Bradley–Terry model with home advantage is approximately 73%. This would indicate that this particular model provides more accurate predictions than just guessing, where one is expected to correctly predict 50% of matches on average. Having said that, forecasting "March Madness" results proved to be a very challenging task, due to an unexpected deep play–off run by the Michigan State Spartans, a team ranked 29th best by the model, as well as poorer than anticipated performance by the Kentucky Wildcats. For example, consider the last three games of the tournament: Kentucky vs Wisconsin, Duke vs Michigan State and the subsequent Championship final Duke vs Wisonsin. Using the model fitted, it is possible to calculate probabilities for each of the match–ups as well as produce posterior probability density plots for these. From Figure 4.4, Kentucky were the favourites to win their semi–final with a probability of approximately 60%. Wisconsin, however, ended Kentucky's unbeaten streak by recording a 71–64 win against the Wildcats. There were no upsets in the second semi–final, where Duke was a clear favourite to beat Michigan State with a probability of approximately 73%. Duke indeed progressed to the National Championship match by defeating Michigan State 81–61. The Bradley–Terry model with home advantage picks Wisconsin as the marginal favourites for the final. However, Duke under the leadership of Mike Krzyzewski took the title by beating Wisconsin 68–63. This illustrates one of many challenges related to sports

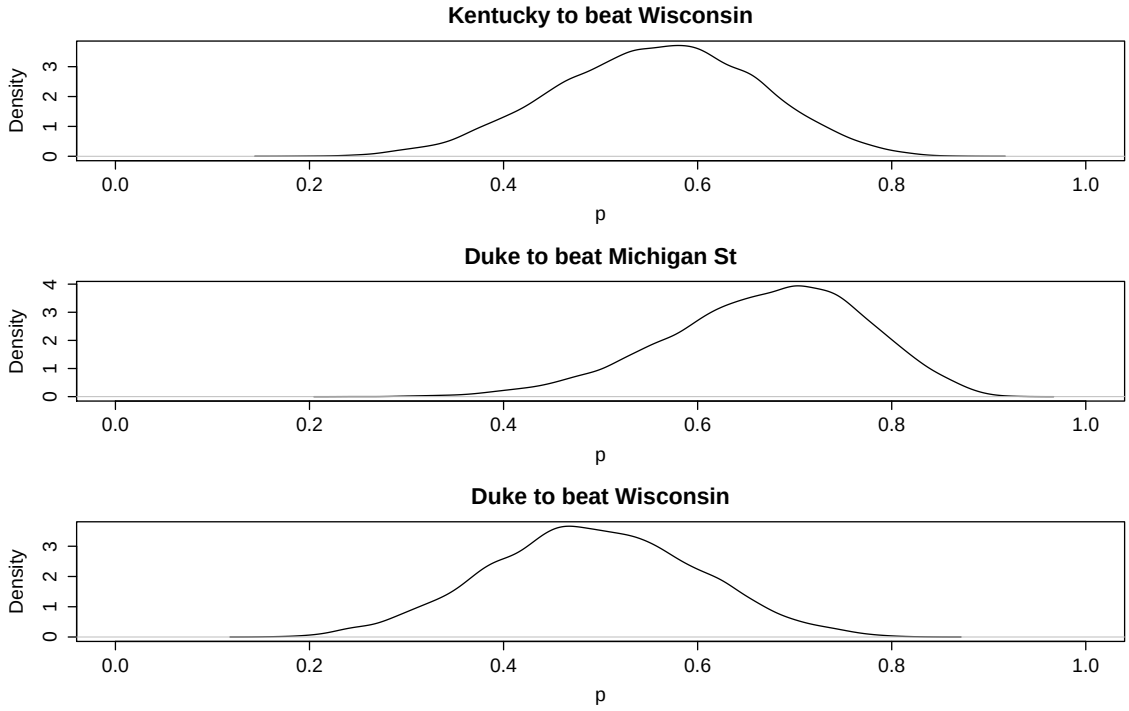


Figure 4.4: Probability densities for the final three games of the tournament

forecasting. It is extremely difficult, if not impossible, to predict all the games correctly. There is no way of accounting for any potential "David beats Goliath" scenarios.

Chapter 5

Conclusions and further discussion

5.1 Conclusions

This project provides an overview of models for paired comparison data. The Bradley–Terry model and its generalisations arise in numerous applications ranging from sport to medicine. A link between paired comparison data models and binary regression models has been explored. The Bradley–Terry and the Thurstone model arise as special cases of the linear paired comparison model or the gamma paired comparison model. A closed form of the latter has been derived. Most of the research on paired comparison data has been performed from the frequentist point of view. Recently, however, several authors proposed to perform Bayesian inference for this type of data. This was also a favoured approach in this report. Fundamental concepts, such as MCMC methods and in particular Metropolis–Hastings and Gibbs sampling algorithms for sampling from a Markov chain whose stationary distribution is the posterior distribution of interest have been discussed. These methods are computationally demanding and cannot be performed without sophisticated statistical software. The **R** package rjags, which provides an **R** interface to the JAGS software was our preferred option. A synthetic dataset was firstly considered to illustrate how inference on paired comparison data can be carried out. Similar methodology was then applied to the NCAA basketball dataset to rank and subsequently determine the best men's collegiate team. It was found that the Bradley–Terry model and the Thurstone model provide virtually the same results when fitted to data. Home advantage proved to be the most significant factor in the analysis. The final model chosen to forecast the 2015 "March Madness" tournament results was the Bradley–Terry model with home advantage. 73% of all games were predicted correctly by this particular model, but the eventual winner of the tournament, Duke, was only ranked as the 5th best team by this model. Such a result is not particularly surprising given the single-elimination format of the March Madness tournament.

5.2 Further discussion

There are many possible extensions of this project. Team–specific home advantage could be considered instead of the global version of this phenomenon. Finding clusters of teams of similar ability and then comparing them in turn is another possibility. Tutz & Schauberger (2014) consider applying this methodology to German Bundesliga data. Rather than comparing two teams at a time, group comparisons could be made. Caron & Doucet (2012) and Huang et al. (2006) discuss this in further detail. As the Bradley–Terry model could be extended to incorporate ties (Rao & Kupper 1967), it is feasible to use this type of a model to analyse sports data with three different outcomes (usually win, loss or a draw), such as football or rugby. Whilst most of the research on paired comparisons in sports data focuses on ranking teams or individuals, other extensions could also arise. For example, uncertainty is rarely, if ever, taken into account in the presentation of sports data. It would be fascinating to test whether the onset of professional eras has had any discernible impact on relevant sports.

Bibliography

Agresti, A. (1990), *Categorical Data Analysis*, Wiley.

Baker, R. D. & McHale, I. G. (2014), 'A dynamic paired comparisons model: Who is the greatest tennis player?', *European Journal of Operational Research* **236**, 677–684.

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

Bäuml, K.-H. (1994), 'Upright versus upside–down faces: How interface attractiveness varies with orientation', *Perception & Psychophysics* **56**, 163–172.

Bradley, R. A. (1953), 'Some Statistical Methods in Taste Testing and Quality Evaluation', *Biometrics* **9**, 22–38.

Bradley, R. A. (1965), 'Another Interpretation of a Model for Paired Comparisons', *Psychometrika* **30**, 315–318.

Bradley, R. & Terry, M. (1952), 'Rank analysis of incomplete block designs. I. The method of paired comparisons', *Biometrika* **39**, 324–345.

Caron, F. & Doucet, A. (2012), 'Efficient Bayesian inference for generalized Bradley-Terry models', *Journal of Computational and Graphical Statistics* **21**, 174–196.

Cattelan, M., Varin, C. & Firth, D. (2013), 'Dynamic Bradley–Terry modelling of sports tournaments', *Journal of Royal Statistical Society Series C (Applied Statistics)* **62**, 135–150.

Chambers, E. A. & Cox, D. (1967), 'Discrimination between alternative binary response models', *Biometrika* **54**, 573–578.

David, H. A. (1988), *The Method of Paired Comparisons*, second edn, Griffin, London.

Davidson, R. R. & Solomon, D. L. (1973), 'A Bayesian approach to paired comparison experimentation', *Biometrika* **60**, 477–487.

Duinevald, C. A. A., Arents, P. & King, B. M. (2000), 'Log–linear modelling of paired comparison data from consumer tests', *Food Quality and Preference* **11**, 63–70.

Elo, A. E. (1978), *The rating of chessplayers, past and present*, Vol. 3, Batsford London.

Firth, D. (2005), 'Bradley-terry models in r', *Journal of Statistical Software* **12**(1), 1–12. **URL:** *http://www.jstatsoft.org/v12/i01*

Ford, L. (1957), 'Solution of a Ranking Problem from Binary Comparisons', *The American Mathematical Monthly* **64**, 28–33.

Gammerman, D. & Lopes, H. (2006), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Texts in Statistical Science, Taylor & Francis, London.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. & Smith, A. F. M. (1990), 'Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling', *Journal of the American Statistical Association* **85**, 972–985.

Gelfand, A. E. & Smith, A. F. (1990), 'Sampling–Based Approaches to Calculating Marginal Densities', *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004), *Bayesian Data Analysis*, second edn, Chapman and Hall/CRC, Boca Raton, Florida.

Geman, S. & Geman, D. (1984), 'Stochastic Relaxation, Gibbs Distibutions, and the Bayesian Restoration of Images', *IEEE Transactions On Pattern Analysis And Machine Intelligence* **PAMI-6**, 721–741.

Glickman, M. E. (1993), 'Paired comparison models with time–varying parameters. phd thesis', *Department of Statistics, Harvard University, Cambridge* .

Glickman, M. E. (1999), 'Parameter estimation in large dynamic paired comparison experiments', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **48**, 377–394.

Greer, D. L. (1983), 'Spectator Booing and the Home Advantage: A Study of Social Influence in the Basketball Arena', *Social Psychology Quaterly* **46**, 252–261.

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**, 97–109.

Henery, R. J. (1992), 'An extension to the Thurstone–Mosteller model for chess', *The Statistician* **41**, 559–567.

Huang, T.-K., Weng, R. C. & Lin, C.-J. (2006), 'Generalized Bradley–Terry Models and Multi–Class Probability Estimates', *The Journal of Machine Learning Research* **7**, 85–115.

Hunter, D. (2004), 'MM algorithms for generalized Bradley–Terry models', *The Annals of Statistics* **32**, 384–406.

Joe, H. (1990), 'Extended Use of Paired Comparison Models, with Application to Chess Rankings', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **39**, 85–93.

Johnson, V. E. & Albert, J. H. (1999), *Ordinal Data Modelling*, Statistics for Social Science and Public Policy, Springer–Verlag, New York.

Jones, M. B. (2014), 'The home disadvantage in championship competitions: team sports', *Psychology of Sport and Exercise* **15**, 392–398.

Kind, P. (1982), 'A Comparison of Two Models for Scaling Health Indicators', *International Journal of Epidemiology* **11**, 271–275.

Kissler, J. & Bäuml, K.-H. (2000), 'Effects of the beholder's age on the perception of facial attractiveness', *Acta Psychologica* **104**, 145–166.

Knorr-Held, L. (2000), 'Dynamic rating of sports teams', *The Statistician* **49**, 261–276.

Krabbe, P. F. M. (2008), 'Thurstone Scaling as a Measurment Method to Quantify Subjective Health Outcomes', *Medical Care* **46**, 357–365.

Kuk, A. Y. C. (1995), 'Modelling Paired Comparison Data with Large Number of Draws and Large Variability of Draw Percenatges Among Players', *Journal of the Royal Statistical Society. Series D (The Statistician)* **44**, 523–528.

Leonard, T. (1977), 'An Alternative Bayesian Approach to the Bradley–Terry Model for Paired Comparisons', *Biometricks* **33**, 121–132.

Matthews, J. N. S. & Morris, K. P. (1995), 'An Application of Bradley–Terry–type Models to the Measurement of Pain', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **44**, 243–255.

Maydeu-Olivares, A. & Böckenholt, U. (2008), 'Modelling Subjective Health Outcomes Top 10 Reasons to Use Thurstone's Method', *Medical Care* **46**, 346–348.

McHale, I. & Morton, A. (2011), 'A Bradley–Terry type model for forecasting tennis match results', *International Journal of Forecasting* **27**, 619–630.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. & Teller, A. H. (1953), 'Equation of State Calculations by Fast Computing Machines', *The Journal of Chemical Physics* **21**, 1087–1092.

Meyn, S. P. & Tweedie, R. L. (1994), 'Computable Bounds For Geometric Convergence Rates Of Markov Chains', *The Annals of Applied Probability* **4**, 981–1011.

Mosteller, F. (1951), 'Remarks on the method of paired comparisons: I, The least squares solution assuming equal deviations and equal correlations. II, The effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed.

III, A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed', *Psychometrika* **16**, 3–9, 203–206, 207–218.

Nelder, J. A. & Wedderburn, R. W. M. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society. Series A (General)* **135**, 370–384.

Plummer, M. (2013), 'rjags: Bayesian graphical models using mcmc. r package version 3–10'.

Rao, P. V. & Kupper, L. L. (1967), 'Ties in paired-comparison experiments: a generalization of the Bradley-Terry model', *Journal of the American Statistical Association* **62**, 194–204.

Snyder, E. E. & Purdy, D. A. (1985), 'The Home Advantage in Collegiate Basketball', *Sociology of Sport Journal* **2**, 352–356.

Stern, H. S. (1990), 'A continuum of paired comparisons models', *Biometrika* **77**, 265–273.

Stuart-Fox, D. M., Firth, D., Moussalli, A. & Whiting, M. J. (2006), 'Multiple signals in chameleon contests: designing and analysing animal contests as a tournament', *Animal Behaviour* **71**, 1263–1271.

Tanner, M. A. & Wong, W. H. (1987), 'The Calculation of Posterior Distributions by Data Augmentation', *Journal of the American Statistical Association* **82**, 528–540.

Thurstone, L. L. (1927), 'A law of comparative judgement', *Psychological Review* **79**, 281–299.

Turner, H. & Firth, D. (2012), 'Bradley-Terry models in r: The bradleyterry2 package', *Journal of Statistical Software* **48**(9), 1–21.
URL: *http://www.jstatsoft.org/v48/i09*

Tutz, G. & Schauberger, G. (2014), 'Extended ordered paired comparison models with applications to football data from German Bundesliga', *Springer-Verlag* **27**, 619–630.

Yao, G. & Böckenholt, U. (1999), 'Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler', *British Journal of Mathematical and Statistical Psychology* **52**, 79–92.

Zermelo, E. (1929), 'Die Berechnung der Turnier–Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung', *Mathematische Zeitschrift* **29**, 436–460.

Appendix: **JAGS** model code

```
## Various models for paired comparisons in jags format

## Bradley-Terry model
bt = "
  model
  {
      for(i in 1:n)
      {
        y[i] ~ dbern(p[i])
        p[i] <- lambda[x[i,1]]/(lambda[x[i,1]] + lambda[x[i,2]])
      }
      for(j in 1:K)
      {
        lambda[j]~dgamma(1,1)
      }
  }
"

## Bradley-Terry model (logistic regression)
bt.logit = "
  model
  {
      for(i in 1:n)
      {
        y[i] ~ dbern(p[i])
        logit(p[i]) <- theta[x[i,1]]-theta[x[i,2]]
      }
      lambda[1] <- 1
      theta[1] <- log(lambda[1])
      for(j in 2:K)
      {
        lambda[j]~dgamma(1,1)
        theta[j] <- log(lambda[j])
      }
  }
"
```

```
## Bradley-Terry model with home advantage (logistic regression)
bt.logit.ha = "
  model
  {
      for(i in 1:n)
      {
        y[i] ~ dbern(p[i])
        logit(p[i]) <- eta*h[i] + theta[x[i,1]]-theta[x[i,2]]
      }
      lambda[1] <- 1
      theta[1] <- log(lambda[1])
      for(j in 2:K)
      {
        lambda[j]~dgamma(1,1)
        theta[j] <- log(lambda[j])
      }
      eta ~ dnorm(0,1)
  }
"


## Bradley-Terry model (logistic regression) with home advantage and
## time-weighting

bt.logit.ha.tw = "
  model
  {
      for(i in 1:n)
      {
        y[i] ~ dbern(p[i])
        logit(p[i]) <- eta*h[i] + (alpha^t[i])*(theta[x[i,1]]-theta[x[i,2]])
      }
      lambda[1] <- 1
      theta[1] <- log(lambda[1])
      for(j in 2:K)
      {
        lambda[j]~dgamma(1,1)
        theta[j] <- log(lambda[j])
      }
      eta ~ dnorm(0,1)
      alpha ~ dbeta(99,1)
  }
"
```

```
## Bradley-Terry model (logistic regression) with home advantage and
## time-weighting (fixed)

bt.logit.ha.twfix = "
  model
  {
      for(i in 1:n)
      {
        y[i] ~ dbern(p[i])
        logit(p[i]) <- eta*h[i] + (alpha^t[i])*(theta[x[i,1]]-theta[x[i,2]])
      }
      lambda[1] <- 1
      theta[1] <- log(lambda[1])
      for(j in 2:K){
        lambda[j]~dgamma(1,1)
        theta[j] <- log(lambda[j])
      }
      eta ~ dnorm(0,1)
  }
"

## Bradley-Terry model (logistic regression) with home advantage,
## time-weighting and an indicator of whether the last game was won

bt.logit.ha.tw.wlg = "
  model
  {
      for(i in 1:n)
      {
        y[i] ~ dbern(p[i])
        logit(p[i]) <- eta*h[i] + (alpha^t[i])*(theta[x[i,1]]-theta[x[i,2]])
                                + beta*(wlg[i,x[i,1]]-wlg[i,x[i,2]])
      }
      lambda[1] <- 1
      theta[1] <- log(lambda[1])
      for(j in 2:K){
        lambda[j]~dgamma(1,1)
        theta[j] <- log(lambda[j])
      }
      eta ~ dnorm(0,1)
      alpha ~ dbeta(99,1)
      beta ~ dnorm(0,1)
  }
"
```

```
## Thurstone-Mosteller model (probit regression)
tm = "
  model
  {
      for(i in 1:n)
      {
        y[i] ~ dbern(p[i])
        probit(p[i]) <- theta[x[i,1]]-theta[x[i,2]]
      }
      for(j in 1:K)
      {
        lambda[j]~dgamma(1,1)
        theta[j] <- log(lambda[j])
      }
  }
"

## Thurstone-Mosteller model with home advantage (probit regression)
tm.ha = "
  model
  {
      for(i in 1:n)
      {
        y[i] ~ dbern(p[i])
        probit(p[i]) <- eta*h[i] + theta[x[i,1]]-theta[x[i,2]]
      }
      for(j in 1:K)
      {
        lambda[j]~dgamma(1,1)
        theta[j] <- log(lambda[j])
      }
      eta ~ dnorm(0,0.1)
  }
"
```