A stochastic rank ordered logit model for rating multi-competitor games and sports

Mark E. Glickman*   Jonathan Hennessy

January 29, 2015

Abstract

Many games and sports, including races, involve outcomes in which competitors are rank ordered. In some sports, competitors may play in multiple events over long periods of time, and it is natural to assume that their abilities change over time. We propose a Bayesian state-space framework for rank ordered logit models to rate competitor abilities over time from the results of multi-competitor games. Our approach assumes competitors' performances follow independent extreme value distributions, with each competitor's ability evolving over time as a Gaussian random walk. The model accounts for the possibility of ties, an occurrence that is not atypical in races in which some of the competitors may not finish and therefore tie for last place. Inference can be performed through Markov chain Monte Carlo (MCMC) simulation from the posterior distribution. We also develop a filtering algorithm that is an approximation to the full Bayesian computations. The approximate Bayesian filter can be used for updating competitor abilities on an ongoing basis. We demonstrate our approach to measuring abilities of 268 women from the results of women's Alpine downhill skiing competitions recorded over the period 2002-2013.

Keywords: Dynamic model, exploded logit, Plackett-Luce, order statistics, ranking

---

*Address for correspondence: Center for Healthcare Organization & Implementation Research, Edith Nourse Rogers Memorial Hospital (152), Bldg 70, 200 Springs Road, Bedford, MA 01730, USA. E-mail address: mg@bu.edu. Phone: (781) 687-2875. Fax: (781) 687-3106. Author affiliations: Glickman – Department of Health Policy and Management, Boston University School of Public Health, and the Center for Healthcare Organization and Implementation Research, a Veteran Administration Center of Innovation; Hennessy – The Houston Rockets. The views expressed in this article are those of the authors and do not necessarily reflect the views of the Department of Veterans Affairs.

1 Introduction

Measuring competitor strength in games and sports has been an area of great interest among professional scouts, sports organizations, and fans. Over the past 20-30 years, most development of statistical methods for assessing competitor strength has been in the context of head-to-head competitions that rely on paired comparison methods. Modern treatment of paired comparisons assume fundamentally that each competitor's strength can be represented as a parameter in a probability model for the outcome of a head-to-head competition. The most common models are the Bradley-Terry model (Bradley and Terry, 1952) and the Thurstone-Mosteller model (Mosteller, 1951). Cattelan (2012) provides a thorough review of the current state of paired comparison modeling. Many games and sports, including various types of races (horse, automobile, human track and field), gymnastics, diving, and golf, involve multiple teams or players competing against each other simultaneously. For such competitions the outcome often of interest is the rank ordering of competitors. Models for rank orderings have been an active area of statistical development, but have received far less attention than modeling results of head-to-head competition. Specific to a sports context, competitors' abilities may be changing over time, and a compelling modeling framework for multi-competitor sports should account for the time-varying nature of ability.

Parametric models for rank orderings have a long history. The common assumption for these models is that each competitor i, $i = 1, \ldots, n$ in an n-player competition has a latent performance Y_i following a specified distribution $F(y|\theta_i)$ with unknown ability parameter θ_i. The probability that player 1 is ranked first, player 2 is ranked second, and so on, can be

expressed in terms of the Y_i as

$$\mathrm{P}(Y_1 > Y_2 > \cdots > Y_n \mid \theta_1, \ldots, \theta_n). \tag{1}$$

Inferences about the θ_i from the results of multiple competitions can then be determined, for example, through likelihood-based approaches involving factors in the form of Equation 1. Early work on parametric models for rank orderings include Plackett (1975) who assume extreme value (i.e., Gumbel) distributions for the Y_i, Henery (1981), Bockenholt (1992), and Bockenholt (1993) who assume the Y_i are normally distributed, and Henery (1983) and Stern (1990) who consider Gamma models for the Y_i. The rank ordering models for the normal and extreme value performance distributions are special cases of the Gamma model (Stern, 1990), though evaluating the permutation probabilities numerically can be difficult for arbitrary Gamma distribution parameters. The rank ordering model based on extreme value distributions is sometimes called the Plackett-Luce model, as an extension of the multinomial logit choice model of Luce (1959) to rank orderings. This model is also commonly called the exploded logit model (Allison and Christakis, 1994), or the rank-ordered logit model (Hausman and Ruud, 1987). Henceforth we will refer to these models as rank ordered logit (ROL) models. Example applications of the ROL model in sports settings include horse-race outcomes (Ali, 1998; Lo and Bacon-Shone, 1994) and NASCAR automobile races (Graves et al., 2003; Guiver and Snelson, 2009).

More recently, interest has focused on models with time-varying parameters. Baker and McHale (2014) assume a ROL model for golfers' abilities and model the change in abilities non-stochastically through barycentric rational interpolants (Taylor, 1945), a particular type of smoother. Herbrich et al. (2007) introduced an approach assuming normal latent performances with normally distributed innovations to the ability parameters, with the ability

parameters estimated through the expectation propagation algorithm (Minka, 2001). Weng and Lin (2011) introduced an approximation procedure based on Stein's method to derive simple updating computations in the context of paired comparison and ROL models. Caron and Teh (2012) developed a Bayesian nonparametric representation of the ROL model, and extend their representation through a Gamma process to account for changes over time.

The approach we develop here is to model multi-competitor game outcomes through a ROL model, and assume competitor strengths evolve over time through a Gaussian random walk. We consider a full Bayesian treatment of the model, and describe how to obtain inferences for the model through Markov chain Monte Carlo (MCMC) simulation from the posterior distribution. An advantage of MCMC simulation in our context is the ease in addressing the occurrence of ties in the rank orderings, a challenge that has previously been computationally problematic. In addition to the full Bayesian analysis of our model, we describe an approximate Bayesian filtering approach that can be used in the context of a large number of competitors or many time periods in which a full Bayesian analysis might be too computationally intensive. The approximate Bayesian filter may also be used to update competitors' abilities as new game results accumulate over time without the need to perform a re-analysis of the entire data set. This filtering approach shares many similarities to the one developed in Glickman (1999) for paired comparisons.

The paper is organized as follows. Section 2 describes the probability model for rank orderings along with the stochastic component for changes in competitor abilities. In this section, we also describe the details of MCMC posterior simulation for competitor abilities. In Section 3, we introduce an approximate Bayesian filter based on the model in the previous section. Both the full Bayesian approach and the approximate Bayesian filter are then applied

4

in Section 4 to a data set on the results of women's Alpine downhill skiing events. The paper concludes in Section 5 with a discussion of the work, along with extensions and limitations.

2 A stochastic model for rank orderings

Consider a population of n competitors who compete in multi-competitor games over T discrete time periods. Assume that during time period t, $t = 1, \ldots, T$, K_t games or contests take place. Also assume that competitor i, $i = 1, \ldots, n$, has an ability parameter θ_{it}, defined formally below, that indicates the competitor's strength during period t. Suppose contest $k = 1, \ldots, K_t$ within period t consist of m_{kt} competitors. Suppressing the dependence on k and t, suppose competitors $1, 2, \ldots, m_{kt}$ are involved in contest k during time period t, and let Y_i be a latent performance by competitor $i = 1, \ldots, m_{kt}$. We assume,

$$Y_i | \theta_{it} \sim \text{Gumbel}(\theta_{it}) \tag{2}$$

with cumulative distribution function of the extreme value/Gumbel distribution

$$F(y|\theta) = \exp\left(-e^{-(y-\theta)}\right) \tag{3}$$

We further assume that the Y_i within contest k are conditionally independent given the θ_{it}.

Suppose that the observed outcome of contest k is a rank ordering of the latent performances $Y_1, \ldots, Y_{m_{kt}}$. It is straightforward to show that, conditional on $\boldsymbol{\theta}_t = (\theta_{1t}, \ldots, \theta_{nt})$,

$$L_{kt} = \text{P}(Y_1 > Y_2 > \cdots > Y_{m_{kt}} | \boldsymbol{\theta}_t) = \prod_{i=1}^{m_{kt}-1} \frac{\exp(\theta_{it})}{\sum_{\ell=i}^{m_{kt}} \exp(\theta_{\ell t})}. \tag{4}$$

The model defined in (2) and (4) is the ROL model. The probability of a particular rank ordering can be understood as the product of multinomial choice probabilities over diminishing choice sets; the product of the probability the winner outperforms all competitors

with the probability the second-place finisher outperforms all but the winner with the probability the third-place finisher outperforms all except the first and second-place finishers, and so on. Thus the likelihood contribution L_{kt} for a single rank ordering of m_{kt} competitors is the product of $m_{kt} - 1$ multinomial logit probability factors; the m_{kt}-th factor in the telescoping product of probabilities is 1 so that it is not necessary to include. The assumption of independent extreme value performance distributions leads to multinomial logit choice probabilities. For the ROL model, it is common to impose a linear constraint on the ability parameters $\boldsymbol{\theta}_t$, such as $\sum_{i=1}^{n} \theta_{it} = 0$, because the model in (4) is uniquely specified only up to an additive constant.

To account for the possibility that competitors' abilities are changing over time, we assume a Gaussian random walk on $\boldsymbol{\theta}$. Together with the ROL model component, the overall model is an instance of a dynamic generalized linear model (Ferreira and Gamerman, 2000; West et al., 1985). Our approach assumes a stochastic process on the $\boldsymbol{\theta}_t$ in which

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{\delta}_{t+1} \tag{5}$$

where

$$\boldsymbol{\delta}_{t+1} \sim \mathrm{N}(\mathbf{0}, \ \boldsymbol{\Upsilon}). \tag{6}$$

The innovation covariance matrix, $\boldsymbol{\Upsilon}$, can be constrained to ensure that the average of the $\boldsymbol{\theta}_{t+1}$ across competitors is the same as the average of the $\boldsymbol{\theta}_t$. Such a constraint acknowledges that the rank orderings provide no information on systematic shifts in the $\boldsymbol{\theta}_t$. This constraint is accomplished by setting

$$\boldsymbol{\Upsilon} = \tau^2 \left(\boldsymbol{I} - \frac{1}{n} \mathbf{1}\mathbf{1}' \right), \tag{7}$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix, $\mathbf{1}$ is the n-vector with 1 as each element, and τ^2 is a

scalar parameter. Thus the marginal variance of the i-th innovation term is $\frac{n-1}{n}\tau^2$, and the correlation between the i-th and j-th innovation terms is $\frac{-1}{n-1}$. Generating a normal vector with the covariance matrix in (7) is identical to simulating values from independent $N(0, \tau^2)$ distributions and then zero-centering the vector by subtracting the sample mean. For a large population of competitors (i.e., with n large), the innovations have a slight negative correlation.

The stochastic process on the $\boldsymbol{\theta}_t$ can be extended in a variety of ways. For example, an autoregressive process on the $\boldsymbol{\theta}_t$ may be assumed such as

$$\boldsymbol{\theta}_{t+1} = \nu\boldsymbol{\theta}_t + \boldsymbol{\delta}_{t+1} \tag{8}$$

where

$$\boldsymbol{\delta}_{t+1} \sim N(\mathbf{0}, \ \tau^2\boldsymbol{I}) \tag{9}$$

and ν is an autoregressive parameter. Such models have been used in the context of measuring ability in sports including Glickman and Stern (1998), and Glickman (1999).

The model specification is completed by a prior distribution. A flexible choice of a prior distribution component on the initial competitor strengths, $\boldsymbol{\theta}_1$, is

$$\boldsymbol{\theta}_1 | \sigma_1^2 \sim N(\mathbf{0}, \ \sigma_1^2\boldsymbol{I}). \tag{10}$$

More generally, an arbitrary multivariate normal prior distribution for $\boldsymbol{\theta}_1$ may be assumed rather than one centered on $\mathbf{0}$ and with independent prior components. A prior distribution for the two variance parameters τ^2 and σ_1^2 may be assumed. Conjugate inverse-Gamma distributions have been used in previous work for Gaussian state model parameters (West et al., 1985).

7

Because the ROL likelihood is a special case of a multinomial logit likelihood, inference for the ROL state-space model can use the same computational approaches as those of dynamic multinomial logit models. Recent work on inference in multinomial logit state-space models have appealed to MCMC simulation from the posterior distribution. Early MCMC approaches for state-space models with non-normal responses relied on sampling time-specific parameters conditional on the neighboring parameter values. Examples included Carlin et al. (1992) and Gamerman and Migon (1993). Cargnoni et al. (1997) proposed an efficient MCMC sampling scheme based on conditionally Gaussian dynamic models.

In many multi-competitor game settings, ties can occur. Sports that involve accrual of discrete point values (e.g., strokes in golf) can result in competitors with identical totals at the end of the competition. Certain types of races and sports settings where competitors have a limited amount of time to achieve a goal may result in competitors who do not finish or complete the desired task. These competitors would then tie for last place. The model in (4) does not directly apply to games and sports in which ties occur.

Two strategies may be considered for adjusting the ROL model. The first approach is to explicitly model the occurrence of ties. In the context of the latent performance model, a tie occurs when latent performances are sufficiently close. For example, if competitors $1, \ldots, d$ tie in a competition, then the model can require

$$\max_{a,b \in \{1,2,\ldots,d\}} |Y_a - Y_b| < \kappa. \tag{11}$$

Given κ and the ability parameters $\boldsymbol{\theta}$, the probability of a tie in (11) can be approximated by Monte Carlo simulation, if not directly. Johnson et al. (2002) in the context of latent Normal performance distributions assume a similar model for ties in which the probability

that two competitors tie is a smoothly decreasing function of the difference between latent performances.

A second strategy, and the approach we assume here, is that the model itself does not recognize that ties exist, but when ties occur the likelihood is specified as a mixture over all possible permutations of the collection of tied competitors. For example, if competitors 1, 2, 3, 4, 5 engage in a contest with competitor 1 winning, competitor 5 in last place, and competitors 2, 3 and 4 tying for second place, then the likelihood would be an average over six ROL probabilities corresponding to the following rankings: $(1, 2, 3, 4, 5)$, $(1, 2, 4, 3, 5)$, $(1, 3, 2, 4, 5)$, $(1, 3, 4, 2, 5)$, $(1, 4, 2, 3, 5)$, and $(1, 4, 3, 2, 5)$, where the position of the value in the 5-tuplet is the rank of that competitor. Recognizing the connection to partial likelihoods for survival data, the mixture likelihood approach was proposed by Kalbfleisch and Prentice (2011) and described in the ROL model context by Allison and Christakis (1994). Glickman (1999) made a similar assumption in the context of paired comparison models with ties.

Allison and Christakis (1994) and Baker and McHale (2014) note that inference for the mixture likelihood is computationally intractable when more than a few competitors are involved in ties. Inference in a Bayesian setting, however, permits a straightforward Monte Carlo estimate of the mixture likelihood. Rather than evaluate the mixture likelihood over all possible permutations of competitors involved in ties, we take the approach of randomly permuting the indices of competitors involved in ties at the start of each MCMC iteration, and then condition on the permutation when simulating model parameters from their conditional posterior distributions. This approach is identical to a common strategy in MCMC simulation for general mixture models in which the mixture component label is treated as a latent variable whose distribution is inferred through Monte Carlo integration (Carlin and

Chib, 1995; Jasra et al., 2005). Upon convergence of MCMC, the simulated draws of the ability and variance parameters are simulations from the mixture distribution.

3   Development of a multi-competitor rating system

Some games and sports settings require inferences about strength to be computed on an ongoing basis for large numbers of competitors. For example, when tracking competitor abilities over time for populations of athletes, or when constructing a rating system for league play, the methods developed in Section 2 may be too computationally intensive to be performed regularly. One common approximation in state-space models for updating the state parameters is the use of particle filters (Doucet et al., 2000, 2001). An early application of particle filters in a sports context involved updating NFL football team strengths (Glickman and Stern, 1998). While updating particles is usually a quick computation, a challenge is that most of the posterior mass tends to concentrate on a small number of particles upon successive filtering steps. An alternative approach that we develop here is based on approximating the posterior distribution of abilities by a normal distribution each time period data are observed, and performing ability parameter updates through a Newton-Raphson algorithm that determines the posterior mode and second derivative at the mode. The updated means are approximated by the posterior mode, and the updated variances are obtained by taking the negative of the inverse of the second derivative. Prior to applying the rating procedure we develop here, the variance parameters τ^2 and σ_1^2 are treated as fixed and known. These parameters may be set at summary estimates (e.g., posterior means) based on a full Bayesian analysis as in Section 2. Alternatively, these parameters may be chosen through optimization based on predictive fit criteria which we discuss below.

10

The rating updating procedure we develop is intended to be applied recursively at the start of each rating period. At the beginning of a new rating period, the prior distribution of competitor strengths is assumed to consist of independent normal distribution components. Game data are observed during the rating period, and then approximate normal posterior distributions are computed for each competitor using the algorithm described below. Finally, to obtain the prior normal distribution for the next rating period, the addition of the innovation variance τ^2 is applied to all competitor strength distributions. This sequence of steps is applied recursively over successive rating periods.

Suppose prior to games in time period t, the ability distributions for competitors $i = 1, \ldots, n$ are specified independently as

$$\theta_{it} \sim \mathrm{N}(\mu_{it}, \ \sigma_{it}^2). \tag{12}$$

Assume K_t competitions occur during period t. We first consider the case in which no ties occur. Suppressing the dependence on k and t, suppose competitors $1, 2, \ldots, m_{kt}$ compete in competition k in which player 1 places first, 2 places second, and so on. The likelihood contribution for competition k for this rank order is given by Equation (4).

Again suppressing dependence on t, we now define two $(m_k - 1) \times n$ matrices essential for the description of the computational algorithm. Let $\boldsymbol{X}_k$ be the $(m_k - 1) \times n$ matrix in which columns are indexed by every competitor in the population and the i-th row encodes the choice set of competitors involved in i-th factor of L_{kt} in (4). More concretely, for $i = 1, \ldots, m_k - 1$, $(\boldsymbol{X}_k)_{ij} = 1$ if competitor j is in the choice set (the indices of the competitors with terms in the denominator) of the j-th multinomial logit probability factor of L_{kt}, and 0 if not.

Let $\boldsymbol{W}_k$ be the $(m_k-1) \times n$ matrix in which columns are indexed by competitors, and all the elements of the i-th row are 0 except for the element corresponding to the i-th place finisher which is set to 1. Therefore $(\boldsymbol{W}_k)_{ij} = 1$ if the numerator of the i-th multinomial logit probability factor of L_{kt} involves competitor j, and 0 otherwise.

Letting $\boldsymbol{\mu}_t = (\mu_{1t}, \ldots, \mu_{nt})$ and $\boldsymbol{\sigma}_t = (\sigma_{1t}, \ldots, \sigma_{nt})$, the log of the posterior distribution up to an additive constant can be written as

$$
\begin{aligned}
\log &p(\boldsymbol{\theta}_t | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{K_t}, \boldsymbol{W}_1, \ldots, \boldsymbol{W}_{K_t}, \boldsymbol{\mu}_t, \boldsymbol{\sigma}_t) \\
&= \ C^* + \log p(\boldsymbol{\theta}_t | \boldsymbol{\mu}_t, \boldsymbol{\sigma}_t) + \sum_{k=1}^{K_t} \log L_{kt} \\
&= \ C^{**} - \left(\sum_{i=1}^{n} \frac{(\theta_{it} - \mu_{it})^2}{2\sigma_{it}^2} \right) + \sum_{k=1}^{K_t} (\boldsymbol{W}_k \boldsymbol{\theta}_t - \log(\boldsymbol{X}_k \boldsymbol{\eta}_t))' \mathbf{1}_{m_k - 1}
\end{aligned}
\tag{13}
$$

where C^* and C^{**} are functions of normalizing constants, $\boldsymbol{\eta}_t = \exp(\boldsymbol{\theta}_t)$, and $\mathbf{1}_{m_k-1}$ is the $(m_k - 1)$-vector with every element set to 1.

We can determine an approximating normal posterior distribution of $\boldsymbol{\theta}_t$ by numerically finding the mode of the log-posterior distribution in (13). The mode can then be used as the approximate normal posterior mean. To obtain the posterior covariance matrix, we evaluate the second derivative matrix of the (13) at the mode, and then find the negative of the matrix inverse to approximate the posterior covariance of $\boldsymbol{\theta}_t$. This optimization can be accomplished through the Newton-Raphson algorithm, though other numerical optimization procedures are possible. The steps of the Newton-Raphson procedure to obtain the approximate normal posterior distribution are outlined in Appendix A.

In practice, two modifications can be made to the above procedure that recognize the computational difficulty of working with large populations of competitors. First, competitors

who do not compete during period t only contribute to (13) through the prior distribution term, and the approximating normal posterior distribution for such competitors is their prior distribution. Therefore, rather than specifying $\boldsymbol{X}_k$, $\boldsymbol{W}_k$, and $\boldsymbol{\theta}_t$ in terms of the full population of n competitors, it is sufficient to redefine these terms involving only competitors who competed in period t. Second, rather than saving the full posterior covariance matrix resulting from the computation, we set the posterior covariances to 0 which results in a normal posterior distribution that is composed of independent competitor-specific normal distributions. This thresholding to 0 may be justified by acknowledging that the covariances generally are likely to be weak, and that retaining the covariances would involve replacing the first term in the sum in (13) with computation requiring the inversion of large covariance matrices.

As a result of the optimization of (13), we obtain approximate normal marginal posterior distributions for each competitor of the form

$$\theta_{it}|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{K_t}, \boldsymbol{W}_1, \ldots, \boldsymbol{W}_{K_t}, \boldsymbol{\mu}_t, \boldsymbol{\sigma}_t \sim \mathrm{N}(\mu_{it}^*, \sigma_{it}^{2*}). \tag{14}$$

For the rating procedure, we assume for each i and t

$$\theta_{i,t+1}|\theta_{it}, \tau^2 \sim \mathrm{N}(\theta_{it}, \tau^2) \tag{15}$$

Therefore, the distribution of $\theta_{i,t+1}$ conditional only on the parameters associated with period t is given

$$
\begin{aligned}
p(\theta_{i,t+1}|\mu_{it}^*, \sigma_{it}^{2*}, \tau^2) &= \int \mathrm{N}(\theta_{it}|\mu_{it}^*, \sigma_{it}^{2*})\mathrm{N}(\theta_{i,t+1}|\theta_{it}, \tau^2)d\theta_{it} \\
&= \mathrm{N}(\theta_{i,t+1}|\mu_{it}^*, \sigma_{it}^{2*} + \tau^2)
\end{aligned} \tag{16}
$$

where $\mathrm{N}(\cdot|\mu, \sigma^2)$ is a normal density with mean μ and variance σ^2.

13

When ties are present in rank orderings, the approach described in Section 2 in the context of MCMC posterior simulation cannot easily be applied to the rating procedure. We instead use an approximation due to Breslow and Crowley (1974). Their approach involves each competitor involved in a (possibly multi-way) tie having a separate factor in the ROL likelihood. Each factor is the multinomial logit probability that each competitor in a tie outperforms all others involved in the tie as well as the other competitors that are ranked lower. The similarity of the performances among tied competitors is therefore captured through factors in the likelihood that have each competitor outperforming the others. As an example, in a competition with six competitors in which player 1 is ranked first, players 2, 3 and 4 tie for second place, player 5 comes in fifth place, and player 6 comes in sixth place, the likelihood contribution would be

$$
\left(\frac{\exp(\theta_1)}{\sum_{i=1}^6 \exp(\theta_i)} \right) \left(\frac{\exp(\theta_2)}{\sum_{i=2}^6 \exp(\theta_i)} \right) \left(\frac{\exp(\theta_3)}{\sum_{i=2}^6 \exp(\theta_i)} \right) \left(\frac{\exp(\theta_4)}{\sum_{i=2}^6 \exp(\theta_i)} \right) \left(\frac{\exp(\theta_5)}{\sum_{i=5}^6 \exp(\theta_i)} \right).
\tag{17}
$$

Note that all three middle factors in (17) have denominators that involve competitor parameters $\theta_2, \ldots, \theta_6$. Baker and McHale (2014) adopt this approach to ties in their ranking model, and demonstrate through simulations that the approximation results in inferences that produce little difference from the mixture likelihood approach described in Section 2.

This approach to modeling ties can be incorporated in a straightforward manner into our rating procedure. As long as competitors who tie are assigned the same rank, the algorithm as described above implements the approximation by Breslow and Crowley (1974). This is because the rows of $\boldsymbol{X}_k$ in (13) recognize that the choice set over which the multinomial logit probabilities are specified include competitors with the same rank (i.e., that are tied). As before, the normal prior distribution of $\boldsymbol{\theta}_t$ is updated to an approximate

14

normal posterior distribution by optimizing (13).

As described, the rating procedure is a filtering algorithm. The algorithm determines an approximate posterior distribution of competitor strength $\boldsymbol{\theta}_t$ based on all competition results prior to or during period t. The rating procedure can be adapted to smooth earlier parameters based on later results, that is, provide inferences about $\boldsymbol{\theta}_t$ based on all competition results through the final time period. This can be accomplished through the Rauch-Tung-Streibel (RTS) smoother (Rauch et al., 1965), a particular version of the Kalman smoother. The computations involved with the RTS smoother assume posterior competitor distributions of θ_{it}, for $t = 1, \ldots, T$

$$\theta_{it}|\boldsymbol{Y}_t \sim \mathrm{N}(\mu_{it}^*, \sigma_{it}^{2*}) \tag{18}$$

where $\boldsymbol{Y}_t$ denotes all game results up through and including period t, and where μ_{it}^* and σ_{it}^{2*} are the posterior parameters of the approximating normal distribution. The RTS smoother is implemented recursively in the following manner. First, let

$$\begin{aligned} \tilde{\mu}_{iT} &= \mu_{iT}^* \\ \tilde{\sigma}_{iT}^2 &= \sigma_{iT}^{2*}. \end{aligned} \tag{19}$$

Then, for $t = T-1, T-2, \ldots, 1$, compute

$$\begin{aligned} \tilde{\mu}_{it} &= \mu_{it}^* + \left(\frac{\sigma_{it}^{2*}}{\sigma_{it}^{2*} + \tau^2}\right)(\tilde{\mu}_{i,t+1} - \mu_{it}^*) \\ \tilde{\sigma}_{it}^2 &= \sigma_{it}^{2*} + \left(\frac{\sigma_{it}^{2*}}{\sigma_{it}^{2*} + \tau^2}\right)^2 (\tilde{\sigma}_{i,t+1}^2 - \sigma_{it}^{2*} - \tau^2). \end{aligned} \tag{20}$$

These computations result in the smoothed parameters of

$$\theta_{it}|\boldsymbol{Y}_T \sim \mathrm{N}(\tilde{\mu}_{it}, \tilde{\sigma}_{it}^2) \tag{21}$$

Our approximate Bayesian approach treats τ^2 and σ_1^2 as fixed in advance. Rather than estimating τ^2 and σ_1^2 through a full Bayesian analysis, estimation may be performed by optimizing a predictive fit criterion. The approach we adopt here is to maximize a weighted average of Spearman rank correlations (Spearman, 1904) between the (filtered) posterior mean strengths and the rank order of competitors in an event during the next time period, and average these values over a validation set of time periods. Specifically, for candidate choices of τ^2 and σ_1^2, we perform the approximate Bayesian filter to obtain the approximate normal posterior for the $\theta_{is} \sim \text{N}(\mu_{is}, \sigma_{is}^2)$ at time period $s < T$. We then compute the Spearman rank correlation between the μ_{is} and the rank order for event k during time period $s + 1$ for each of the K_{s+1} events to obtain predictive measures of fit $\rho_{k,s+1}$ for $k = 1, \ldots, K_{s+1}$. This process is repeated for each subsequent time period to obtain predictive Spearman correlations. Thus, for each $t = s, s + 1, \ldots, T - 1$, we compute the approximate posterior means μ_{it} from the filtering algorithm, and calculate the Spearman correlation with the competitors' ranks in each event in period $t + 1$. The weighted average of correlations over the validation periods $t = s + 1, \ldots, T$ is then given by

$$\rho_W = \frac{\sum_{t=s+1}^{T} \sum_{k=1}^{K_t} (m_{kt} - 1)\rho_{k,t}}{\sum_{t=s+1}^{T} \sum_{k=1}^{K_t} (m_{kt} - 1)} \tag{22}$$

This approach to constructing weighted averages of Spearman rank correlations as an overall correlation measure is described in Taylor (1987).

With the computation for ρ_W, the variance parameters can be optimized through common optimization algorithms, such as the Nelder-Mead optimization algorithm (Nelder and Mead, 1965). In our context, the Nelder-Mead algorithm optimizes τ^2 and σ_1 to result in the largest value of ρ_W given the data and the choice of the time period s at which the predictive measure is computed. While the Spearman rank correlation takes on only finitely

many values, we have found in our examples that this does not hinder the Nelder-Mead algorithm which assumes a continuous objective function.

4 Application to women's Alpine downhill skiing

We apply the methods developed in Sections 2 and 3 to the results of women's competitive Alpine downhill skiing over the period from February 12, 2002 to December 7, 2013. The data set consists of the results of 103 elite women's skiing events administered by the Fédération Internationale de Ski (FIS), including the Olympic games, World championships, World Cup, and a variety of regional events. A total of 268 women skiers competed in these events, averaging 1.578 events per year. Many of the women competed infrequently; 51 of the women competed in only one event, and 26 competed in only two events over the twelve year period. However, two skiers competed in 90 or more of the 103 events. The data were provided to us by the U.S. Olympic Committee.

Alpine downhill skiing competitions within the FIS are governed by the World Cup scoring system for each event. Race completion times are converted to integer point scores. We were not provided with actual race completion times. Depending on the event, each competitor may have had multiple scored runs, and the total score for a competitor in an event was the sum of the points earned for each run. The discretized point scoring therefore frequently resulted in ties, typically for lower finish positions in events. In our competition results data, a total of 4.4% of the final positions in events were ties. In addition to ties based on equal total FIS points, many competitors in events did not receive any points. The scoring system awards points on a given run to the top 30 finishers, so those who did not finish in

the top 30 in any run did not receive any FIS points in the event. These competitors could be treated as tying for last place in an event. A total of 8.7% of the standings in events were ties based on not receiving points for the event. Thus a total of 13.1% of the final standings in events were ties.

For our main analyses, we divided the 12-year period of results into 24 six-month rating periods, with an average of 4.3 events per period. Two periods (July-December 2002 and January-June 2003) had no events recorded, while three of the periods consisted of as many as seven competitions (January-June 2005, 2009 and 2010) which was the maximum in a 6-month period. Using six month rating periods is a compromise between having periods long enough over which skiers abilities are not changing appreciably, and short enough to detect changes in ability.

We fit the model in Section 2 simulating from the posterior distribution of skiers' abilities via MCMC using the Bayesian software package JAGS (Plummer, 2003) called from within the computing package R (R Foundation for Statistical Computing, 2012). Two parallel chains were run with dispersed starting values, each with a burn-in of 10000 iterations. Each chain ran for an additional 20000 iterations, and posterior inferences were based on these sets of simulated parameter draws. Convergence diagnostics (Cowles and Carlin, 1996) indicated that the MCMC simulation had reached stationarity.

We also ran our approximate Bayesian filter to the same game outcome data as described in Section 3. We used the final three rating periods, equal to 1.5 years of competitions, to compute a predictive weighted average Spearman rank correlation within a Nelder-Mead optimization algorithm. Assuming a different numbers of validation periods on which to

| Parameter | Full Bayes | | Approximate Bayes |
	Posterior Mean	95% Central Posterior Interval	Optimized Value
τ	0.352	(0.310, 0.399)	0.295
σ_1	0.451	(0.083, 0.668)	0.313

Table 1: Posterior inferences for τ, the innovation standard deviation, and σ_1, the standard deviation of skier abilities in 2002. First two columns are the results for the full Bayesian analysis, and the final column summarizes the optimized values from the approximate Bayesian filter.

compute the correlation measure did not result in substantively different optimized variance parameters.

Table 1 displays the resulting estimated standard deviation parameters from the full Bayesian and the approximate Bayesian methods. The first two columns contain the MCMC-estimated posterior means and 95% central posterior intervals for the standard deviation of skiers' initial abilities, σ_1, and the innovation standard deviation, τ, as displayed in Equations (7) and (10). The third column displays the optimized values from the approximate Bayesian filter. The standard deviation of initial abilities among elite women skiers indicates an appreciable amount of variation, but it is imprecisely estimated from the data based on the full Bayesian analysis. The innovation standard deviation is more precisely estimated as indicated by the full Bayesian analysis, and it suggests the possibility of large shifts in ability between 6-month periods. The optimization for the approximate Bayesian analysis resulted in standard deviation estimates that were somewhat lower than the posterior means. The optimized value of σ_1 from the approximation algorithm was within the 95% central posterior interval, but the optimized value of τ was lower than the 95% central posterior interval.

We label the 24 time periods for our analyses as $(2002.1, 2002.2, 2003.1, \ldots, 2013.2)$. In Table 2 we summarize in the first two columns the posterior mean and standard deviations for the top 20 women skiers ranked according to their MCMC-estimated posterior means among the subset of 108 skiers who competed in events since January 2012. The third and fourth columns of Table 2 summarize the posterior means and standard deviations from the approximate Bayesian algorithm for the skiers in the first two columns. The distribution of posterior means of the $\theta_{i,2013.2}$ for this group ranged from -1.517 to 2.707 for the full Bayesian analysis, and from -1.813 to 2.144 for the approximate Bayesian analysis. From Table 2, most of the top twenty skiers have posterior means of the $\theta_{i,2013.2}$ that are close. Accounting for the posterior uncertainty of these values, the relative abilities between adjacent skiers are nearly indistinguishable, though skiers who are further apart on the list have more clearly distinguishable abilities. It is worth noting that two skiers not on the list (Hilde Gerg and Michaela Dorfmeister) had posterior mean ability parameters of 3.264 and 2.229, respectively, from the full Bayesian analysis, which were quite a bit higher than those on Table 2. Both of these skiers had consistently impressive performances in their last set of competitions, but Dorfmeister last competed in 2006 and Gerg in 2005.

The approximate posterior means in the third column are ordered similarly to the MCMC-estimated means in the first column, suggesting a strong correspondence between the rankings of top skiers from both analyses. The biggest exception occurs with Elena Fanchini who is estimated to be four places lower on the list. The distribution of approximate posterior means among the top 20 is shifted down by about 0.5–0.55 compared to the means in the full Bayesian analysis. This constant difference does not affect the probability calculation for rank orderings because the ROL probabilities are the same up to an additive constant

	Full Bayes		Approximate Bayes	
	Posterior	Posterior	Posterior	Posterior
Skier	Mean	Std Dev	Mean	Std Dev
Tina Maze	2.707	0.414	2.144	0.380
Marion Rolland	2.556	0.537	2.009	0.467
Anna Fenninger	2.386	0.477	1.749	0.435
Elena Fanchini	2.121	0.463	1.494	0.420
Julia Mancuso	2.072	0.355	1.623	0.313
Lara Gut	2.061	0.439	1.507	0.389
Elisabeth Görgl	2.053	0.414	1.520	0.374
Maria Höfl-Riesch	2.045	0.452	1.552	0.395
Lindsey Vonn	1.816	0.399	1.467	0.349
Johanna Schnarf	1.682	0.526	1.163	0.478
Marianne Kaufmann-Abderhalden	1.675	0.476	1.058	0.434
Viktoria Rebensburg	1.663	0.464	1.095	0.425
Regina Sterz	1.487	0.423	0.919	0.389
Stacey Cook	1.482	0.382	0.954	0.355
Dominique Gisin	1.433	0.441	0.888	0.403
Ilka Štuhec	1.343	0.441	0.753	0.407
Carolina Ruiz Castillo	1.296	0.396	0.763	0.407
Stefanie Moser	1.235	0.448	0.680	0.403
Fabienne Suter	1.223	0.375	0.750	0.338
Verena Stuffer	1.181	0.422	0.604	0.390

Table 2: Posterior mean and standard deviation of skier abilities in the second half of 2013, for 20 skiers with highest means among 108 active skiers in 2012-2013 based on the full Bayesian analysis and based on the approximate Bayesian approximation.

on the scale of the strength parameters. The approximate posterior standard deviations in the fourth column are estimated to be slightly smaller than the corresponding values in the second column, but not by an amount that would result in conflicting conclusions.

Two high-profile skiers in our data set who have actively competed over most of the 12-year period are the American athletes Lindsey Vonn and Julia Mancuso. Figure 1 displays the posterior mean strength of each skier along with 95% central posterior intervals around the means for both the full Bayesian analysis and the smoothed approximate Bayesian rating system analysis. In both cases, Mancuso and Vonn appear to have similar abilities up through about 2008, at which point Vonn experienced substantially improved performances from 2009 to 2013. The strength summaries from the full Bayesian analysis (left figure) appear to span a greater range compared to the approximate Bayesian analysis (right figure), as the peak mean strength for Lindsey Vonn in 2011 reaches over 5.0 in the full Bayesian analysis, but only 4.0 in the approximate Bayesian analysis. The overall trends of both skiers match closely in the two approaches.

Let $\theta_{V,t}$ and $\theta_{M,t}$ be the ability parameters for Vonn and Mancuso, respectively, at time t. Figure 2 displays the pointwise posterior means over time for $\exp(\theta_{V,t})/(\exp(\theta_{V,t}) + \exp(\theta_{M,t}))$, the probability Vonn would outperform Mancuso (i.e., obtain a higher place in an event) at time t. The figure shows the posterior means computed both for the full Bayesian analysis and the approximate Bayesian analysis. Despite the more compressed estimated posterior means from the approximate Bayesian analysis, the posterior probabilities Vonn outperformed Mancuso are close for both analyses as can be seen by the coinciding probability trends. The probabilities differ by no more than 0.05 for all estimated probabilities with the exception of the second 6-month period in 2007 in which the probability difference was 0.076.
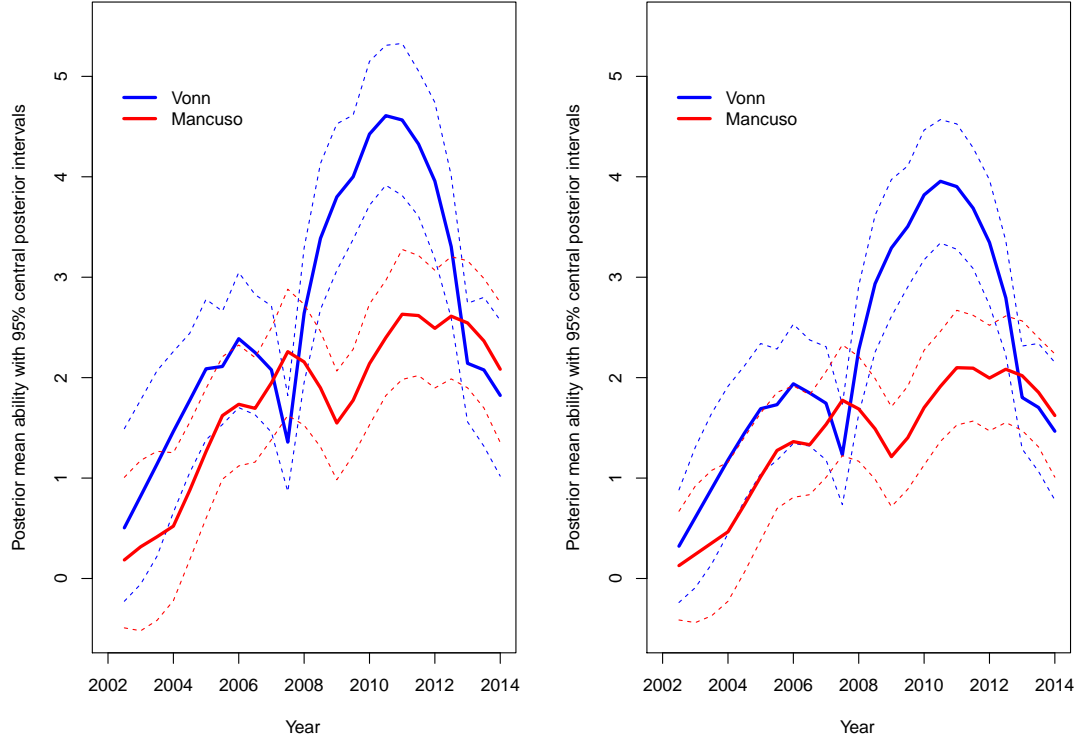
Figure 1: Pointwise means and 95% central posterior intervals of ability parameters over time for Julia Mancuso and Lindsey Vonn. *Left*: Summaries based on full Bayesian analysis. *Right*: Summaries based on approximate Bayesian analysis.

Consistent with Figure 1, Vonn would be expected to outperform Mancuso prior to 2007, at which point Mancuso appeared to be somewhat better. Then from 2009 to 2013, the probability Vonn would outperform Mancuso is close to 0.9. By 2013, the two skiers are inferred to be of essentially similar strengths.

We assessed the predictability of the full Bayesian and approximate Bayesian approaches by predicting rank orders of seven events held between December 21, 2013 and March 12, 2014. The seven events, which were not used in the previous model fitting, consisted of 73 of the 268 women skiers. Six skiers in these seven events were not in the model
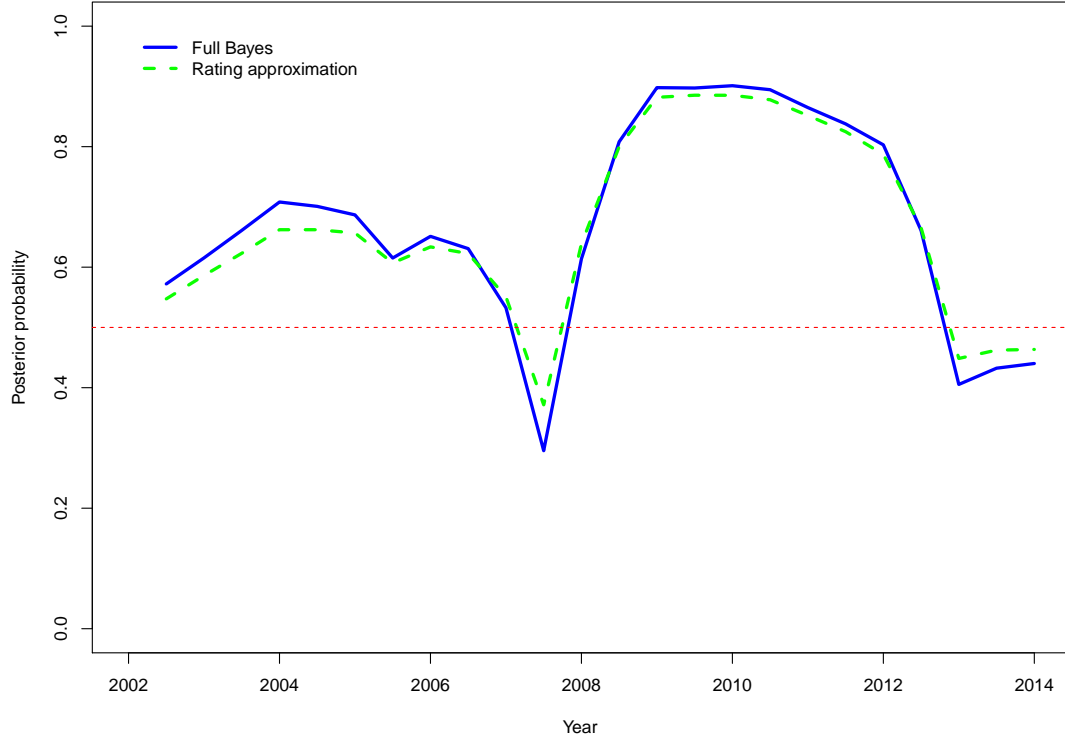
Figure 2: Pointwise posterior probability by year Lindsey Vonn would outperform Julia Mancuso based on full Bayesian analysis and approximate Bayesian rating analysis. The horizontal dashed line is drawn at a probability of 0.5.

development sample of 268, so their results in the seven events were removed.

We computed the weighted average of Spearman correlations between the rank orderings of the $\mu_{i,2013.2}$ and the final finish places across the seven events as our measure of predictive accuracy, as described in Section 3. The weighted average of the Spearman correlations are summarized in Table 3. In addition to the full Bayesian and approximate Bayesian analyses based on dividing the original sample into 6-month periods, we performed the full Bayesian and approximate Bayesian analyses dividing the sample into 3-month periods for one set of analyses and 12-month periods for another set. To ensure correspondence

24

Approach	Weighted Spearman's correlation
Full Bayes, 3-month rating periods	0.580
Full Bayes, 6-month rating periods	0.586
Full Bayes, 12-month rating periods	0.614
Approximate Bayes, 3-month rating periods	0.601
Approximate Bayes, 6-month rating periods	0.587
Approximate Bayes, 12-month rating periods	0.596

Table 3: Performance of six different updating procedures on predicting event results for 7 future events. Results are summarized by averaging Spearman rank correlations between skiers' standings and the rank order of the 2013 posterior mean strength weighted inversely by the variance of the correlations.

with the analysis for 6-month periods, the variance parameters for the approximate Bayesian analyses for 3-month and 12-month periods were optimized for 6 validation periods and 2 validation periods, respectively, corresponding to the final 1.5 years of game results and 2 years of game results. As seen in the table, the average rank correlations are comparable with values ranging between 0.580 and 0.614. The most predictive method by this measure is the full Bayesian analysis of 12-month periods, followed by the approximate Bayesian analysis of 3-month periods. Because a conservative standard error estimate of a Spearman rank correlation is $1/(m-1)$, the value when the samples are uncorrelated and where m is the number of objects involved in the ranking (Zar, 1972), an estimate of the standard error of the weighted Spearman rank correlations is computed to be about 0.06. With a standard error of this magnitude, none of the predictions in Table 3 substantially outperforms any of the others.

5 Discussion

The approaches described in this paper to infer competitor abilities in multi-competitor games assume a flexible set of assumptions; competitor performances follow extreme value distributions, and the mean performance varies stochastically over time through a Gaussian random walk. Our full Bayesian approach applies standard computational machinery via MCMC simulation from the posterior distribution to infer model parameters. By fitting models through MCMC simulation, rank orderings with ties pose no difficulties. The approximate Bayesian filter we developed produces parameter estimates that do not always match the full Bayesian approach, but in our applications the probabilities of rank orderings are quite close.

One main difference between the full Bayesian analysis and the approximate Bayesian filter is that the former retains covariance information between pairs of competitors over time. The approximate Bayesian filter assumes after each time period that the covariances return to 0. If a pair of competitors participate in multiple events with any regularity, a positive covariance between the strength parameters will be induced. The positive covariance leads to more precise inferences about the relative abilities of the pair of competitors. We found in our analyses with Lindsey Vonn and Julia Mancuso, who both competed frequently and in the same events, that the loss of covariance information in the approximate Bayesian filter did not degrade the performance of the probability one outperforms the other relative to the full Bayesian analysis. It may very well be that an application in which competitions occur more frequently and that pairs of players compete more regularly would be needed for the approximate Bayesian filter to evidence noticeably less reliable inferences.

Our approach is specifically designed to model games and sports in which the outcome is a rank ordering. It is worth noting that this approach can be easily modified to address multi-competitor sports in which one competitor is singled out as a winner, and rank orders are not relevant. Because the ROL likelihood in Equation (4) is a telescoping product of multinomial logit probabilities of each competitor outperforming the rest, the setting with a winner and no other players ranked is simply the first multinomial logit probability factor of the ROL likelihood. Thus the methods in this paper apply to this revised setting. Furthermore, more arcane game variants in which multiple players tie for first place by design (e.g., games in which several competitors survive or exceed a threshold criterion and therefore tie for first place) can equally be addressed within our framework.

One consideration in applying the full Bayesian approach versus the approximate Bayesian filter is the trade-off between speed and accuracy. Our experience was that running the full Bayesian analysis on a Windows PC laptop (Intel Core i7 CPU Q 720 @ 1.6 GHz, 6.0GB RAM) took two days to complete, whereas optimizing the approximate Bayesian filter and then performing the filter and smoother on optimized variance parameters took less than 1 minute. Thus the approximate Bayesian approach results in enormous savings in time. Our analyses suggest that the correspondence in estimated ability parameters between the full and approximate Bayesian approaches is strong, and the outperformance probabilities between pairs of competitors calculated using the two different approaches differ by amounts that are negligible for practical purposes. If a goal of the analysis is to update competitor abilities on an ongoing basis as new game outcome data are collected, the difference in computational speed between the full and approximate Bayesian approaches justifies the use of the filtering approach despite the small level of inaccuracy.

In many multi-competitor sports settings, actual performance measures are recorded which provide greater detail than merely the rank ordering. For example, in human, automobile, and horse races, race completion times are often available. In these situations, it is usually preferable to model the actual performance measures as the additional detail can result in more accurate descriptions of competitor strength. However if such information is not available or unreliably recorded, or if a more robust approach that does not rely on precise model formulations for performance measures is desired, then methods that assume only rank orderings for game outcomes are appropriate. Similarly, if a goal is ultimately to make predictions for rank orderings in games and sports, then our framework is a potentially useful approach and worthy of consideration.

Acknowledgments

A   Newton-Raphson algorithm for optimizing log-posterior

We outline the steps for implementing the Newton-Raphson algorithm to find the posterior mode of $\boldsymbol{\theta}_t$ in Equation (13). Let the first and second derivatives of Equation (13) as functions

of $\boldsymbol{\theta}_t$ be

$$D_1(\boldsymbol{\theta}_t) = (D_{1.1}(\boldsymbol{\theta}_t), \ldots, D_{1.n}(\boldsymbol{\theta}_t))$$

$$D_2(\boldsymbol{\theta}_t) = \begin{pmatrix} D_{2.11}(\boldsymbol{\theta}_t) & \cdots & D_{2.1n}(\boldsymbol{\theta}_t) \\ \vdots & \ddots & \vdots \\ D_{2.n1}(\boldsymbol{\theta}_t) & \cdots & D_{2.nn}(\boldsymbol{\theta}_t). \end{pmatrix}$$

For event k at time t, let

$$p_{ik}(\boldsymbol{\theta}_t) = \frac{\exp(\theta_{it}) \sum_{\ell=1}^{m_k-1}(\boldsymbol{X}_k)_{\ell i}}{(\boldsymbol{X}_k \boldsymbol{\eta}_t)' \boldsymbol{1}_{m_k-1}}$$

where $\boldsymbol{\eta}_t = \exp(\boldsymbol{\theta}_t)$. Then

$$D_{1.i}(\boldsymbol{\theta}_t) = -\left(\frac{\theta_{it} - \mu_{it}}{\sigma_{it}^2}\right) + \sum_{k=1}^{K_t} \sum_{\ell=1}^{m_k} (\boldsymbol{W}_k)_{\ell i} - \sum_{k=1}^{K_t} p_{ik}(\boldsymbol{\theta}_t) \tag{23}$$

$$D_{2.ii}(\boldsymbol{\theta}_t) = \frac{-1}{\sigma_{it}^2} - \sum_{k=1}^{K_t} p_{ik}(\boldsymbol{\theta}_t)(1 - p_{ik}(\boldsymbol{\theta}_t)) \tag{24}$$

$$D_{2.ih}(\boldsymbol{\theta}_t) = \sum_{k=1}^{K_t} p_{ik}(\boldsymbol{\theta}_t) p_{hk}(\boldsymbol{\theta}_t) \tag{25}$$

The Newton-Raphson algorithm proceeds in the following manner.

1. Select starting vector of posterior means, $\boldsymbol{\mu}_t^{*0} = (\mu_{1t}^{*0}, \ldots, \mu_{nt}^{*0})$. We have found that a good choice is to perform the following sequence of calculations.

   (a) Calculate $\pi_{it}^{*0} = \frac{\sum_{k=1}^{K_t}(\sum_{\ell=1}^{m_k-1}(\boldsymbol{X}_k)_{\ell i})}{\sum_{k=1}^{K_t}(m_k-1)}$, the proportion of the times competitor i is outperformed by his/her opponents during period t. Note that $1 - \pi_{it}^{*0}$ is therefore the proportion of times competitor i outperforms his/her opponents.

   (b) Let F be the cumulative distribution function (cdf) for a standard logistic distribution, and F^{-1} the inverse cdf. Let $q_{it}^{*0} = F^{-1}(0.01 + 0.98(1 - \pi_{it}^{*0}))$ be the quantile of the standard logistic distribution evaluated at the outperformance

29

probability scaled to stay between 0.01 and 0.99. The scaling ensures that the quantiles are not infinite if the player always outperforms his/her opponents.

(c) Let $\mu_{it}^{*0} = \frac{q_{it}^{*0} + \mu_{it}/\sigma_{it}^2}{1 + 1/\sigma_{it}^2}$, a weighted average of q_{it}^{*0} with the prior mean μ_{it}.

2. At iteration j, $j = 1, 2, \ldots$, let

$$\boldsymbol{\mu}_t^{*j} = \boldsymbol{\mu}_t^{*j-1} - D_2^{-1}(\boldsymbol{\mu}_t^{*j-1})D_1(\boldsymbol{\mu}_t^{*j-1}). \tag{26}$$

The iteration is repeated until $\boldsymbol{\mu}_t^{*j}$ changes by a negligible amount. The final estimated posterior means and standard deviations at iteration J are given by

$$\boldsymbol{\mu}_t^* = \boldsymbol{\mu}_t^{*J} \tag{27}$$

$$\boldsymbol{\sigma}_t^* = \sqrt{\mathrm{diag}(D_2^{-1}(\boldsymbol{\mu}_t^{*J}))}. \tag{28}$$

References

Ali, M. M. (1998): "Probability models on horse-race outcomes," *Journal of Applied Statistics*, 25, 221–229.

Allison, P. D. and N. A. Christakis (1994): "Logit models for sets of ranked items," *Sociological methodology*, 24, 199–228.

Baker, R. D. and I. G. McHale (2014): "Deterministic evolution of strength in multiple comparisons models: Who is the greatest golfer?" *Scandinavian Journal of Statistics*, n/a–n/a.

Bockenholt, U. (1992): "Thurstonian representation for partial ranking data," *British Journal of Mathematical and Statistical Psychology*, 45, 31–49.

Bockenholt, U. (1993): "Applications of thurstonian models to ranking data," in *Probability models and statistical analyses for ranking data*, Springer, 157–172.

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

Breslow, N. and J. Crowley (1974): "A large sample study of the life table and product limit estimates under random censorship," *The Annals of Statistics*, 2, 437–453.

Cargnoni, C., P. Muller, and M. West (1997): "Bayesian forecasting of multinomial time series through conditionally gaussian dynamic models," *Journal of the American Statistical Association*, 92, 640–647.

Carlin, B. P. and S. Chib (1995): "Bayesian model choice via markov chain monte carlo methods," *Journal of the Royal Statistical Society. Series B (Methodological)*, 473–484.

Carlin, B. P., N. G. Polson, and D. S. Stoffer (1992): "A monte carlo approach to nonnormal and nonlinear state-space modeling," *Journal of the American Statistical Association*, 87, 493–500.

Caron, F. and Y. W. Teh (2012): "Bayesian nonparametric models for ranked data," in *Advances in Neural Information Processing Systems*, 1520–1528.

Cattelan, M. (2012): "Models for paired comparison data: A review with emphasis on dependent data," *Statistical Science*, 27, 412–433.

Cowles, M. K. and B. P. Carlin (1996): "Markov chain monte carlo convergence diagnostics: a comparative review," *Journal of the American Statistical Association*, 91, 883–904.

Doucet, A., N. De Freitas, and N. Gordon (2001): *Sequential Monte Carlo methods in practice*, Springer.

Doucet, A., S. Godsill, and C. Andrieu (2000): "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and computing*, 10, 197–208.

Ferreira, M. A. and D. Gamerman (2000): "Dynamic generalized linear models," *BIOSTATISTICS-BASEL-5*, 57–72.

Gamerman, D. and H. S. Migon (1993): "Dynamic hierarchical models," *Journal of the Royal Statistical Society. Series B (Methodological)*, 629–642.

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.

Glickman, M. E. and H. S. Stern (1998): "A state-space model for national football league scores," *Journal of the American Statistical Association*, 93, 25–35.

Graves, T., C. S. Reese, and M. Fitzgerald (2003): "Hierarchical models for permutations: Analysis of auto racing results," *Journal of the American Statistical Association*, 98, 282–291.

Guiver, J. and E. Snelson (2009): "Bayesian inference for plackett-luce ranking models," in *Proceedings of the 26th annual international conference on machine learning*, ACM, 377–384.

Hausman, J. A. and P. A. Ruud (1987): "Specifying and testing econometric models for rank-ordered data," *Journal of Econometrics*, 34, 83–104.

Henery, R. J. (1981): "Permutation probabilities as models for horse races," *Journal of the Royal Statistical Society. Series B (Methodological)*, 86–91.

Henery, R. J. (1983): "Permutation probabilities for gamma random variables," *Journal of applied probability*, 822–834.

Herbrich, R., T. Minka, and T. Graepel (2007): "TrueSkill: A bayesian skill rating system," in *Advances in Neural Information Processing Systems*, 569–576.

Jasra, A., C. C. Holmes, and D. A. Stephens (2005): "Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling," *Statistical Science*, 50–67.

Johnson, V. E., R. O. Deaner, and C. P. Van Schaik (2002): "Bayesian analysis of rank data with application to primate intelligence experiments," *Journal of the American Statistical Association*, 97, 8–17.

Kalbfleisch, J. D. and R. L. Prentice (2011): *The statistical analysis of failure time data*, John Wiley & Sons.

Lo, V. S. and J. Bacon-Shone (1994): "A comparison between two models for predicting ordering probabilities in multiple-entry competitions," *The Statistician*, 317–327.

Luce, R. D. (1959): *Individual Choice Behavior a Theoretical Analysis*, john Wiley and Sons.

Minka, T. P. (2001): *A family of algorithms for approximate Bayesian inference*, Ph.D. thesis, Massachusetts Institute of Technology.

Mosteller, F. (1951): "Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations," *Psychometrika*, 16, 3–9.

Nelder, J. A. and R. Mead (1965): "A simplex method for function minimization," *The computer journal*, 7, 308–313.

Plackett, R. L. (1975): "The analysis of permutations," *Applied Statistics*, 193–202.

Plummer, M. (2003): "JAGS: A program for analysis of bayesian graphical models using gibbs sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March*, 20–22.

R Foundation for Statistical Computing (2012): "R: A language and environment for statistical computing," *Vienna, Austria: R Foundation for Statistical Computing*.

Rauch, H. E., C. T. Striebel, and F. Tung (1965): "Maximum likelihood estimates of linear dynamic systems," *AIAA journal*, 3, 1445–1450.

Spearman, C. (1904): "The proof and measurement of association between two things," *The American journal of psychology*, 15, 72–101.

Stern, H. (1990): "Models for distributions on permutations," *Journal of the American Statistical Association*, 85, 558–564.

Taylor, J. M. (1987): "Kendall's and spearman's correlation coefficients in the presence of a blocking variable," *Biometrics*, 409–416.

Taylor, W. J. (1945): "Method of lagrangian curvilinear interpolation," *Journal of Research of the National Bureau of Standards*, 35, 151–155.

Weng, R. C. and C.-J. Lin (2011): "A bayesian approximation method for online ranking," *The Journal of Machine Learning Research*, 12, 267–300.

West, M., P. J. Harrison, and H. S. Migon (1985): "Dynamic generalized linear models and bayesian forecasting," *Journal of the American Statistical Association*, 80, 73–83.

Zar, J. H. (1972): "Significance testing of the spearman rank correlation coefficient," *Journal of the American Statistical Association*, 67, 578–580.