Paired Comparison Models with Time-Varying Parameters

Mark E. Glickman

Department of Statistics
Harvard University

Paired Comparison Models with Time-Varying Parameters

A thesis presented

by

Mark E. Glickman

to

The Department of Statistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of

Statistics

Harvard University
Cambridge, Massachusetts
May, 1993

Abstract

Paired comparison experiments involve comparisons among a set of objects, treatments, or competitors in blocks of size two to determine relative merits among the objects. Applications to tournament chess competition or team sports differ from the usual paired comparison framework in that competitors' abilities may change over time. A dynamic model is developed for the usual paired comparison experiment in which only a binary indicator of the preferred object is available (or trinary, if ties are allowed), and for alternative paired comparison experiments in which more information (perhaps the score) is available. Posterior distributions of the rating parameters and posterior prediction intervals for the results of future comparisons are obtained using methods that are closely related to the Kalman filter. The methodology is applied to the analysis of professional football game outcomes and to chess tournament results from the 1988–1989 World Cup.

Contents

Acknowledgements

When I first began playing chess in tournaments at the age of eleven, I became quickly fascinated with chess ratings. I remember spending much of my free time wondering how the rating system worked, and occasionally calculating hypothetical updates to my own rating. This interest in chess ratings lead me to the field of statistics, and I decided to major in statistics in college and subsequently enter a Ph. D. program in statistics. It only seemed natural that my doctoral thesis would involve developing models for analyzing chess tournament data. I am very grateful to the Harvard Statistics department for having allowed me the opportunity to write a scholarly work on a topic I have been interested in even before I was a teenager.

This thesis would have been merely a jumble of words and equations if it weren't for my advisor, Hal Stern. Hal, the Bruce Springsteen of statistics, gave my thesis the breath of life. He was particularly important in helping me avoid a natural tendency to be ultra-critical of my own work. Hal also was extremely helpful in getting me to think about the "big picture," but without forsaking details. I am quite grateful for the time Hal made available to me, often at a moment's notice. Among his many contributions to my thesis, Hal provided me with the NFL data analyzed in Chapter 3.

I owe thanks to quite a few others who have helped with my research both academically and financially. I would like to thank both Art Dempster and Herman Chernoff for their financial support throughout my years at Harvard. Art, my second reader, was particularly helpful in getting me to think about modeling aspects of the thesis. I greatly appreciate the help I received from professors Andrew Gelman, Jun Liu and Alan Zaslavsky, all of whom played important roles in solidifying my thesis, especially in the area of Bayesian computation. Thanks to Chris Chabris, a fellow chess player, who suggested the World Cup chess data set analyzed in Chapter 5. Thanks also to John Hartmann, my apartment-mate, who helped me to write Unix shell scripts that circumvented tiresome

programming problems. I am indebted to Tom Blackwell, Eric Bradlow, Emery Brown, David Hoaglin, Chuanhai Liu, Patricia Meehan and Mark Vangel for many helpful discussions.

I would like to thank the people who have made my experience these last six years enjoyable and rewarding, including my fellow statistics graduate students, past and present, as well as friends outside the department. My experience would not have been complete without help from the current department administrator, Judy Rapping, her predecessor, Claude Poux, and the department secretaries Dale Rinkel and Shelly Weiner. Most importantly, I would like to thank my family; my sister Jill, my brother Andrew, and, of course, my mom and dad, for encouraging me to pursue a doctoral degree, and for being supportive and caring throughout my life.

Chapter 1

Introduction

Accurate estimation of a tournament chess competitor's playing strength is an important goal of tournament chess organizers like the U. S. Chess Federation (USCF), and is also of interest to players, as they would like to have their strength quantified. Suppose the results of tournament chess games over many years are observed. One approach to rating chess players' strengths is to use the results from the most recent competitions only and rank players based on these game outcomes. This may not be the most desirable approach because it ignores potentially useful information from earlier competitions. Another approach is to combine all the data and rank players based on the aggregated data. This seems unsatisfactory as well because previous competition data are given as much weight as current data in ranking players. The best approach seems to be a compromise between these two methods, so that earlier competition data is given some weight in estimating current ability, but not as much as the most recent competition data. Ideally, the degree of compromise should be determined from the data itself. The thesis formalizes this idea, and explores an approach to estimating the abilities of competitors from game outcomes observed over time using dynamic paired comparison models.

Paired comparison models are used to analyze data from experiments in which a set of p objects or competitors are compared in blocks of size 2. Applications of paired comparison models include studies in choice behavior, preference testing, sports, and sensory discrimination. The inferential goals of paired comparison analyses include ranking objects based on the results of comparisons, and predicting the results of future comparisons. Most work on paired comparison modeling assumes the outcome is a binary variable indicating which of two objects is preferred. This thesis examines binary outcome models, but also examines models for comparisons with a continuous outcome measure. Two of the most commonly used paired comparison models are the Bradley-Terry model (Bradley and Terry 1952) and the Thurstone-Mosteller model (Mosteller 1951; Thurstone 1927). David (1988) provides a thorough introduction and an extensive bibliography on topics in paired comparison modeling.

The thesis examines paired comparison models in which the merits or abilities of objects being compared may change over time. Competitive sports and games, including chess, football, hockey, and basketball, involve teams or players whose abilities change over time due to effects such as player trades, coaching staff changes, and aging of players. The Bradley-Terry and Thurstone-Mosteller models assume that paired comparison data occur simultaneously or within a short period of time, or that objects' merits do not change over time, so that blind use of these models is not appropriate. The models developed here assign a rating parameter to each object or competitor which describes the object's strength or merit, and the rating parameter may change over time. Methods are developed for paired comparisons where the outcome variable is either a difference in measurements between the two objects, or where the outcome is a binary indicator of which object is preferred (or trinary, if ties are allowed).

The models developed in this thesis have strong connections to the discrete-time Kalman filter (Kalman 1960; Kalman and Bucy 1961), and to Bayesian dynamic models (West and Harrison 1990). Our models assume an "observation equation," which is the probability model for an outcome of a comparison at time t, and a "system equation" which describes the stochastic evolution of parameters from time t to time $t + 1$. The foundational assumptions for this model are that when objects are compared or players compete, information is obtained about their rating parameters, but between competitions, changes may occur that affect the merits of the objects or the abilities of players, so that the distribution of the rating parameters becomes less certain due to the passage of time without comparisons. This uncertainty is reflected in a greater variance of the posterior distribution of the rating parameters.

The system variance, describing the magnitude of the parameter changes over time, is an unknown parameter, and this poses some difficulties in the analyses of the dynamic paired comparison models. To facilitate the analyses, the posterior distribution of the system variance parameter is approximated by a discrete distribution. Two methods for carrying out the approximation are examined. In one analysis, iterative simulation via the Gibbs sampler (Geman and Geman 1984) is employed to obtain samples from the posterior distribution of parameters of interest. A second approach involves approximating the prior distribution of the system variance by a discrete distribution, and performing a conditional analysis for each value of the system variance. The proposed models in this thesis differ in an important respect from the usual dynamic models of West and Harrison (1990) in the assumption that the system variance is unknown a priori. Previous uses of models like this typically assume the variance is known or can be well estimated in advance. In this respect the models developed in this thesis are less restrictive than the usual models. Placing probability distributions on all rating and variance parameters allows for a flexible approach to modeling paired comparison data.

While the thesis specifically focuses on paired comparison models, the approach to analyzing dynamic models has a wide range of applicability. For example, the methods developed for analyzing paired comparison models with continuous outcomes can be generalized to normal linear dynamic models of West and Harrison (1990), so that these more general models can be utilized without assuming the system variance is known. Furthermore, methodology developed to analyze paired comparison models with indicator outcomes can be applied to a wide class of non-linear dynamic models, including dynamic versions of generalized linear models (McCullagh and Nelder 1989). Thus, the combination of dynamic probability models and Bayesian calculations that are feasible in modern computing environments offers a powerful tool in analyzing experimental time series data.

This paper is organized into two main parts. The first part, which includes Chapters 2 and 3, discusses paired comparison models with a normally distributed outcome variable, while the second part, which includes Chapters 4 and 5, considers paired comparison models with indicator outcomes. In particular, Chapter 2 develops the methodology for paired comparison models where the outcome variable is the difference between measurements or scores. Chapter 3 applies the methodology of Chapter 2 to the analysis of NFL football game results. Chapter 4 examines the Bradley-Terry model and its dynamic extensions, and discusses the methodology to analyze data using such models. In Chapter 5, the methodology of Chapter 4 is applied to analyzing the results of chess tournament games from the 1988–9 World Cup Chess events and some simulated data. Chapter 6 concludes the thesis with a discussion of the utility of the models and possible extensions for more complex paired comparison experiments.

Chapter 2

Paired Comparisons with Measured Outcomes

In this chapter, we consider several models for paired comparison data in which the outcome variable is the difference between measurements of the objects or teams. Our approach assumes that the measurements or scores of each object are from independent normal distributions. The analysis of the normal model is described in Section 2.1, first assuming that the variances from the normal distributions are identical, and then considering unrestricted variances. For the remainder of the development in this chapter, we assume that the variances are identical. We examine the Bayesian analysis of the normal model in Section 2.2 using a conjugate prior distribution. In Section 2.3, we introduce the dynamic model which assumes that the means of the normally distributed measurements can change over time. The magnitude of the changes is reflected by a variance parameter of the dynamic component in the model. We develop the analysis of the dynamic model in Sections 2.4 and 2.5. Because the analysis of the dynamic model is intractable in closed form, we examine two methods to facilitate the analysis. Each method involves an approximation of the posterior distribution for the variance parameter of the dynamic component. One approach uses the Gibbs sampler to draw a random sample from the marginal posterior distribution of the variance parameter. A second approach, incorporating a non-iterative technique, approximates the continuous prior distribution on the variance parameter by a discrete distribution. Extending the model to allow for the inclusion of covariate information is discussed in Section 2.6. We then show in Section 2.7 how to obtain approximate marginal posterior distributions for parameters of interest, and describe how to obtain the forecast distribution. Sequential updating of the posterior distribution of parameters when new data becomes available is discussed in Section 2.8. For developing of the models, we use language that assumes paired comparison data are scores resulting from team competitions, but the models are understood to have more general applicability.

2.1 A Normal Model

Suppose a set of p teams participate in a series of competitions, and let $S_{i_k k}$ be the random variable corresponding to the score of the k-th game produced by team i_k. We consider models appropriate when only $Y_{i_k j_k k} = S_{i_k k} - S_{j_k k}$ can be observed, or when the only relevant information concerning mean team scores is the difference in the scores. This latter situation occurs, for example, in experiments with paired designs. The only information about the treatments that is not confounded with blocks are the differences in measurements within blocks. We assume that the $S_{i_k k}$ come from independent normal distributions. We consider a model that assumes equal variances in Section 2.1.1, and then examine a model with unrestricted variances in Section 2.1.2.

2.1.1 Equal Variances

The normal model with equal variances assumes team i produces scores that are independently distributed

$$(S_i|\theta_i, \tau^2) \sim \mathrm{N}(\theta_i, \tau^2),$$

with possibly different means, but equal variances.

Suppose a total of n comparisons among the p teams are observed. As before, denote by $S_{i_k k}$ the score produced by team i_k on the k-th comparison, $k = 1, \ldots, n$, where i_k can take on values $1, \ldots, p$. We have under the assumptions above, given θ_{i_k}, θ_{j_k} and τ^2,

$$Y_{i_k j_k k} = S_{i_k k} - S_{j_k k} \sim \mathrm{N}(\theta_{i_k} - \theta_{j_k}, 2\tau^2), \tag{2.1}$$

where $Y_{i_k j_k k}$ is the score difference of the k-th competition between teams i_k and j_k. We assume that the result of competitions are independent from each other, so that the $Y_{i_k j_k k}$ are independent random variables.

This model for score differences can be derived from a slightly different set of assumptions. If we assume that for the k-th competition,

$$\begin{aligned} S_{i_k k} &= \theta_{i_k} + \alpha_k + \epsilon_{i_k k} \\ S_{j_k k} &= \theta_{j_k} + \alpha_k + \epsilon_{j_k k}, \end{aligned} \tag{2.2}$$

where $\epsilon_{i_k k}$ and $\epsilon_{j_k k}$ are normally distributed uncorrelated random variables with mean 0 and variance τ^2, then we obtain the model in (2.1) above. We can think of α_k as a nuisance parameter particular to competition k. For instance, the weather conditions for a competition or other external factors may affect both teams' final scores, and it is not unreasonable to assert that the effect will be equal for both teams. The model in (2.2) is common for incomplete block designs with blocks of size 2. The block effects are not of much interest, and the analysis of such models can proceed by taking differences of observations within blocks to remove the block effects.

The model can be represented in matrix form as follows. Let $\boldsymbol{y}$ be the vector of the n score differences, and let $\boldsymbol{\theta}$ be the vector of the p team parameters $(\theta_1, \ldots, \theta_p)^{\mathsf{T}}$. We introduce an $n \times p$ design matrix, $\boldsymbol{X}$. Suppose that the k-th element of $\boldsymbol{y}$ is the score difference between teams i_k and j_k. Then the k-th row of $\boldsymbol{X}$ is given by $\boldsymbol{X}_{ki_k} = 1$, $\boldsymbol{X}_{kj_k} = -1$, and $\boldsymbol{X}_{k\ell} = 0$ for $\ell \neq i_k$ and $\ell \neq j_k$. We can now restate our model in the more compact form

$$(\boldsymbol{y}|\boldsymbol{\theta}, \tau^2) \sim \mathrm{N}(\boldsymbol{X\theta}, 2\tau^2\mathbf{I}_n), \tag{2.3}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

In the next section, we propose a complete probability model for $\boldsymbol{\theta}$ and describe approaches to posterior inference. We begin our discussion of the model by considering maximum likelihood estimation. It is straightforward to compute maximum likelihood estimates (MLE's) for $\boldsymbol{\theta}$ as long as the singularities in the design matrix $\boldsymbol{X}$ are removed. Because the data only supply information on score differences, we can only estimate the θ_i uniquely up to an additive constant. We impose a single constraint to identify $\boldsymbol{\theta}$. One possibility is the linear constraint, $\sum_{i=1}^{p} \theta_i = 0$. We incorporate this into the design matrix by augmenting $\boldsymbol{X}$ with the row $(1, 1, \ldots, 1)$ and $\boldsymbol{y}$ with the extra observation 0 with 0 variance. Singularities will arise if teams can be partitioned into two subsets in which none of the teams in one subset competes against any team in the other subset. In such a situation, team parameters can be estimated within subsets, but not between subsets. No information is provided in the data to simultaneously estimate all of the parameters. This suggests that an important consideration in designing a paired comparison experiment is to prevent such a situation from occurring. Necessary and sufficient conditions are discussed in Ford (1957).

If we call the augmented matrix $\tilde{\boldsymbol{X}}$ and the augmented score difference vector $\tilde{\boldsymbol{y}}$, then the maximum likelihood estimates of $\boldsymbol{\theta}$ and τ^2 are given by the ordinary linear model estimates

$$\hat{\boldsymbol{\theta}} = (\tilde{\boldsymbol{X}}^{\mathsf{T}}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^{\mathsf{T}}\tilde{\boldsymbol{y}} \tag{2.4}$$

$$\hat{\tau}^2 = \frac{1}{2n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}). \tag{2.5}$$

Estimated variances and covariances of the parameter estimates are given by

$$\widehat{\mathrm{Var}(\hat{\boldsymbol{\theta}})} = 2\hat{\tau}^2(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}. \tag{2.6}$$

Extensions to include linear covariates are straightforward. Covariate information can be appended as columns to the design matrix $\boldsymbol{X}$, and the formulas in (2.4), (2.5), and (2.6) still hold, as long as $\boldsymbol{X}$ remains nonsingular.

2.1.2 Unrestricted Variances

We now explore a model where team scores are not necessarily homoskedastic. Let

$$S_{i_k k} \sim \mathrm{N}(\theta_{i_k}, \tau_{i_k}^2),$$

so that

$$Y_{i_k j_k k} = S_{i_k k} - S_{j_k k} \sim \mathrm{N}(\theta_{i_k} - \theta_{j_k}, \tau^2_{i_k} + \tau^2_{j_k}).$$

The likelihood for this model can be written as

$$\mathrm{L}_{\mathrm{obs}} \propto \prod_k \frac{1}{\sqrt{\tau^2_{i_k} + \tau^2_{j_k}}} \exp\left(-\frac{(y_{i_k j_k k} - \theta_{i_k} + \theta_{j_k})^2}{2(\tau^2_{i_k} + \tau^2_{j_k})} \right),$$

where $\mathrm{L}_{\mathrm{obs}}$ denotes the likelihood given the observed data.

Maximum likelihood estimates can be found using the EM algorithm (Dempster, Laird and Rubin 1977). The Newton-Raphson algorithm may also be used, but the ease of implementation of the EM algorithm encourages its use in this case. We now briefly describe the procedure for obtaining maximum likelihood estimates using EM.

We note that the observed data are the score differences, $y_{i_k j_k k}$, and that the score of team i_k from game k, which we denote $s_{i_k k}$, can be thought of as missing data. Therefore the result of the k-th comparison is a score $s_{i_k k}$ for team i_k, and a score $s_{i_k k} - y_{i_k j_k k}$ for team j_k, with $s_{i_k k}$ unobserved. To carry out the M-step of the EM algorithm, notice that the complete-data likelihood can be expressed as the product of $2n$ independent normal densities, that is,

$$\mathrm{L}_{\mathrm{com}} \propto \prod_k \left\{ \frac{1}{\tau_{i_k}} \exp\left(-\frac{1}{2\tau^2_{i_k}} (s_{i_k k} - \theta_{i_k})^2 \right) \right\} \cdot \left\{ \frac{1}{\tau_{j_k}} \exp\left(-\frac{1}{2\tau^2_{j_k}} ((s_{i_k k} - y_{i_k j_k k}) - \theta_{j_k})^2 \right) \right\}.$$

The maximum likelihood estimates of the θ_i and the τ^2_i conditional on the missing data are easy to compute. The maximum likelihood estimates of the θ_i are the sample means of the complete data observations that correspond to θ_i in the likelihood, and the maximum likelihood estimates of the τ^2_i are the sample variances of the complete data.

The E-step of the EM algorithm requires finding expressions for the expected sufficient statistics, which in this case is equivalent to finding

$$\mathrm{E}(s_{i_k k} | \theta_1, \ldots, \tau^2_1, \ldots, \boldsymbol{y})$$

and

$$\mathrm{E}(s^2_{i_k k} | \theta_1, \ldots, \tau^2_1, \ldots, \boldsymbol{y}).$$

We can derive

$$\mathrm{E}(s^2_{i_k k} | \theta_1, \ldots, \tau^2_1, \ldots, \boldsymbol{y})$$

by first obtaining

$$\mathrm{Var}(s_{i_k k} | \theta_1, \ldots, \tau^2_1, \ldots, \boldsymbol{y}).$$

These expressions can be shown to be equal to

$$\mathrm{E}(s_{i_k k} | \theta_1, \ldots, \theta_p, \tau^2_1, \ldots, \tau^2_p, \boldsymbol{y}) = \frac{\theta_{i_k}/\tau^2_{i_k} + (\theta_{j_k} + y_{i_k j_k k})/\tau^2_{j_k}}{1/\tau^2_{i_k} + 1/\tau^2_{j_k}}$$

$$\begin{aligned}\mathrm{Var}(s_{i_k k}|\theta_1,\ldots,\theta_p,\tau_1^2,\ldots,\tau_p^2,\boldsymbol{y}) &= \frac{1}{1/\tau_{i_k}^2+1/\tau_{j_k}^2}\\ \Longrightarrow \mathrm{E}(s_{i_k k}^2|\theta_1,\ldots,\theta_p,\tau_1^2,\ldots,\tau_p^2,\boldsymbol{y}) &= \frac{1}{1/\tau_{i_k}^2+1/\tau_{j_k}^2}+\left(\frac{\theta_{i_k}/\tau_{i_k}^2+(\theta_{j_k}+y_{i_k j_k k})/\tau_{j_k}^2}{1/\tau_{i_k}^2+1/\tau_{j_k}^2}\right)^2.\end{aligned}$$

The EM algorithm proceeds as follows. Pick a set of p starting values for the θ_i and for the τ_i^2. Denote these $\theta_1^{(1)},\ldots,\theta_p^{(1)}$ and $\tau_1^{2\,(1)},\ldots,\tau_p^{2\,(1)}$. At the t-th iteration, the E-step requires

$$\begin{aligned}\mathrm{E}^{(t+1)}(s_{i_k k}) &\leftarrow \frac{\theta_{i_k}^{(t)}/\tau_{i_k}^{2\,(t)}+(\theta_{j_k}^{(t)}+y_{i_k j_k k})/\tau_{j_k}^{2\,(t)}}{1/\tau_{i_k}^{2\,(t)}+1/\tau_{j_k}^{2\,(t)}}\\ \mathrm{E}^{(t+1)}(s_{i_k k}^2) &\leftarrow \frac{1}{1/\tau_{i_k}^{2\,(t)}+1/\tau_{j_k}^{2\,(t)}}+(\mathrm{E}^{(t+1)}(s_{i_k k}))^2,\end{aligned}$$

which are the expected values and expected squared values of the missing data given the parameter estimates after the t-th iteration. The M-step then replaces the current parameter estimates with the maximum likelihood estimates from the "complete" data:

$$\begin{aligned}\theta_i^{(t+1)} &\leftarrow \text{Sample mean of expected data involving team } i\\ \tau_i^{2\,(t+1)} &\leftarrow \text{Sample mean of expected squared data involving team } i\text{, minus } (\theta_i^{(t+1)})^2.\end{aligned}$$

In the M-step, we choose to force the θ_i to have mean 0 by subtracting off the mean of the θ_i at every iteration. This guarantees that the θ_i will be centered around 0. Otherwise, the EM algorithm will produce estimates that may be translated from the estimates obtained by centering around 0, and the translation will depend on the starting values.

This iteration is continued until convergence. Depending on the size of the problem, the algorithm may converge slowly because the problem assumes that one-half of the complete data is missing. Asymptotic variances and covariances of the the parameters may be found using SEM (Meng and Rubin 1991) in conjunction with the EM algorithm. An alternative would be to use gradient methods which may converge more quickly, but require specification of the Hessian matrix which may be cumbersome.

2.2 A Bayesian Formulation

A complete Bayesian formulation of the normal model incorporates prior distributions on team abilities, as well as on τ^2. For the following discussion, we now let τ^2 be the variance of the score difference, and let $\phi = 1/\tau^2$ be the precision of score differences. We assume a normal-gamma prior distribution on $(\boldsymbol{\theta},\phi)$ with density

$$f(\boldsymbol{\theta},\phi)\propto\phi^{p/2}\exp\{-\frac{\phi}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{R}(\boldsymbol{\theta}-\boldsymbol{\mu})\}\cdot\phi^{v/2-1}\exp(-\frac{1}{2}\phi v\xi)$$

which we write as

$$(\boldsymbol{\theta}, \phi) \sim \mathrm{NG}(\boldsymbol{\mu}, \xi, \boldsymbol{R}, v).$$

This representation is used by Raiffa and Schlaifer (1961, Section 13.5). The normal-gamma distribution, besides accommodating a wide range of prior beliefs, is the conjugate distribution for normal regression models. The distribution acquires its name from the factorization

$$\begin{aligned} \phi &\sim \mathrm{Gamma}(v/2, v\xi/2) \\ (\boldsymbol{\theta}|\phi) &\sim \mathrm{N}(\boldsymbol{\mu}, (\phi\boldsymbol{R})^{-1}). \end{aligned}$$

In the normal-gamma distribution, the parameter $\boldsymbol{\mu}$ is the mean of $\boldsymbol{\theta}$, the matrix $\boldsymbol{R}$ is the information matrix for $\boldsymbol{\theta}$ in units of ϕ, the parameter ξ is the harmonic mean of τ^2, and v is equivalent to the number of degrees of freedom associated with ϕ.

Suppose we observe n score differences, $\boldsymbol{y}$, along with the $n \times p$ design matrix $\boldsymbol{X}$. We assume as in (2.3), suppressing the conditioning on $\boldsymbol{X}$,

$$(\boldsymbol{y}|\theta, \phi) \sim \mathrm{N}(\boldsymbol{X\theta}, 1/\phi).$$

Then following Raiffa and Schlaifer (1961), the posterior distribution for $(\boldsymbol{\theta}, \phi)$ is

$$(\boldsymbol{\theta}, \phi|\boldsymbol{y}) \sim \mathrm{NG}(\boldsymbol{\mu}', \xi', \boldsymbol{R}', v')$$

where

$$\begin{aligned} \boldsymbol{R}' &= \boldsymbol{R} + \boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} \\ \boldsymbol{\mu}' &= \boldsymbol{R}'^{\,-1}(\boldsymbol{R\mu} + \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}) \\ v' &= v + n \\ \xi' &= \frac{1}{v'}[(v\xi + \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{R\mu}) + \boldsymbol{y}^{\mathsf{T}}\boldsymbol{y} - \boldsymbol{\mu}'^{\mathsf{T}}\boldsymbol{R}'\boldsymbol{\mu}'] \end{aligned} \tag{2.7}$$

Raiffa and Schlaifer (1961) also show that the posterior distribution of $\boldsymbol{\theta}$ marginalized over ϕ is a multivariate t-distribution on v' degrees of freedom,

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \propto |\boldsymbol{R}|^{1/2}(v' + (\boldsymbol{\theta} - \boldsymbol{\mu}')^{\mathsf{T}}\boldsymbol{R}(\boldsymbol{\theta} - \boldsymbol{\mu}')/\xi')^{(v'+p)/2}$$

with mean vector $\boldsymbol{\mu}'$ and scale matrix $\xi'\boldsymbol{R}'^{\,-1}$. Inference on $\boldsymbol{\theta}$ can be obtained from the multivariate t marginal posterior.

Typically, information about teams' abilities is available prior to competition, and this information can be incorporated into the prior distribution of the parameters. If no prior information exists, a multivariate normal prior with large variances can reflect this lack of information, and the analysis proceeds without difficulty. Note that in the Bayesian model it is not necessary to add a constraint on the θ_i in order to obtain a unique posterior mean $\boldsymbol{\mu}'$. However, it is necessary to require the prior parameter $\boldsymbol{R}$ to be of full rank so that $\boldsymbol{R}'$ is invertible.

The inclusion of linear covariate information is accompanied by a trivial extension to the Bayesian model. We place a normal-gamma prior on the regression parameters as before, and (2.7) again describes the posterior distribution.

2.3 A Dynamic Model

In many paired comparison situations, teams will compete against each other over time, and during this time it is reasonable to believe that team abilities may change. For example, between seasons of professional sports, team abilities may change dramatically as a result of player trades, training, changes in coaching staff, and so on. We want to reflect this possibility in modeling the paired comparison data.

The Bayesian model of the previous section is not appropriate for paired comparisons where team abilities may change over time. It would not be reasonable, for example, to continually update the posterior distribution using (2.7) as new data is observed, because such an analysis treats the observations as time-exchangeable. Instead, an appropriate model should give greater importance to more recent results.

We define a tournament (or season) as a collection of paired comparisons that are made simultaneously, or close enough in time to treat the observations as exchangeable. For simplicity, assume that occurrences of tournaments are separated by equal amounts of time. We use superscript (t) to index tournaments, $t = 1, 2, \ldots, T$. Let $\boldsymbol{y}^{(t)}$ be the vector of $n^{(t)}$ score differences resulting from tournament t, and let $\boldsymbol{X}^{(t)}$ be the design matrix, or tournament schedule, associated with tournament t, as defined in Section 2.1. Also, let $\boldsymbol{\theta}^{(t)}$ be the vector of team parameters associated with tournament t. The model for observations is

$$(\boldsymbol{y}^{(t)} | \boldsymbol{X}^{(t)}, \boldsymbol{\theta}^{(t)}, \tau^2) \sim \mathrm{N}(\boldsymbol{X}^{(t)}\boldsymbol{\theta}^{(t)}, \tau^2 \mathbf{I}_{n^{(t)}}). \tag{2.8}$$

Notice that τ^2 does not depend on time. We now introduce a probability model that describes the evolution of the parameter $\boldsymbol{\theta}$,

$$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} + \boldsymbol{\nu}^{(t)}, \tag{2.9}$$

where $\boldsymbol{\nu}^{(t)}$ is the amount by which team abilities change between tournaments $t-1$ and t. The parameter $\boldsymbol{\nu}^{(t)}$ is assumed to be stochastically independent of $\boldsymbol{\theta}^{(t-1)}$. We further assume

$$(\boldsymbol{\nu}^{(t)} | \sigma^2) \sim \mathrm{N}(\boldsymbol{\alpha}_t, \sigma^2 \mathbf{I}_p), \tag{2.10}$$

where the vector parameter $\boldsymbol{\alpha}_t$ is the mean amount by which teams change between tournaments $t-1$ and t, and σ^2 is the between-season variance for a team's ability. The parameter $\boldsymbol{\alpha}_t$ may be a function of covariate information.

We specify an initial prior distribution

$$(\boldsymbol{\theta}^{(1)}, \phi) \sim \mathrm{NG}(\boldsymbol{\mu}^{(1)}, \xi^{(1)}, \boldsymbol{R}^{(1)}, v^{(1)}), \tag{2.11}$$

where $\boldsymbol{\mu}^{(1)}$, $\xi^{(1)}$, $\boldsymbol{R}^{(1)}$,and $v^{(1)}$ are the hyperparameters for the prior distribution. A vague prior on $\boldsymbol{\theta}^{(1)}$ is obtained by choosing parameters $\boldsymbol{R}^{(1)}$ and $\xi^{(1)}$ such that the diagonal elements of $\mathrm{Var}(\boldsymbol{\theta}^{(1)}) = \frac{v^{(1)}}{v^{(1)}-2}\xi^{(1)}(\boldsymbol{R}^{(1)})^{-1}$ are large. A prior mean of ϕ is set by choosing an appropriate value of $\xi^{(1)}$, and the uncertainty in ϕ is captured by the value of $v^{(1)}$.

Lower values of $v^{(1)}$ connote less certainty about ϕ. The off-diagonal elements of $\boldsymbol{R}^{(1)}$ describe the prior correlation believed to exist between the ratings of teams. If no known similarities between teams exist a priori, we set the covariances of teams ratings, and therefore the off-diagonal elements of $\boldsymbol{R}^{(1)}$, to 0.

We also assume a prior distribution on $\omega = 1/\sigma^2$ having density

$$f(\omega) \propto \omega^{a_0 - 1} e^{-b_0 \omega}, \tag{2.12}$$

which, for $a_0 > 0$, $b_0 > 0$, gives the Gamma density. Proper inferences are possible with other values of a_0 and b_0, however, and even an improper prior distribution with $b_0 = 0$ may be a sensible choice. Without much prior knowledge concerning ω, we may choose a_0 small to reflect the initial uncertainty of ω. For proper prior distributions, the value of b_0 can be chosen so that b_0/a_0 is an initial guess (the prior harmonic mean) of σ^2.

West and Harrison (1990) define a normal dynamic linear model as consisting of the sequence of quadruples $H = \{F^{(t)}, G^{(t)}, V^{(t)}, W^{(t)}\}$ where

$$\begin{aligned} (\boldsymbol{y}^{(t)} | \boldsymbol{\theta}^{(t)}, H) &\sim \mathrm{N}(F^{(t)}\boldsymbol{\theta}^{(t)}, V^{(T)}), \\ (\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}^{(t-1)}, H) &\sim \mathrm{N}(G^{(t)}\boldsymbol{\theta}^{(t-1)}, W^{(t)}). \end{aligned}$$

The model defined in (2.8)–(2.12) is an example of a dynamic linear model with $F^{(t)} = \boldsymbol{X}^{(t)}$, $G^{(t)} = \mathbf{I}_p$, $V^{(t)} = \tau^2 \mathbf{I}_{n(t)}$, and $W^{(t)} = \sigma^2 \mathbf{I}_p$. Note that in our model τ^2 is unknown, so we impose a joint prior on $(\boldsymbol{\theta}^{(t)}, \phi = 1/\tau^2)$ to account for our uncertainty.

Equation (2.8) is the so-called "observation equation" of the dynamic model, and defines the sampling distribution of the score differences. Equation (2.9), along with the distributional specification of (2.10) is often called the "system equation" or "state equation" of the dynamic model. The assumption underlying these equations is that team's abilities change on average by $\boldsymbol{\alpha}_t$ from tournament $t-1$ to t. The parameter $\boldsymbol{\alpha}_t$ may be known a priori, or it may involve estimable parameters. We assume $\boldsymbol{\alpha}_t$ may depend on characteristics of the teams that relate to the evolution of their abilities over time, such as age of the players on the team. However, we assume for purposes of developing our model that $\boldsymbol{\alpha}_t = 0$ for all t, meaning that teams' abilities are expected to remain the same relative to other teams. This assumption defines a random walk model for the parameter $\boldsymbol{\theta}$.

2.4 Analysis with σ^2 known

This section illustrates the analysis of the model when σ^2 is assumed to be known. It is often unrealistic to make this assumption in practice, although much of the literature on dynamic models treats σ^2 as either known (see, for example, Meinhold and Singpurwalla 1983), or well estimated, so that the analysis of the model can be performed conditional on the estimate. We develop here the methodology for analyzing data under the dynamic model with known σ^2, and apply these methods in Section 2.5 where we assume σ^2 is an unknown parameter.

Let

$$\begin{aligned} d_t &= \{\text{Observed data for season } t\} \\ D_t &= (d_1, d_2, \ldots, d_t) \\ &= \{\text{Observed data through season } t\} \end{aligned}$$

By convention, we let D_0 be the state of the dynamic system before data is observed.

For this section, we are primarily interested in obtaining the distribution of $(\boldsymbol{\theta}^{(T)}|D_T)$, the posterior distribution of $\boldsymbol{\theta}^{(T)}$ marginalized over all other parameters, and $(d_{T+1}|D_T)$, the forecast distribution of score differences. The posterior distribution $(\boldsymbol{\theta}^{(T)}|D_T)$ may be of interest in ranking teams after a season, making use of previous seasons' results. The forecast distribution for next season's results may be used to find

$$\Pr(y_{ij}^{(T+1)} > 0|D_T) = \Pr(i \text{ defeats } j \text{ in season } (T+1) \text{ given available data}).$$

In this section, we obtain these results conditional on σ^2.

The full posterior distribution given σ^2 can be written as

$$\begin{aligned} & f(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi|\sigma^2, D_T) \\ & \quad \propto \{f(\phi|\sigma^2, D_0)\} \times \{f(\boldsymbol{\theta}^{(1)}|\phi, \sigma^2, D_0)\} \\ & \qquad \times \{f(d_1|\boldsymbol{\theta}^{(1)}, \phi, \sigma^2, D_0)\} \\ & \qquad \times \prod_{t=2}^{T} \Big\{ \{f(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)}, \ldots, \boldsymbol{\theta}^{(1)}, \phi, \sigma^2, D_{t-1})\} \\ & \qquad\qquad \times \{f(d_t|\boldsymbol{\theta}^{(t)}, \ldots, \boldsymbol{\theta}^{(1)}, \phi, \sigma^2, D_{t-1})\} \Big\} \end{aligned}$$

or equivalently under the probability model as

$$\begin{aligned} & f(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi|\sigma^2, D_T) \\ & \quad \propto \{\exp(-v^{(1)}\xi^{(1)}\phi/2)\, \phi^{v^{(1)}/2-1}\} \\ & \qquad \times \{\exp(-\frac{\phi}{2}(\boldsymbol{\theta}^{(1)} - \boldsymbol{\mu}^{(1)})^{\mathsf{T}} \boldsymbol{R}^{(1)} (\boldsymbol{\theta}^{(1)} - \boldsymbol{\mu}^{(1)}))\} \\ & \qquad \times \{\phi^{n^{(1)}/2} \exp(-\frac{\phi}{2}(\boldsymbol{y}^{(1)} - \boldsymbol{X}^{(1)}\boldsymbol{\theta}^{(1)})^{\mathsf{T}}(\boldsymbol{y}^{(1)} - \boldsymbol{X}^{(1)}\boldsymbol{\theta}^{(1)}))\} \\ & \qquad \times \prod_{t=2}^{T} \Big\{ \{\exp(-\frac{1}{2\sigma^2}(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)})^{\mathsf{T}}(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}))\} \\ & \qquad\qquad \times \{\phi^{n^{(t)}/2} \exp(-\frac{\phi}{2}(\boldsymbol{y}^{(t)} - \boldsymbol{X}^{(t)}\boldsymbol{\theta}^{(t)})^{\mathsf{T}}(\boldsymbol{y}^{(t)} - \boldsymbol{X}^{(t)}\boldsymbol{\theta}^{(t)}))\} \Big\} \qquad (2.13) \end{aligned}$$

We are primarily interested in obtaining the distribution of $(\boldsymbol{\theta}^{(T)}, \phi|\sigma^2, D_T)$, so we integrate the full posterior distribution in (2.13) over the parameters $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T-1)}$. This is accomplished through a recursive updating and forecasting operation. An updating operation is the Bayesian calculation that obtains the distribution of $(\boldsymbol{\theta}^{(t)}, \phi|\sigma^2, D_t)$ from $(\boldsymbol{\theta}^{(t)}, \phi|\sigma^2, D_{t-1})$ after acquiring data at time t. A forecasting operation is the Bayesian calculation that obtains the distribution of $(\boldsymbol{\theta}^{(t+1)}, \phi|\sigma^2, D_t)$ from the distribution $(\boldsymbol{\theta}^{(t)}, \phi|\sigma^2, D_t)$, reflecting the additional uncertainty due to the passage of time. Repeated application of updating and forecasting operations results in the distribution of $(\boldsymbol{\theta}^{(T)}, \phi|\sigma^2, D_T)$ from that of $(\boldsymbol{\theta}^{(1)}, \phi|\sigma^2, D_0)$.

2.4.1 Updating and Forecasting Recursions

We demonstrate the updating and forecasting recursions by considering the state of the dynamic system prior to tournament t. The prior distribution on $(\boldsymbol{\theta}^{(t)}, \phi|\sigma^2, D_{t-1})$ is

$$(\boldsymbol{\theta}^{(t)}, \phi|\sigma^2, D_{t-1}) \sim \mathrm{NG}(\boldsymbol{\mu}^{(t)}, \xi^{(t)}, \boldsymbol{R}^{(t)}, v^{(t)})$$

as in (2.11). We now observe d_t, consisting of the $n^{(t)}$ outcomes of paired comparisons of tournament t, and by the methodology of Section 2.2 we obtain

$$(\boldsymbol{\theta}^{(t)}, \phi|\sigma^2, D_t) \sim \mathrm{NG}(\boldsymbol{\mu}'^{(t)}, \xi'^{(t)}, \boldsymbol{R}'^{(t)}, v'^{(t)}) \tag{2.14}$$

where

$$\begin{aligned}
\boldsymbol{R}'^{(t)} &= \boldsymbol{R}^{(t)} + \boldsymbol{X}^{(t)\mathsf{T}}\boldsymbol{X}^{(t)} \\
\boldsymbol{\mu}'^{(t)} &= \boldsymbol{R}'^{(t)\,-1}(\boldsymbol{R}^{(t)}\boldsymbol{\mu}^{(t)} + \boldsymbol{X}^{(t)\mathsf{T}}\boldsymbol{y}^{(t)}) \\
v'^{(t)} &= v^{(t)} + n^{(t)} \\
\xi'^{(t)} &= \frac{1}{v'^{(t)}}[(v^{(t)}\xi^{(t)} + \boldsymbol{\mu}^{(t)\mathsf{T}}\boldsymbol{R}^{(t)}\boldsymbol{\mu}^{(t)}) + \boldsymbol{y}^{(t)\mathsf{T}}\boldsymbol{y}^{(t)} - \boldsymbol{\mu}'^{(t)\mathsf{T}}\boldsymbol{R}'^{(t)}\boldsymbol{\mu}'^{(t)}]
\end{aligned} \tag{2.15}$$

This completes an updating calculation.

To demonstrate the forecasting calculation, we assume the distribution in (2.14) has been obtained. The conditional distribution of $\boldsymbol{\theta}^{(t+1)}$ is specified as

$$(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}, \phi, \sigma^2, D_t) \sim \mathrm{N}(\boldsymbol{\theta}^{(t)}, \sigma^2\mathbf{I}_p),$$

which can be interpreted as incorporating the additional uncertainty in the team's abilities at time $t+1$. The joint distribution of parameters before observing paired comparison data at time $t+1$ is

$$\begin{aligned}
f(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}, \phi|\sigma^2, D_t) &= f(\boldsymbol{\theta}^{(t)}, \phi|\sigma^2, D_t) \times f(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}, \phi, \sigma^2, D_t) \\
&\propto \exp\{-\frac{\phi}{2}(\boldsymbol{\theta}^{(t)} - \boldsymbol{\mu}'^{(t)})^{\mathsf{T}}\boldsymbol{R}'^{(t)}(\boldsymbol{\theta}^{(t)} - \boldsymbol{\mu}'^{(t)})\}\phi^{p/2} \\
&\quad \times \exp(-\frac{1}{2}\phi v'^{(t)}\xi'^{(t)})\phi^{v'^{(t)}/2-1} \\
&\quad \times \exp(-\frac{1}{2\sigma^2}(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)})^{\mathsf{T}}(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)})).
\end{aligned}$$

Marginalizing over $\boldsymbol{\theta}^{(t)}$ yields

$$\begin{aligned} f(\boldsymbol{\theta}^{(t+1)}, \phi|\sigma^2, D_t) &= \int f(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}, \phi|\sigma^2, D_t)\, d\boldsymbol{\theta}^{(t)} \\ &\propto \exp\{-\frac{\phi}{2}(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\mu}^{(t+1)})^{\mathsf{T}}\boldsymbol{R}^{(t+1)}(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\mu}^{(t+1)})\}\phi^{p/2} \\ &\quad \times \exp(-\frac{1}{2}\phi v^{(t+1)}\xi^{(t+1)})\phi^{v^{(t+1)}/2-1}, \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{R}^{(t+1)} &= (\boldsymbol{R}'^{(t)\,-1} + \frac{\sigma^2}{\xi'^{(t)}}\mathbf{I}_p)^{-1} \\ \boldsymbol{\mu}^{(t+1)} &= \boldsymbol{\mu}'^{(t)} \\ v^{(t+1)} &= v'^{(t)} \\ \xi^{(t+1)} &= \xi'^{(t)}. \end{aligned}$$

This marginalization gives the distribution of $(\boldsymbol{\theta}^{(t+1)}, \phi)$ given D_t, that is,

$$(\boldsymbol{\theta}^{(t+1)}, \phi|\sigma^2, D_t) \sim \mathrm{NG}(\boldsymbol{\mu}^{(t+1)}, \xi^{(t+1)}, \boldsymbol{R}^{(t+1)}, v^{(t+1)}).$$

This defines a single iteration of the forecasting calculation. Note that the only difference between the distributions of $(\boldsymbol{\theta}^{(t)}, \phi|\sigma^2, D_t)$ and $(\boldsymbol{\theta}^{(t+1)}, \phi|\sigma^2, D_t)$ is that the conditional variance of the rating parameters in the latter distribution has been increased by $\sigma^2\mathbf{I}_p$.

2.4.2 Predictive distribution of $(d_{t+1}|\sigma^2, D_t)$

The forecast distribution $(d_{t+1}|\sigma^2, D_t)$, $t = 0, \ldots, T$, for the $n^{(t+1)}$ observations to be observed during season $t+1$ can be obtained in the following manner. We have

$$\begin{aligned} (d_{t+1}|\boldsymbol{X}^{(t+1)}, \boldsymbol{\theta}^{(t+1)}, \phi, \sigma^2, D_t) &\sim \mathrm{N}(\boldsymbol{X}^{(t+1)}\boldsymbol{\theta}^{(t+1)}, \frac{1}{\phi}\mathbf{I}_{n^{(t+1)}}) \\ (\boldsymbol{\theta}^{(t+1)}|\phi, \sigma^2, D_t) &\sim \mathrm{N}(\boldsymbol{\mu}^{(t+1)}, (\phi\boldsymbol{R}^{(t+1)})^{-1}) \\ (\phi|\sigma^2, D_t) &\sim \mathrm{Gamma}(v^{(t+1)}/2, v^{(t+1)}\xi^{(t+1)}/2). \end{aligned}$$

The conditional distribution of d_{t+1} given ϕ, but marginal over $\boldsymbol{\theta}^{(t+1)}$, is normal, and has mean and variance given by

$$\begin{aligned} &\mathrm{E}(d_{t+1}|\boldsymbol{X}^{(t+1)}, \boldsymbol{\mu}^{(t+1)}, \phi, \sigma^2, D_t) \\ &\quad= \mathrm{E}(\mathrm{E}(d_{t+1}|\boldsymbol{\theta}^{(t+1)}, \phi, \sigma^2, D_t)) \\ &\quad= \mathrm{E}(\boldsymbol{X}^{(t+1)}\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\mu}^{(t+1)}, \phi, \sigma^2, D_t) = \boldsymbol{X}^{(t+1)}\boldsymbol{\mu}^{(t+1)} \end{aligned}$$

$$\begin{aligned} &\mathrm{Var}(d_{t+1}|\boldsymbol{X}^{(t+1)}, \boldsymbol{\mu}^{(t+1)}, \boldsymbol{R}^{(t+1)}, \phi, \sigma^2, D_t) \\ &\quad= \mathrm{E}(\mathrm{Var}(d_{t+1}|\boldsymbol{\theta}^{(t+1)}, \boldsymbol{R}^{(t+1)}, \phi, \sigma^2, D_t)) + \mathrm{Var}(\mathrm{E}(d_{t+1}|\boldsymbol{\theta}^{(t+1)}, \boldsymbol{R}^{(t+1)}, \phi, \sigma^2, D_t)) \end{aligned}$$

$$= \mathrm{E}(\frac{1}{\phi}\mathbf{I}_{n^{(t+1)}}) + \mathrm{Var}(\boldsymbol{X}^{(t+1)}\boldsymbol{\theta}^{(t+1)})$$
$$= \frac{1}{\phi}(\mathbf{I}_{n^{(t+1)}} + \boldsymbol{X}^{(t+1)}\boldsymbol{R}^{(t+1)\ -1}\boldsymbol{X}^{(t+1)\mathsf{T}}).$$

This implies

$$(d_{t+1}, \phi|\sigma^2, D_t) \sim \mathrm{NG}(\boldsymbol{X}^{(t+1)}\boldsymbol{\mu}^{(t+1)}, \xi^{(t+1)}, \mathbf{I}_{n^{(t+1)}} + \boldsymbol{X}^{(t+1)}\boldsymbol{R}^{(t+1)\ -1}\boldsymbol{X}^{(t+1)\mathsf{T}}, v^{(t+1)}),$$

from which we obtain the standard result that the distribution marginal over ϕ is given by the t-distribution

$$(d_{t+1}|\sigma^2, D_t) \sim T_{v^{(t+1)}}(\boldsymbol{X}^{(t+1)}\boldsymbol{\mu}^{(t+1)}, \xi^{(t+1)}(\mathbf{I}_{n^{(t+1)}} + \boldsymbol{X}^{(t+1)}\boldsymbol{R}^{(t+1)\ -1}\boldsymbol{X}^{(t+1)\mathsf{T}})). \quad (2.16)$$

The univariate predictive distribution for $y_{ij}^{(t+1)}$ is

$$(y_{ij}^{(t+1)}|\sigma^2, D_t) \sim T_{v^{(t+1)}}(\mu_i^{(t+1)} - \mu_j^{(t+1)}, \xi^{(t+1)}(1 + (\boldsymbol{R}^{(t+1)})_{ii}^{-1} + (\boldsymbol{R}^{(t+1)})_{jj}^{-1} - 2(\boldsymbol{R}^{(t+1)})_{ij}^{-1})).$$

From this distribution, it is straightforward to calculate $\Pr(y_{ij}^{(t+1)} > 0|\sigma^2, D_t)$ numerically.

2.5 Analysis with σ^2 unknown

In most paired comparison experiments the extra variance, σ^2, describing the magnitude of changes in the rating parameters between successive competitions is unknown, and should realistically be treated as an unknown parameter in the analysis of the model. Furthermore, it is often of direct interest to learn from the data what are plausible values of σ^2. In this section, we explore two approaches to analyzing the model with unknown σ^2.

To motivate our two approaches, suppose we assume the dynamic model of (2.8)–(2.12). We observe T tournaments, and are interested in the marginal posterior distribution $(\boldsymbol{\theta}^{(T)}, \phi|D_T)$. To obtain this distribution, we can marginalize over the conditional distribution given $\omega = 1/\sigma^2$,

$$f(\boldsymbol{\theta}^{(T)}, \phi|D_T) = \int f(\boldsymbol{\theta}^{(T)}, \phi|\omega, D_T)\ f(\omega|D_T)d\omega.$$

The first density in the integrand is normal-gamma, and is calculated by the updating and forecasting recursions for known σ^2 of Section 2.4. The second density is difficult to obtain in closed form. We propose, therefore, two approaches to the analysis. One approach involves drawing a large random sample from the marginal distribution of $(\omega|D_T)$ and uses the sample as an approximation to the exact distribution. This is accomplished using the Gibbs sampler as described in Section 2.5.1. Alternatively, we can approximate the prior distribution of ω by a discrete distribution, so that the posterior probability distribution of $(\omega|D_T)$ is easily calculated. This approach, a non-iterative analysis, is described in Section 2.5.2. In either case, to compute the marginal posterior distribution of

$(\boldsymbol{\theta}^{(T)}, \phi | D_T)$, the first conditional density in the integrand is integrated over the approximate distribution of $(\omega | D_T)$, and the computation of marginal posterior distributions of $\boldsymbol{\theta}^{(T)}$ or d_{T+1} becomes a more tractable problem. Section 2.7 discusses how to obtain these marginal posterior distributions.

2.5.1 Analysis using Iterative Simulation

The Gibbs sampler, an iterative simulation technique for drawing samples and summarizing parameters of distributions that would otherwise be difficult to examine, is a computational approach that is being used increasingly in the analysis of Bayesian models. The algorithm was first described in detail in Geman and Geman (1984), though many of the ideas underlying the Gibbs sampler have been developed in previous work (Besag 1974; Hastings 1970; Metropolis et al. 1953). More recently, Gelfand and Smith (1990) develop the Gibbs sampler methodology in a general framework and discuss the properties of the algorithm. The flexibility of the iterative simulation approach has enabled the analysis of Bayesian models in which closed form posterior distributions are difficult or impossible to obtain.

To describe the use of the Gibbs sampler, suppose we have m random variables $Z_1, \ldots, Z_m$, and we are able to specify and generate random samples from the conditional distributions $(Z_i | Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_m)$, $i = 1, \ldots, m$, that is, the distribution of each of the Z_i conditional on the rest. The Gibbs sampler begins with an arbitrary set of starting values $Z_1^{(0)}, \ldots, Z_m^{(0)}$. Then at iteration q, we draw

$$\begin{array}{ccl} Z_1^{(q)} & \sim & f(Z_1 | Z_2^{(q-1)}, \ldots, Z_m^{(q-1)}) \\ Z_2^{(q)} & \sim & f(Z_2 | Z_1^{(q)}, Z_3^{(q-1)}, \ldots, Z_m^{(q-1)}) \\ \vdots & & \vdots \\ Z_m^{(q)} & \sim & f(Z_m | Z_1^{(q)}, \ldots, Z_{m-1}^{(q)}). \end{array}$$

Under regularity conditions specified in Geman and Geman (1984), the distribution of $(Z_1^{(l)}, \ldots, Z_m^{(l)})$ converges to the joint distribution of $(Z_1, \ldots, Z_m)$ as $l \to \infty$.

Assessing convergence of the algorithm is still under investigation (see, for example, Gelman and Rubin 1992a; Geyer 1992). The approach used throughout this paper, proposed by Gelman and Rubin (1992a), involves running multiple Gibbs samplers with overdispersed starting values, and monitors convergence by comparing the within-sample variance of single samplers to the between-sample variance. Other approaches to monitoring convergence are mentioned in Gelfand et al. (1990) and Zeger and Karim (1991).

We use the Gibbs sampler in the analysis of the dynamic paired comparison model to approximate the marginal posterior distribution of $(\omega | D_T)$. This is done by alternately sampling from the conditional posterior distribution of $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi | \omega, D_T)$ and the conditional posterior distribution of $(\omega | \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi, D_T)$. Liu (1992) notes that the

Gibbs sampler will converge more quickly if parameters can be combined, as in this case, so that rather than alternating among a total of $T+2$ conditional distributions, we only alternate between two conditional distributions. After the Gibbs sampler converges, we continue the iterative simulation to obtain a random sample from the marginal posterior distribution of $(\omega|D_T)$. We now describe the methods used to draw samples from each of the conditional distributions.

Steps of the Gibbs Sampler

The Gibbs sampler, which we describe in detail below, proceeds as follows (the subscript c is used to indicate parameters part of the current Gibbs draw):

1. Pick a starting value, ω_c, for the innovation precision.

2. Obtain the normal-gamma distribution for $(\boldsymbol{\theta}^{(T)}, \phi|\omega_c, D_T)$ via the updating and forecasting operations of Section 2.4.

3. Draw ϕ_c from
$$(\phi|\omega_c, D_T) \sim \mathrm{Gamma}(v'^{(T)}/2, v'^{(T)}\xi'^{(T)}/2).$$

4. Draw $\boldsymbol{\theta}_c^{(T)}$ from
$$(\boldsymbol{\theta}^{(T)}|\phi_c, \omega_c, D_T) \sim \mathrm{N}(\boldsymbol{\mu}'^{(T)}, (\phi_c \boldsymbol{R}'^{(T)})^{-1}).$$

5. For $t = T-1, \ldots, 1$, successively draw $\boldsymbol{\theta}_c^{(t)}$ from
$$(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}_c^{(t+1)}, \ldots, \boldsymbol{\theta}_c^{(T)}, \phi_c, \omega_c, D_T) \sim \mathrm{N}(V^{(t)}(\phi \boldsymbol{R}'^{(t)} \boldsymbol{\mu}'^{(t)} + \omega_c \boldsymbol{\theta}^{(t+1)}), V^{(t)}),$$
where $V^{(t)} = (\phi \boldsymbol{R}'^{(t)} + \omega_c \mathbf{I}_p)^{-1}$. This completes the first half of a Gibbs sampler iteration, that is, a draw from the distribution of $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi|\omega, D_T)$.

6. Draw ω_c from
$$(\omega|\boldsymbol{\theta}_c^{(1)}, \ldots, \boldsymbol{\theta}_c^{(T)}, \phi_c, D_T) \sim \mathrm{Gamma}(a_0 + p(T-1)/2, b_0 + \frac{1}{2}\sum_{t=1}^{T-1}(\boldsymbol{\theta}_c^{(t+1)} - \boldsymbol{\theta}_c^{(t)})^{\mathsf{T}}(\boldsymbol{\theta}_c^{(t+1)} - \boldsymbol{\theta}_c^{(t)})).$$
This completes the second half of a Gibbs sampler iteration.

Repeat steps (2) through (6) until convergence.

Sampling from $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi|\omega, D_T)$

In Section 2.4, we describe how to obtain the posterior distribution of $(\boldsymbol{\theta}^{(T)}, \phi|\omega, D_T)$ by the updating and forecasting recursions (steps (2)–(4)). Now we are interested in the posterior distribution of the remaining parameters conditional on ω.

Let ω_c be the current draw of ω in the Gibbs sampler. To draw from the joint distribution of $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi|\omega_c, D_T)$, we invoke the following argument. Suppose we want to draw from the joint distribution of random variables $(X_1, X_2, \ldots, X_n)$. To do so, we can first draw from the marginal distribution of X_n, and then draw from the conditional distribution of $(X_{n-1}|X_n)$, $(X_{n-2}|X_{n-1}, X_n)$, and so on, until we draw from $(X_1|X_2, \ldots, X_n)$. This process gives a single draw from the joint distribution of all variables.

For the current problem, we can rewrite the joint posterior of $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi|\omega_c, D_T)$ as a product of conditionally independent densities,

$$\begin{aligned} f(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi|\omega_c, D_T) &= f(\boldsymbol{\theta}^{(T)}, \phi|\omega_c, D_T) \times f(\boldsymbol{\theta}^{(T-1)}|\boldsymbol{\theta}^{(T)}, \phi, \omega_c, D_T) \\ &\quad \times \cdots \times f(\boldsymbol{\theta}^{(1)}|\boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi, \omega_c, D_T). \end{aligned} \tag{2.17}$$

This factorization suggests that to draw from the joint distribution of $T+1$ parameters on the left-hand side of (2.17), we can successively draw from each of the conditional distributions on the right-hand side. Given a particular value, ω_c, of the system precision, we can compute the distribution of $(\boldsymbol{\theta}^{(T)}, \phi|\omega_c, D_T)$ using the updating methodology of Section 2.4.

To understand how to draw from the distribution of $(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}_c^{(t+1)}, \ldots, \boldsymbol{\theta}_c^{(T)}, \phi_c, \omega_c, D_T)$, $t = 1, \ldots, T-1$, it is helpful to write down the joint posterior distribution of all parameters. Conditioning now on ω_c, we have

$$\begin{aligned} &f(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi|\omega_c, D_T) \\ &\propto \ \{f(\phi|\omega_c, D_0)\} \\ &\quad \times\{f(\boldsymbol{\theta}^{(1)}|\phi, \omega_c, D_0)\} \\ &\quad \times\{f(d_1|\boldsymbol{\theta}^{(1)}, \phi, \omega_c, D_0)\} \\ &\quad \times\prod_{t=2}^{T}\Big\{\{f(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)}, \ldots, \boldsymbol{\theta}^{(1)}, \phi, \omega_c, D_{t-1})\} \\ &\qquad\qquad \times\ \{f(d_t|\boldsymbol{\theta}^{(t)}, \ldots, \boldsymbol{\theta}^{(1)}, \phi, \omega_c, D_{t-1})\}\Big\} \end{aligned}$$

which we can write as

$$\begin{aligned} &f(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi|\omega_c, D_T) \\ &\propto \ \{\exp(-v^{(1)}\xi^{(1)}\phi/2)\,\phi^{v^{(1)}/2-1}\} \\ &\quad \times\{\exp(-\frac{\phi}{2}(\boldsymbol{\theta}^{(1)} - \boldsymbol{\mu}^{(1)})^{\mathsf{T}}\boldsymbol{R}^{(1)}(\boldsymbol{\theta}^{(1)} - \boldsymbol{\mu}^{(1)}))\} \\ &\quad \times\{\phi^{n^{(1)}/2}\exp(-\frac{\phi}{2}(\boldsymbol{y}^{(1)} - \boldsymbol{X}^{(1)}\boldsymbol{\theta}^{(1)})^{\mathsf{T}}(\boldsymbol{y}^{(1)} - \boldsymbol{X}^{(1)}\boldsymbol{\theta}^{(1)}))\} \\ &\quad \times\prod_{t=2}^{T}\Big\{\{\exp(-\frac{\omega_c}{2}(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)})^{\mathsf{T}}(\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}))\} \\ &\qquad\qquad \times\ \{\phi^{n^{(t)}/2}\exp(-\frac{\phi}{2}(\boldsymbol{y}^{(t)} - \boldsymbol{X}^{(t)}\boldsymbol{\theta}^{(t)})^{\mathsf{T}}(\boldsymbol{y}^{(t)} - \boldsymbol{X}^{(t)}\boldsymbol{\theta}^{(t)}))\}\Big\}. \end{aligned}$$

We now compute the joint posterior distribution marginalized over $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(t-1)}$. This is accomplished by performing the updating and forecasting computations of Section 2.4

to obtain the marginal distribution of $(\boldsymbol{\theta}^{(t)}, \phi | \omega_c, D_t)$. This results in

$$\begin{aligned}
&f(\boldsymbol{\theta}^{(t)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi | \omega_c, D_T) \\
&\propto \; f(\boldsymbol{\theta}^{(t)}, \phi | \omega_c, D_t) \times f(\boldsymbol{\theta}^{(t+1)}, \ldots, \boldsymbol{\theta}^{(T)} | \boldsymbol{\theta}^{(t)}, \phi, \omega_c, D_T) \\
&\propto \; \{\exp(-v'^{(t)} \xi'^{(t)} \phi/2)\, \phi^{v'^{(t)}/2-1}\} \\
&\quad \times \{\exp(-\frac{\phi}{2}(\boldsymbol{\theta}^{(t)} - \boldsymbol{\mu}'^{(t)})^{\mathsf{T}} \boldsymbol{R}'^{(t)} (\boldsymbol{\theta}^{(t)} - \boldsymbol{\mu}'^{(t)}))\} \\
&\quad \times \{\phi^{n^{(t+1)}/2} \exp(-\frac{\phi}{2}(\boldsymbol{y}^{(t+1)} - \boldsymbol{X}^{(t+1)} \boldsymbol{\theta}^{(t+1)})^{\mathsf{T}} (\boldsymbol{y}^{(t+1)} - \boldsymbol{X}^{(t+1)} \boldsymbol{\theta}^{(t+1)}))\} \\
&\quad \times \{\exp(-\frac{\omega_c}{2}(\boldsymbol{\theta}^{(t+2)} - \boldsymbol{\theta}^{(t+1)})^{\mathsf{T}} (\boldsymbol{\theta}^{(t+2)} - \boldsymbol{\theta}^{(t+1)}))\} \\
&\quad \vdots \\
&\quad \times \{\exp(-\frac{\omega_c}{2}(\boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}^{(T-1)})^{\mathsf{T}} (\boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}^{(T-1)}))\} \\
&\quad \times \{\phi^{n^{(T)}/2} \exp(-\frac{\phi}{2}(\boldsymbol{y}^{(T)} - \boldsymbol{X}^{(T)} \boldsymbol{\theta}^{(T)})^{\mathsf{T}} (\boldsymbol{y}^{(T)} - \boldsymbol{X}^{(T)} \boldsymbol{\theta}^{(T)}))\}. \qquad (2.18)
\end{aligned}$$

The conditional distribution of $(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}^{(t+1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi, \omega_c, D_T)$ is proportional to the marginal posterior distribution in (2.18) treating the parameters $\boldsymbol{\theta}^{(t+1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi, \omega$ as constants. This is simplified by neglecting the terms that are constant with respect to $\boldsymbol{\theta}^{(t)}$, which yields

$$\begin{aligned}
&f(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}^{(t+1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi, \omega_c, D_T) \\
&\propto \; \{\exp(-\frac{\phi}{2}(\boldsymbol{\theta}^{(t)} - \boldsymbol{\mu}'^{(t)})^{\mathsf{T}} \boldsymbol{R}'^{(t)} (\boldsymbol{\theta}^{(t)} - \boldsymbol{\mu}'^{(t)}))\} \\
&\quad \times \{\exp(-\frac{\omega_c}{2}(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)})^{\mathsf{T}} (\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}))\}.
\end{aligned}$$

Thus, conditional on $\boldsymbol{\theta}^{(t+1)}$, $\boldsymbol{\theta}^{(t)}$ is independent of the data $d_{t+1}, \ldots, d_T$. We therefore have

$$(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}^{(t+1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi, \omega_c, D_T) \sim \mathrm{N}(V^{(t)}(\phi \boldsymbol{R}'^{(t)} \boldsymbol{\mu}'^{(t)} + \omega_c \boldsymbol{\theta}^{(t+1)}), V^{(t)}), \qquad (2.19)$$

where $V^{(t)} = (\phi \boldsymbol{R}'^{(t)} + \omega_c \mathbf{I}_p)^{-1}$.

So to perform the iterative sampling procedure for drawing from $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi | \omega_c, D_T)$, we first draw $\boldsymbol{\theta}_c^{(T)}$ and ϕ_c given ω_c and D_T as in steps (3) and (4). In obtaining the normal-gamma parameters for $(\boldsymbol{\theta}^{(T)}, \phi | \omega_c, D_T)$, we retain during the updating and forecasting operations the normal-gamma parameters for $(\boldsymbol{\theta}^{(t)}, \phi | \omega_c, D_t)$, $t = 1, \ldots, T$. These parameters are computed in the intermediate steps as in Section 2.4. Now from (2.19), we can draw $\boldsymbol{\theta}_c^{(T-1)}$ from $(\boldsymbol{\theta}^{(T-1)} | \boldsymbol{\theta}_c^{(T)}, \phi_c, \omega_c, D_T)$, and recursively, we can draw $\boldsymbol{\theta}_c^{(t)}$ from $(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}_c^{(t+1)}, \ldots, \boldsymbol{\theta}_c^{(T)}, \phi_c, \omega_c, D_T)$. This process gives us a complete single draw from the conditional distribution of $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi | \omega_c, D_T)$, which completes the first half of a single Gibbs sampler iteration (step 5).

Sampling from $(\omega | \boldsymbol{\theta}_c^{(1)}, \ldots, \boldsymbol{\theta}_c^{(T)}, \phi_c, D_T)$

To perform the second half of one iteration of the Gibbs sampler, assuming we have draws $\boldsymbol{\theta}_c^{(1)}, \ldots, \boldsymbol{\theta}_c^{(T)}, \phi_c$, we now want to draw ω_c from the conditional distribution of $(\omega|\boldsymbol{\theta}_c^{(1)}, \ldots, \boldsymbol{\theta}_c^{(T)}, \phi_c, D_T)$. Suppressing the conditioning on hyperparameters, we have

$$\begin{aligned}
&f(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi, \omega|D_T)\\
&\quad\propto \;\; \{\omega^{a_0-1}e^{-b_0\omega}\}\\
&\qquad\quad \times\{\exp(-v^{(1)}\xi^{(1)}\phi/2)\,\phi^{v^{(1)}/2-1}\}\\
&\qquad\quad \times\{\exp(-\frac{\phi}{2}(\boldsymbol{\theta}^{(1)}-\boldsymbol{\mu}^{(1)})^{\mathsf{T}}\boldsymbol{R}^{(1)}(\boldsymbol{\theta}^{(1)}-\boldsymbol{\mu}^{(1)}))\}\\
&\qquad\quad \times\{\phi^{n^{(1)}/2}\exp(-\frac{\phi}{2}(\boldsymbol{y}^{(1)}-\boldsymbol{X}^{(1)}\boldsymbol{\theta}^{(1)})^{\mathsf{T}}(\boldsymbol{y}^{(1)}-\boldsymbol{X}^{(1)}\boldsymbol{\theta}^{(1)}))\}\\
&\qquad\quad \times\prod_{t=2}^{T}\Big\{\{\omega^{p/2}\exp(-\frac{\omega}{2}(\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^{(t-1)})^{\mathsf{T}}(\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^{(t-1)}))\}\\
&\qquad\qquad\quad \times\;\{\phi^{n^{(t)}/2}\exp(-\frac{\phi}{2}(\boldsymbol{y}^{(t)}-\boldsymbol{X}^{(t)}\boldsymbol{\theta}^{(t)})^{\mathsf{T}}(\boldsymbol{y}^{(t)}-\boldsymbol{X}^{(t)}\boldsymbol{\theta}^{(t)}))\}\Big\}
\end{aligned}$$

The conditional posterior density, $f(\omega|\boldsymbol{\theta}_c^{(1)}, \ldots, \boldsymbol{\theta}_c^{(T)}, \phi_c, D_T)$, is proportional to the above density. Ignoring terms that are constant with respect to ω, we have

$$\begin{aligned}
&f(\omega|\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi, D_T)\\
&\quad\propto \;\; \{\omega^{a_0-1}e^{-b_0\omega}\}\\
&\qquad\quad \times\prod_{t=2}^{T}\Big\{\{\omega^{p/2}\exp(-\frac{\omega}{2}(\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^{(t-1)})^{\mathsf{T}}(\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^{(t-1)}))\}\Big\}\\
&\quad\propto \;\; \{\omega^{a_0-1}e^{-b_0\omega}\}\\
&\qquad\quad \times\{\omega^{p(T-1)/2}\exp(-\frac{\omega}{2}\sum_{t=1}^{T-1}(\boldsymbol{\theta}^{(t+1)}-\boldsymbol{\theta}^{(t)})^{\mathsf{T}}(\boldsymbol{\theta}^{(t+1)}-\boldsymbol{\theta}^{(t)}))\},
\end{aligned}$$

so that

$$(\omega|\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi, D_T) \sim \text{Gamma}(a_0+p(T-1)/2, b_0+\frac{1}{2}\sum_{t=1}^{T-1}(\boldsymbol{\theta}^{(t+1)}-\boldsymbol{\theta}^{(t)})^{\mathsf{T}}(\boldsymbol{\theta}^{(t+1)}-\boldsymbol{\theta}^{(t)})). \tag{2.20}$$

Therefore, to complete the second half of one iteration of the Gibbs sampler (step 6), we randomly draw ω_c from a Gamma distribution conditional on the parameters drawn from the first half of the Gibbs sampler iteration. When the successive $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\theta}^{(t+1)}$ are close to one another, the sum of squared deviations in (2.20) will be small, so the distribution of ω has large mean. This indicates that the amount of variability between competitions is small, which seems consistent with the successive $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\theta}^{(t+1)}$ being close. Conversely, when the sum in (2.20) is large, the value of ω drawn from the distribution is likely to be small, corresponding to a large system variance, σ^2.

Carlin, Polson and Stoffer (1992) develop a methodology for modeling a large class of state-space models using the Gibbs sampler to approximate the posterior distribution of parameters. Their approach, which can be used to analyze the normal dynamic linear models of the form (2.8)–(2.11) as a special case, involves constructing a Gibbs sampler that alternately draws from many conditional distributions. However, as demonstrated in Liu (1992), if some parameters can be "grouped" into joint distributions (conditional on the rest), then the resulting Gibbs sampler is a more efficient procedure in the sense of faster convergence to the unconditional posterior distribution. While Carlin, Polson and Stoffer (1992) iterate among many conditional distributions, our approach involves iterating between exactly two, so that our Gibbs sampler will converge at least as quickly.

2.5.2 Analysis using a Discretized Prior on ω

An alternative approach to obtaining the marginal posterior distribution of $\boldsymbol{\theta}^{(t)}$ involves using a discrete approximation to the prior distribution of ω. This approximation may be done in any of several reasonable ways. First, we may draw a random sample of size m from a continuous prior distribution, in which case the empirical distribution of the sample approximates the true prior distribution. Another possibility is to partition the range of ω into m intervals, compute the probability associated with each interval, and compute the centroid of each interval. A third approach, which seems the most reasonable to use in practice, is to select m grid values of ω that are overdispersed for the distribution of ω, and assign probabilities for each value corresponding to prior beliefs.

Suppose we have selected a set of m values of ω, $\{\omega_1, \ldots, \omega_m\}$, and assume we have approximated the prior distribution by the probability function

$$f(\omega|D_0) = \begin{cases} \pi_i^{(0)} & \text{if } \omega = \omega_i \\ 0 & \text{if } \omega \notin \{\omega_1, \ldots, \omega_m\} \end{cases}$$

where we assume $f(\omega|D_0)$ has mass only on the set $\{\omega_1, \ldots, \omega_m\}$. For any season t, we have by Bayes' theorem

$$f(\omega|D_t) = \frac{f(d_t|D_{t-1}, \omega)}{f(d_t|D_{t-1})}\, f(\omega|D_{t-1}).$$

The marginal distribution of ω given data through season t has a simple relationship to the distribution of ω given data through season $t-1$. Continuing the recursion, we have for all t

$$\begin{aligned} f(\omega|D_t) &= \frac{\prod_{j=1}^{t} f(d_j|D_{j-1}, \omega)}{\prod_{j=1}^{t} f(d_j|D_{j-1})} f(\omega|D_0) \\ &\propto f(\omega|D_0) \prod_{j=1}^{t} f(d_j|D_{j-1}, \omega). \end{aligned}$$

Thus the marginal density of $(\omega|D_T)$ is proportional to the product of the prior density of $(\omega|D_0)$ and a reweighting factor, which is the product of conditional likelihoods of the form $f(d_t|D_{t-1}, \omega)$. The reweighting factor for a specific value of ω is a measure of how likely the data were to occur given the previous data and the specific value of ω. When the reweighting factor is small for a given value of ω (relative to other values of ω), the data is not well predicted from the given value of ω, so the posterior probability of that specific value of ω is small. From Section 2.4, $(d_t|D_{t-1}, \omega)$ has a multivariate t-distribution with density

$$f(d_t|\omega, D_{t-1}) \propto \frac{1}{|C^{(t)}|^{1/2}}(v^{(t)} + (d_t - \boldsymbol{X}^{(t)}\boldsymbol{\mu}^{(t)})^{\mathsf{T}}C^{(t)\,-1}(d_t - \boldsymbol{X}^{(t)}\boldsymbol{\mu}^{(t)})/\xi^{(t)})^{-(v^{(t)}+n^{(t)})/2},$$

with $C^{(t)} = (I_{n^{(t)}} + \boldsymbol{X}^{(t)}\boldsymbol{R}^{(t)\,-1}\boldsymbol{X}^{(t)\mathsf{T}})$. Therefore, the posterior distribution of ω is obtained by reweighting the prior distribution of ω by a product of multivariate t-densities, so that

$$\pi_i^{(T)} = \frac{\pi_i^{(0)}\prod_{j=1}^{T} f(d_j|\omega_i, D_{j-1})}{\sum_{j=1}^{m}\pi_j^{(0)}\prod_{j=1}^{T} f(d_T|\omega_j, D_{T-1})}.$$

It is therefore a straightforward procedure to approximate the posterior distribution of ω.

The attractiveness to this approach over the iterative simulation approach is its lower computational cost. In performing iterative simulation, let N denote the number of iterations performed in a single Gibbs sampler, and let g be the number of parallel Gibbs samplers run to assess convergence. We must perform Ng forward filters, that is, the computation that produces the parameters for the distribution of $(\boldsymbol{\theta}^{(T)}, \phi|\omega, D_T)$ from $(\boldsymbol{\theta}^{(1)}, \phi|\omega, D_0)$. In addition, we must run Ng back filters, that is, the computation that produces the distributions of $(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t+1)}, \ldots, \boldsymbol{\theta}^{(T)}, \phi, \omega, D_T)$, $t = T-1, \ldots, 1$. In the non-iterative analysis, we only need to run m forward filters and no back filters. If we set $m = 150$, $g = 6$ and $N = 500$, all reasonable parameter values, then the iterative approach is 20 times slower than the non-iterative simulation approach of this section, and 40 times slower if back filtering is considered as expensive as forward filtering.

The main disadvantage of the non-iterative approach relative to the Gibbs sampler stems from choosing the initial set of values of ω. If the data show that the support of $(\omega|D_T)$ is not well represented by the set $\{\omega_1, \ldots, \omega_m\}$ so that the prior distribution on ω is misspecified, the posterior distribution of $(\omega|D_T)$ will not be approximated accurately.

2.6 Inclusion of Non-dynamic Linear Covariates

Paired comparison experiments often include data which might be useful predictors of the outcome of future comparisons. The model stipulated in Section 2.3 can be extended to include linear covariates whose distribution does not change due to the passage of time. In this section, we describe a model that includes linear non-dynamic covariates into the dynamic model, and show how to modify the analyses of Sections 2.4 and 2.5 in the presence of these covariates.

Suppose we observe with each paired comparison a set of q variables. Let $W^{(t)}$ to be the $n^{(t)} \times q$ covariate data matrix, and let β be the vector of q parameters associated with the variables where β is assumed to remain constant over time. The model for score differences at time t is now given by

$$\boldsymbol{y}^{(t)} \sim \mathrm{N}(\boldsymbol{X}_*^{(t)}\boldsymbol{\theta}_*^{(t)}, \frac{1}{\phi}\mathbf{I}_{n^{(t)}}),$$

where $\boldsymbol{X}_*^{(t)}$ is the $n \times (p+q)$ concatenation of $\boldsymbol{X}^{(t)}$ and $W^{(t)}$, and $\boldsymbol{\theta}_*^{(t)}$ is the $(p+q)-$vector $(\boldsymbol{\theta}^{(t)}, \beta)$. The probability model that describes the evolution of the parameters, $\boldsymbol{\theta}_*^{(t)}$, is given by

$$\boldsymbol{\theta}_*^{(t)} = \boldsymbol{\theta}_*^{(t-1)} + \boldsymbol{\nu}_*^{(t)}.$$

The parameters β, the last q components of $\boldsymbol{\theta}_*^{(t)}$, are assumed to remain constant over time, so we assume

$$(\boldsymbol{\nu}_*^{(t)}|\sigma^2) \sim \mathrm{N}(\alpha_{*t}, \sigma^2\mathbf{J}_{p,q}),$$

where $\alpha_{*t} = (\alpha_t, 0, 0, \ldots, 0)$, that is, the vector α_t followed by q zeroes, and $J_{p,q}$ is the $p+q$ dimensional square matrix consisting of zeroes everywhere, except that the first p diagonal elements are equal to one. This formulation of the model specifies that the first p components of $\boldsymbol{\theta}_*^{(t)}$ are dynamic and are allowed to change over time according to the model in (2.9). The last q elements of $\boldsymbol{\theta}_*^{(t)}$ remain fixed over time, and as more data are accumulated these parameters are estimated with greater precision.

Analysis conditional on σ^2

The analysis of Section 2.4 where σ^2 is known remains essentially the same with a few modifications in the updating and forecasting formulas. The formulas in (2.15), which update the normal-gamma parameters upon observing new data, do not change with the addition of non-dynamic parameters,

$$\begin{aligned}
\boldsymbol{R}_*'^{(t)} &= \boldsymbol{R}_*^{(t)} + \boldsymbol{X}_*^{(t)\mathsf{T}}\boldsymbol{X}_*^{(t)} \\
\boldsymbol{\mu}_*'^{(t)} &= \boldsymbol{R}_*'^{(t)\,-1}(\boldsymbol{R}_*^{(t)}\boldsymbol{\mu}_*^{(t)} + \boldsymbol{X}_*^{(t)\mathsf{T}}\boldsymbol{y}^{(t)}) \\
v'^{(t)} &= v^{(t)} + n^{(t)} \\
\xi'^{(t)} &= \frac{1}{v'^{(t)}}[(v^{(t)}\xi^{(t)} + \boldsymbol{\mu}_*^{(t)\mathsf{T}}\boldsymbol{R}_*^{(t)}\boldsymbol{\mu}_*^{(t)}) + \boldsymbol{y}^{(t)\mathsf{T}}\boldsymbol{y}^{(t)} - \boldsymbol{\mu}_*'^{(t)\mathsf{T}}\boldsymbol{R}_*'^{(t)}\boldsymbol{\mu}_*'^{(t)}]
\end{aligned} \tag{2.21}$$

The forecasting equations change slightly because the non-dynamic parameters do not increase in variance due to the passage of time. We therefore have

$$\begin{aligned}
\boldsymbol{R}_*^{(t+1)} &= (\boldsymbol{R}_*'^{(t)\,-1} + \frac{\sigma^2}{\xi'^{(t)}}\mathbf{J}_{p,q})^{-1} \\
\boldsymbol{\mu}_*^{(t+1)} &= \boldsymbol{\mu}_*'^{(t)} \\
v^{(t+1)} &= v'^{(t)} \\
\xi^{(t+1)} &= \xi'^{(t)}.
\end{aligned} \tag{2.22}$$

This demonstrates how to calculate the marginal distribution of $\boldsymbol{\theta}_*^{(T)}$ given σ^2 through an updating and forecasting recursion.

Gibbs Sampler analysis

The use of iterative simulation in Section 2.5.1 undergoes several important changes in the presence of non-dynamic parameters. The goal of the Gibbs sampler is to sample alternately between the distribution of $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \beta, \phi | \omega, D_T)$ and $(\omega | \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \beta, \phi, D_T)$. The difficulty in this analysis is that drawing $\boldsymbol{\theta}^{(t)}$ is affected by the presence of the non-dynamic parameter β. The strategy we employ is to factor the distribution of $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \beta, \phi | \omega, D_T)$ into a product of conditional densities,

$$f(\beta, \phi | \omega, D_T) \times f(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)} | \beta, \phi, \omega, D_T)$$

To perform the sampling, we obtain the normal-gamma distribution for $(\boldsymbol{\theta}_*^{(T)} = (\boldsymbol{\theta}^{(T)}, \beta), \phi | \omega, D_T)$ via the updating and forecasting formulas described above. We next sample ϕ_c from its gamma distribution, and then sample β_c conditional on ϕ_c from its normal distribution. To sample the $\boldsymbol{\theta}^{(t)}$, $t = 1, \ldots, T$, we need to recompute the conditional normal distributions of $(\boldsymbol{\theta}^{(t)} | \beta_c, \phi_c, d_t)$. This can easily be shown (see, for example, Dillon and Goldstein 1984, pg. 546) to have the distribution

$$(\boldsymbol{\theta}^{(t)} | \beta_c, \phi_c) \sim \mathrm{N}(\boldsymbol{\mu}_\theta'^{(t)} + \Sigma_{\theta,\beta}'^{(t)} \Sigma_{\beta,\beta}'^{(t)\,-1} (\boldsymbol{\mu}_\beta'^{(t)} - \beta_c), \Sigma_{\theta,\theta}'^{(t)} - \Sigma_{\theta,\beta}'^{(t)} \Sigma_{\beta,\beta}'^{(t)\,-1} \Sigma_{\beta,\theta}'^{(t)}),$$

where

$$\Sigma'^{(t)} = (\phi_c \boldsymbol{R}'^{(t)})^{-1}.$$

The matrix $\Sigma'^{(t)}$ and the vector $\boldsymbol{\mu}'^{(t)}$ are indexed by θ and β to denote the respective subsets corresponding to these parameters. Once these conditional distributions have been obtained, we need to perform a recursion analysis similar to the conjugate normal-gamma updating and forecasting equations of Section 2.4. This analysis is conditional on ϕ_c, so that updating and forecasting corresponds to calculating parameters for posterior normal distributions rather than normal-gamma distributions. This needs to be carried out in order to calculate the parameters of the distributions of $(\boldsymbol{\theta}^{(t)} | \phi_c, \beta_c, \omega, D_t)$ for all t, from which we will obtain the desired joint posterior distribution.

The normal-based updating and forecasting recursions are straightforward. Suppose we have

$$(\boldsymbol{\theta}^{(t)} | \phi_c, \beta_c, \omega_c, D_{t-1}) \sim \mathrm{N}(\boldsymbol{m}_\theta^{(t)}, \mathbf{C}_\theta^{(t)}).$$

Then upon observing new data, d_t, the updating recursion obtains the distribution

$$(\boldsymbol{\theta}^{(t)} | \phi_c, \beta_c, \omega_c, D_t) \sim \mathrm{N}(\boldsymbol{m}_\theta'^{(t)}, \mathbf{C}_\theta'^{(t)}),$$

where

$$\begin{aligned} \mathbf{C}_\theta'^{(t)} &= (\mathbf{C}_\theta^{(t)\,-1} + \Sigma^{(t)\,-1})^{-1} \\ \boldsymbol{m}_\theta'^{(t)} &= \mathbf{C}_\theta'^{(t)} (\mathbf{C}_\theta^{(t)\,-1} \boldsymbol{m}_\theta^{(t)} + \Sigma^{(t)\,-1} \boldsymbol{\mu}^{(t)}), \end{aligned}$$

where $\Sigma^{(t)}$ and $\boldsymbol{\mu}^{(t)}$ are the conditional means and variances for the t-th competition conditional on β_c and ϕ_c.

The forecasting recursion which obtains the distribution of parameters before the $t+1$-th competition is given by

$$(\boldsymbol{\theta}^{(t+1)}|\phi_c, \beta_c, \omega_c, D_t) \sim \mathrm{N}(\boldsymbol{m}_\theta^{(t+1)}, \mathbf{C}_\theta^{(t+1)}),$$

where

$$\begin{aligned} \boldsymbol{m}_\theta^{(t+1)} &= \boldsymbol{m}_\theta^{\prime(t)} \\ \mathbf{C}_\theta^{(t+1)} &= \mathbf{C}_\theta^{\prime(t)} + \sigma^2 \mathbf{I}_p. \end{aligned}$$

These two recursions are performed alternately in succession to obtain the posterior distribution of $(\boldsymbol{\theta}^{(T)}|\phi_c, \beta_c, \omega_c, D_T)$.

To obtain a sample of draws of the $\boldsymbol{\theta}^{(t)}$, $t = T-1, \ldots, 1$, we first draw $\boldsymbol{\theta}_c^{(T)}$ from the distribution of $(\boldsymbol{\theta}^{(T)}|\phi_c, \beta_c, \omega_c, D_T)$. Then for $t = T-1, \ldots, 1$, we successively draw from

$$(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}_c^{(t+1)}, \ldots, \boldsymbol{\theta}_c^{(T)}, \phi_c, \beta, \omega_c, D_T) \sim \mathrm{N}(V^{(t)}(\mathbf{C}_\theta^{\prime(t)\,-1}\boldsymbol{m}_\theta^{\prime(t)} + \omega\boldsymbol{\theta}_c^{(t+1)}),\ V^{(t)}),$$

where $V^{(t)} = (\mathbf{C}_\theta^{\prime(t)\,-1} + \omega\mathbf{I}_p)^{-1}$. The result of this process is a single draw from the distribution of $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \beta, \phi|\omega, D_T)$.

Drawing from the distribution of $(\omega|\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}, \beta, \phi, D_T)$ is identical to (2.20) in Section 2.5. Only the p dynamic parameters of $\boldsymbol{\theta}^{(t)}$ are used to compute the parameters of the gamma distribution because the conditional distribution of ω does not depend on the non-dynamic components.

Non-iterative analysis

The addition of non-dynamic parameters to the analysis of Section 2.5.2 presents no obstacles. Here the non-iterative analysis requires the computation of $f(d_t|D_{t-1}, \omega)$, for all t, in order to obtain the posterior distribution of ω. This is accomplished by computing the normal-gamma parameters for the distribution of $(\boldsymbol{\theta}_*^{(t)}, \phi|\omega, D_{t-1})$ using the updating and forecasting equations (2.21) and (2.22), and calculating the values of the t-densities given these parameter values for each ω_i and each season t. No other modifications are necessary to the analysis of Section 2.5.2.

2.7 Marginal and Forecast Distributions

In practice, we are primarily interested in making inferences on $\boldsymbol{\theta}^{(T)}$ and on $\boldsymbol{y}^{(T+1)}$. This involves marginalizing the distribution of $(\boldsymbol{\theta}^{(T)}, \omega|D_T)$ or $(\boldsymbol{y}^{(T+1)}, \omega|D_T)$ over the nuisance parameter ω. In this section, we describe an approach using the results of Section 2.5 to marginalize over ω.

2.7.1 Marginal Distribution of $(\sigma|D_T)$

The methods of Sections 2.5.1 and 2.5.2 provide an approximate posterior distribution for $(\omega|D_T)$, and therefore for $(\sigma|D_T)$, as a discrete distribution. In the case of Section 2.5.1, the Gibbs sampler results in a random sample of m draws from the posterior distribution of $(\omega|D_T)$, whereas Section 2.5.2 results in a discrete probability distribution on the m values of ω which approximates the true continuous distribution of $(\omega|D_T)$. In both cases, we have a discrete distribution on a set of values of ω which approximates the continuous distribution of $(\omega|D_T)$. We assume, therefore, that

$$f(\omega|D_T) = \begin{cases} \pi_i^{(T)} & \text{for } \omega = \omega_i \\ 0 & \text{if } \omega \notin \{\omega_1, \ldots, \omega_m\} \end{cases} . \tag{2.23}$$

When $\{\omega_1, \ldots, \omega_m\}$ are drawn from the Gibbs sampler, we assume that $\pi_i^{(T)} = 1/m$ for all i. The approximate mean and variance of $(\sigma|D_T)$, where $\sigma_i = 1/\sqrt{\omega_i}$, are given by

$$\begin{aligned} \mathrm{E}(\sigma|D_T) &= \int \sigma f(\sigma|D_T)\, d\sigma = \sum_{j=1}^{m} \sigma_j \pi_j^{(T)}. \\ \mathrm{Var}(\sigma|D_T) &= \int (\sigma - \mathrm{E}(\sigma|D_T))^2 f(\sigma|D_T)\, d\sigma \\ &= \sum_{i=1}^{m} (\sigma_i - \mathrm{E}(\sigma|D_T))^2 \pi_i^{(T)}. \end{aligned}$$

Higher order moments can be calculated analogously.

2.7.2 Marginal Distribution of $\theta^{(T)}$

In order to make inferences on the distribution of $(\boldsymbol{\theta}^{(T)}|D_T)$, we need to marginalize over ω. After tournament T,

$$(\boldsymbol{\theta}^{(T)}, \phi|\omega, D_T) \sim \mathrm{NG}(\boldsymbol{\mu}'^{(T)}, \xi'^{(T)}, \boldsymbol{R}'^{(T)}, v'^{(T)}).$$

Marginalizing over ϕ gives

$$(\boldsymbol{\theta}^{(T)}|\omega, D_T) \sim T_{v'^{(T)}}(\boldsymbol{\mu}'^{(T)}, \xi'^{(T)} \boldsymbol{R}'^{(T)\,-1}).$$

Therefore,

$$\begin{aligned} f(\boldsymbol{\theta}^{(T)}|D_T) &= \int f(\boldsymbol{\theta}^{(T)}|\omega, D_T)\; f(\omega|D_T) d\omega \\ &= \sum_{i=1}^{m} \pi_i^{(T)} f(\boldsymbol{\theta}^{(T)}|\omega_i, D_T), \end{aligned}$$

which is a mixture of multivariate t densities.

The mean of this mixture distribution is computed as

$$\begin{aligned}\mathrm{E}(\boldsymbol{\theta}^{(T)}|D_T) &= \mathrm{E}(\mathrm{E}(\boldsymbol{\theta}^{(T)}|\omega, D_T)) \\ &= \mathrm{E}(\boldsymbol{\mu}^{(T)}) = \sum_{i=1}^{m} \pi_i^{(T)} \boldsymbol{\mu}_i^{(T)},\end{aligned}$$

where $\boldsymbol{\mu}_i^{(T)}$ is the mean of $(\boldsymbol{\theta}^{(T)}|\omega_i, D_T)$. To compute the variance, we first note that for multivariate t distributions,

$$\mathrm{Var}(\boldsymbol{\theta}^{(T)}|\omega_i, D_T) = \frac{v'^{(T)}}{v'^{(T)} - 2} \xi_i'^{(T)} \boldsymbol{R}_i'^{(T)\,-1},$$

where $\xi_i'^{(T)}$ and $\boldsymbol{R}_i'^{(T)}$ are indexed to show the dependence on ω_i. We therefore have

$$\begin{aligned}&\mathrm{Var}(\boldsymbol{\theta}^{(T)}|D_T) \\ &= \mathrm{E}(\mathrm{Var}(\boldsymbol{\theta}^{(T)}|\omega, D_T)) + \mathrm{Var}(\mathrm{E}(\boldsymbol{\theta}^{(T)}|\omega, D_T)) \\ &= \sum_{i=1}^{m} \pi_i^{(T)} \mathrm{Var}(\boldsymbol{\theta}^{(T)}|\omega_i, D_T) + \mathrm{Var}(\boldsymbol{\mu}^{(T)}) \\ &= \sum_{i=1}^{m} \pi_i^{(T)} \mathrm{Var}(\boldsymbol{\theta}^{(T)}|\omega_i, D_T) + \left(\sum_{i=1}^{m} \pi_i^{(T)} (\boldsymbol{\mu}_i^{(T)} - \mathrm{E}(\boldsymbol{\theta}^{(T)}|D_T))(\boldsymbol{\mu}_i^{(T)} - \mathrm{E}(\boldsymbol{\theta}^{(T)}|D_T))^{\mathsf{T}} \right)\end{aligned} \tag{2.24}$$

These computations allow us to summarize the first and second moments of the marginal distribution of $(\boldsymbol{\theta}^{(T)}|D_T)$.

Posterior inferences on $(\boldsymbol{\theta}^{(T)}|D_T)$ from the mixture distribution are difficult to perform analytically. Instead, we simulate draws from the mixture distribution, and make inferences empirically from the obtained sample. To simulate draws from the mixture, we can draw from $(\omega|D_T)$ first, and then given this value of ω we can draw from the exact multivariate t distribution of $(\boldsymbol{\theta}^{(T)}|\omega, D_T)$. This gives us a single draw from the marginal distribution of $(\boldsymbol{\theta}^{(T)}|D_T)$. To draw from a multivariate t-distribution, we can first draw from the gamma distribution of $(\phi|\omega, D_T)$, and then draw from the multivariate normal distribution of $(\boldsymbol{\theta}^{(T)}|\phi, \omega, D_T)$. In either case, this process is repeated until we obtain a sample large enough so that empirical inferences, such as credible intervals for parameters, can be obtained.

2.7.3 Forecast Distribution for d_{T+1}

We now show how to obtain forecast means and variances, and prediction intervals for unobserved data d_{T+1} of length $n^{(T+1)}$. We first note that from (2.16) in Section 2.4,

$$(d_{T+1}|\omega, D_T) \sim T_{v^{(T+1)}}(\boldsymbol{X}^{(T+1)} \boldsymbol{\mu}^{(T+1)}, \xi^{(T+1)}(\mathbf{I}_{n^{(T+1)}} + \boldsymbol{X}^{(T+1)} \boldsymbol{R}^{(T+1)\,-1} \boldsymbol{X}^{(T+1)\mathsf{T}})) \tag{2.25}$$

so that the predictive distribution for d_{T+1}

$$\begin{aligned} f(d_{T+1}|D_T) &= \int f(d_{T+1}|\omega, D_T) f(\omega|D_T)\, d\omega \\ &= \sum_{i=1}^{m} \pi_i^{(T)} f(d_{T+1}|\omega_i, D_T) \qquad (2.26) \end{aligned}$$

is again a mixture of t distributions. The mean is computed as

$$\begin{aligned} &\mathrm{E}(d_{T+1}|D_T) \\ &= \mathrm{E}(\mathrm{E}(d_{T+1}|\omega, D_T)) \\ &= \sum_{i=1}^{m} \pi_i^{(T)} \boldsymbol{X}^{(T+1)} \boldsymbol{\mu}_i^{(T+1)} = \boldsymbol{X}^{(T+1)} (\sum_{i=1}^{m} \pi_i^{(T)} \boldsymbol{\mu}_i^{(T+1)}), \end{aligned}$$

where $\boldsymbol{X}^{(T+1)}$ is the design matrix for the new data, and $\boldsymbol{\mu}_i^{(T+1)}$ is the mean of the normal-gamma distribution that forecasts $\boldsymbol{\theta}^{(T+1)}$ from data through time T. For computing the variance, we first note that

$$\mathrm{Var}(d_{T+1}|\omega_i, D_T) = \frac{v^{(T+1)}}{v^{(T+1)} - 2} \xi_i^{(T+1)} (\mathbf{I}_{n(T+1)} + \boldsymbol{X}^{(T+1)} \boldsymbol{R}_i^{(T+1)\,-1} \boldsymbol{X}^{(T+1)\mathsf{T}})$$

so that

$$\begin{aligned} &\mathrm{Var}(d_{T+1}|D_T) \\ &= \mathrm{E}(\mathrm{Var}(d_{T+1}|\omega, D_T)) + \mathrm{Var}(\mathrm{E}(d_{T+1}|\omega, D_T)) \\ &= \sum_{i=1}^{m} \pi_i^{(T)} \mathrm{Var}(d_{(T+1)}|\omega_i, D_T) + \mathrm{Var}(\boldsymbol{X}^{(T+1)} \boldsymbol{\mu}^{(T+1)}) \\ &= \sum_{i=1}^{m} \pi_i^{(T)} \mathrm{Var}(d_{(T+1)}|\omega_i, D_T) + \boldsymbol{X}^{(T+1)} \mathrm{Var}(\boldsymbol{\mu}^{(T+1)}) \boldsymbol{X}^{(T+1)\,\mathsf{T}} \\ &= \sum_{i=1}^{m} \pi_i^{(T)} \mathrm{Var}(d_{(T+1)}|\omega_i, D_T) \\ &\quad + \boldsymbol{X}^{(T+1)} \left(\sum_{i=1}^{m} \pi_i^{(T)} (\boldsymbol{\mu}_i^{(T+1)} - \mathrm{E}(\boldsymbol{\theta}^{(T)}|D_T))(\boldsymbol{\mu}_i^{(T+1)} - \mathrm{E}(\boldsymbol{\theta}^{(T)}|D_T))^{\mathsf{T}} \right) \boldsymbol{X}^{(T+1)\,\mathsf{T}}. \end{aligned}$$

$$(2.27)$$

Thus calculations for the estimated mean and variance are straightforward from these formulas.

We obtain inferences from the predictive distribution in (2.26) by drawing random samples of $\boldsymbol{y}^{(T+1)}$ and reporting empirically based predictive intervals. We draw from $(\omega|D_T)$, and then given this value we can draw from the exact multivariate t distribution of $f(d_{T+1}|\omega, D_T)$ as given in (2.25). This results in a single draw from the marginal distribution. Repeated applications of this process yield a sample on which to perform empirically based inferences.

It should be noted that treating ω, and therefore σ^2, as an unknown parameter appropriately increases the estimated variability in both $(\boldsymbol{\theta}^{(T)}|D_T)$ and $(d_{T+1}|D_T)$. This is reflected by the inclusion in (2.24) and (2.27) of the second term, the variance of the conditional mean. If ω were treated as known, this second term would vanish, and the variability of the forecasts would be underestimated.

2.8 Sequential Updating with New Data

In this section we consider the situation in which we have observed T tournaments of paired comparison data and have performed one of the analyses of this chapter thereby obtaining the distribution of $(\boldsymbol{\theta}^{(T)}, \phi|\omega, D_T)$ and the distribution of $(\omega|D_T)$. We now observe data at tournament $T+1$ and want updated distributions of $(\boldsymbol{\theta}^{(T+1)}, \phi|\omega, D_{T+1})$ and $(\omega|D_{T+1})$.

Naturally, one method that gives the desired result is to apply the Gibbs sampler methodology of Section 2.5.1 by performing the entire Gibbs sampler analysis on all $T+1$ tournaments. This will give the desired posterior distributions, but it is extremely inefficient in that it does not make use of the analysis of the first T tournaments.

A more efficient approach is based on a non-iterative updating algorithm similar to the method of Section 2.5.2. In the analysis of Section 2.5.2, we update the distribution of $(\omega|D_0)$ to $(\omega|D_T)$ by reweighting the initial probabilities, $\pi_i^{(0)}$, by the product of marginal likelihoods of the data to obtain the posterior probabilities, $\pi_i^{(T)}$. The analysis here is similar, as we want to update the distribution of $(\omega|D_T)$ to $(\omega|D_{T+1})$ from data observed at tournament $T+1$.

We have at time T

$$(\boldsymbol{\theta}^{(T)}, \phi|\omega, D_T) \sim \mathrm{NG}(\boldsymbol{\mu}'^{(T)}, \xi'^{(T)}, \boldsymbol{R}'^{(T)}, v'^{(T)}),$$

and we also have a sample of m values, $\omega_1, \ldots, \omega_m$, for which

$$f(\omega|D_T) = \begin{cases} \pi_i^{(T)} & \text{if } \omega = \omega_i \\ 0 & \text{if } \omega \notin \{\omega_1, \ldots, \omega_m\} \end{cases}$$

Obtaining the distribution of $(\boldsymbol{\theta}^{(T+1)}, \phi|\omega, D_{T+1})$ is a simple application of the forecasting and updating formulas given in Section 2.4. To calculate the posterior probabilities, $\pi_i^{(T+1)}$, for the distribution of $(\omega|D_{T+1})$, we have

$$\begin{aligned} f(\omega|D_{T+1}) &= \frac{f(d_{T+1}|D_T, \omega)}{f(d_{T+1}|D_T)} f(\omega|D_T) \\ &\propto f(d_{T+1}|D_T, \omega)\, f(\omega|D_T). \end{aligned}$$

The posterior distribution of $(\omega|D_{T+1})$ is given by

$$f(\omega|D_{T+1}) = \begin{cases} \pi_i^{(T+1)} & \text{if } \omega = \omega_i \\ 0 & \text{if } \omega \notin \{\omega_1, \ldots, \omega_m\} \end{cases}$$

with

$$\pi_i^{(T+1)} = \frac{\pi_i^{(T)} f(d_{T+1}|\omega_i, D_T)}{\sum_{j=1}^m \pi_j^{(T)} f(d_{T+1}|\omega_j, D_T)}.$$

Thus the updating of the discrete posterior distribution from $\pi_i^{(T)}$ to $\pi_i^{(T+1)}$ is obtained by reweighting by the relative likelihoods of the new data, which have a multivariate t density. This approach appears to be an efficient way to obtain the approximate posterior distribution of $(\omega|D_{T+1})$ as it makes use of the previous computation that resulted in the posterior distribution of $(\omega|D_T)$. The methods of Section 2.7 can be used to obtain the marginal posterior distribution of $(\boldsymbol{\theta}^{(T+1)}|D_{T+1})$ as well as forecast distributions for subsequent tournaments.

Chapter 3

Analysis of NFL Football Game Scores

Using the methodology described in Chapter 2, we analyze in this chapter National Football League (NFL) game outcomes from 1981–1991. As a result of our analysis, we obtain rankings of teams at the end of the 1991 season and game predictions for the 1992 season. In Section 3.1 we describe a probability model for football scores that incorporates changes in team abilities over time, and the implementation of both the Gibbs sampler analysis and non-iterative analysis of the data under the model. Section 3.2 reports the results of the analyses of the model. We obtain the posterior distribution of the between-season variance, the parameter describing the magnitude of the team ability changes from season to season. We also compute posterior distribution summaries and inferences for team ratings, and obtain prediction intervals for 1992 NFL games. To test the adequacy of the model, we perform diagnostic checks on the forecast residuals. In Section 3.3, we compare the results of analyses based on different amounts of data. We demonstrate in Section 3.4 how to update the distribution of the model parameters to account for new data (the first four weeks of 1992 NFL game results). We conclude the chapter in Section 3.5 with a discussion of the utility of the model.

The data used in the analysis consists of the complete regular season results of NFL football games for the years 1981, 1983–1986, 1988–1991. We did not include games for the years 1982 and 1987 because these were years in which strikes occurred; games played during these years may not be representative of how teams would perform under usual conditions. The NFL is comprised of 28 teams, and during the regular season each team plays 16 games, resulting in a total of 224 games played per season. For each game, we recorded the final score for each team, and the site of the game. Use of covariate information could likely improve the predictive power of the model (game statistics, for example), but no additional information was recorded.

3.1 A Model for NFL Score Differences

The probability model that we assume for differences in final scores during season t, $t = 1981, \ldots, 1991$, is given by

$$y^{(t)}_{i_k j_k k} \sim \mathrm{N}(\theta^{(t)}_{i_k k} - \theta^{(t)}_{i_k k} + \delta^{(t)}_i \theta_{(\mathrm{hfa})}, \tau^2), \tag{3.1}$$

where, at time t,

$$\begin{aligned}
y^{(t)}_{i_k j_k k} &= \text{the score difference of game } k \\
\theta^{(t)}_i &= \text{rating parameter for team } i \\
\theta_{(\mathrm{hfa})} &= \text{home field advantage parameter} \\
i_k, j_k &= \text{indices of teams involved in the } k\text{-th game} \\
\delta^{(t)}_i &= \begin{cases} 1 & \text{if team } i_k \text{ plays on their home field} \\ -1 & \text{if team } j_k \text{ plays on their home field} \end{cases} \\
\tau^2 &= \text{variance of } y^{(t)}_{i_k j_k k} \text{ conditional on the mean}
\end{aligned}$$

The team parameters, $\theta^{(t)}_i$, $i = 1, \ldots, 28$, indicate the relative scoring abilities of the teams during season t. The difference $\theta^{(t)}_{i_k} - \theta^{(t)}_{j_k}$ is the score difference expected to occur between teams i_k and j_k during season t, not accounting for the site of the game. The parameter $\theta_{(\mathrm{hfa})}$, the home-field advantage (HFA) parameter, specifies the amount on average the home team will outscore the visiting team, and is assumed constant for all teams and seasons. The variance of the score difference about the mean, τ^2, is also assumed independent of teams involved in a game and of season.

The model in (3.1) is intended only as an approximation. NFL games scores can take on only integer values, so the assumption of normality conditional on the parameter values cannot be considered exact. Modeling football game score differences as discrete outcomes may be more realistic, but the analysis of such a model may prove to be intractable. An example of such an analysis appears in Rosner (1976). The normal approximation allows for the use of the methodology developed in Chapter 2.

For the NFL data, we assume a model for the evolution of the parameters given by

$$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} + \boldsymbol{\nu}^{(t)},$$

where $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\theta}^{(t-1)}$ denote the vectors of 28 team parameters for seasons t and $t-1$, respectively, and $\boldsymbol{\nu}^{(t)}$ is the amount of change incurred in the parameter values between seasons $t-1$ and t. We assume for the analysis of NFL game scores that

$$\boldsymbol{\nu}^{(t)} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

The model assumes that team abilities change independently from season to season, the expected change being 0, and the variance constant for all teams. The HFA parameter is assumed to remain constant over time.

The model we choose for the evolution of team parameters suggests that teams with low abilities undergo roughly the same amount of change in ability from season $t-1$ to t as teams with higher abilities. This assumption is probably not quite right because poorly performing teams at the end of season $t-1$ have greater chance for improvement, for example, by obtaining better players in the player selection draft before the following season. Thus the mean change from season to season may be a function of the team parameter values or records of the previous season. Furthermore, the mean change from season to season may be a function of other covariates that are difficult to quantify, such as player trades and coaching staff changes. The model as postulated also assumes that teams do not undergo ability changes during a season, but instead only change between seasons. While team abilities do change during a season, the magnitude of the change will likely be small relative to the innovations occurring between seasons, so the simplifying assumption of constant ability within a season is not unreasonable.

3.1.1 Prior Parameters

We use a vague prior distribution on the parameters to reflect our initial uncertainty. This choice allows the likelihood of the data to determine the distribution of the parameters. The normal-gamma parameters for the distribution of $(\boldsymbol{\theta}^{(1981)}, \theta_{(\mathrm{hfa})})$ are

$$\begin{aligned}
\boldsymbol{\mu}^{(1981)} &= (0,\ldots,0,3) \\
\xi^{(1981)} &= 100 \\
\boldsymbol{R}^{(1981)} &= \mathbf{I}_{29} \\
v^{(1981)} &= .5.
\end{aligned}$$

The prior means for the team parameters are all zero, that is, no team is assumed a priori better than another, and the mean of the HFA parameter equivalent to a 3 point advantage for the home team. The variance of the prior means is assumed to be 100, which corresponds to a large amount of uncertainty. The initial uncertainty of the observation variance estimate of 100 is reflected by only .5 degrees of freedom. The harmonic mean variance of a score difference is assumed to be 100.

We assumed a vague prior for the distribution of $\omega = 1/\sigma^2$ with improper density

$$f(\omega) = \frac{1}{\omega}.$$

which corresponds to a prior distribution of $1/\sigma$ on σ. This choice gives greater weight a priori to smaller values of σ^2. This prior distribution on ω is sufficiently vague to allow the likelihood to strongly dominate.

3.1.2 Model Implementation

In this section, we describe the implementation of the iterative and non-iterative analyses. The model accounts for the lack of 1982 and 1987 NFL data by assuming the parameter means do not change for these years, while the team parameter variances increase by σ^2. When performing the recursive iterations of updating and forecasting as described in Section 2.4, the updating computation is omitted for 1982 and 1987. Otherwise, the computation proceeds as usual.

Gibbs Sampler Implementation

We performed an analysis of the model for NFL data using the Gibbs sampler techniques of Section 2.5.1 with the extension for non-dynamic covariate information described in Section 2.6. We ran three parallel Gibbs samplers with overdispersed starting values of ω: $1/100^2$, $1/5^2$, and $1/.2^2$. Each Gibbs sampler proceeds by drawing the parameters $\theta_{(\text{hfa})}, \boldsymbol{\theta}^{(1981)}, \ldots, \boldsymbol{\theta}^{(1991)}, \tau^2$, given the initial value of ω, then drawing ω conditional on these parameter values, and so on. This was continued for 450 iterations for each sampler. Convergence of the algorithm was checked by examining the potential scale reduction of ω, as described in Gelman and Rubin (1992a). The potential scale reduction for ω is an estimate of the factor by which the variance of the current distribution of ω in the Gibbs sampler will decrease with continued iterations. Values of the potential scale reduction near 1 are indicative of convergence. For the NFL model, it suffices to examine the potential scale reduction for the parameter ω, because once the Gibbs sampler has reached the stationary distribution for ω, the distribution of the remaining parameters is completely specified. Transforming ω by taking logarithms to make the distribution less skewed, the potential scale reduction for ω is computed to be 1.027. This value indicates that continued iterative simulation would not lead to improved inferences. We therefore conclude 450 iterations is sufficient to reach the target distribution. Each of the samplers were continued 50 more iterations to produce 150 more draws of ω, and then a random subsample of 100 of these were chosen to be the final sample drawn from the marginal posterior distribution of ω. The sample values are then reexpressed as values from the posterior distribution σ by setting $\sigma = 1/\sqrt{\omega}$.

Non-iterative Analysis

We also analyzed the model using the non-iterative methods of Section 2.5.2, allowing for the inclusion of a non-dynamic covariate as described in Section 2.6. We approximated the prior distribution of σ to be

$$f(\sigma) \propto \frac{1}{\sigma},$$

where σ has been discretized and takes only the 20 equally spaced values $(2.0, 2.16, \ldots, 5.0)$. This set of values appears to be overdispersed for the posterior distribution of σ, as indicated by the Gibbs sampler analysis. In the absence of such information, a wider range of values would be required. Thus we have comparable prior distributions for the Gibbs sampler and non-iterative analyses. We obtain the approximate posterior distribution of σ using the techniques of Section 2.5.2 by computing the reweighting factors, $f(\boldsymbol{y}^{(1992)}|\omega_i, D_{1991}) \cdots f(\boldsymbol{y}^{(1981)}|\omega_i, D_{1980})$.

3.2 Results

We discuss in this section the results of the iterative simulation analysis and the non-iterative analysis. We examine the marginal posterior distributions of σ and of $\boldsymbol{\theta}^{(1991)}$ along with $\theta_{(\mathrm{hfa})}$, and compute forecast distributions of 1992 results. We also perform model diagnostics to assess the validity of the model.

3.2.1 Marginal Posterior Distribution of σ

Figure 3.1 displays the distribution of σ, the system standard deviation, for both the iterative simulation analysis and the non-iterative analysis. The top histogram shows the posterior distribution based on the sample of 100 draws from the Gibbs sampler, while the bottom histogram shows the posterior distribution calculated from the grid of 20 equally spaced values. The similarity of the distributions indicates general agreement of the methodologies in producing comparable results. As we show in subsequent analyses, the slight difference in the approximated distributions for σ does not appear to have large effect on the marginal distribution of the team parameters or on predictive distributions.

The estimated posterior mean and standard deviation of σ, the system standard deviation of the model, calculated from the iterative analysis are 3.24 and 0.304, respectively. From the non-iterative analysis, the mean is estimated to be 3.16 and the standard deviation is 0.307. Thus the results are consistent. An interpretation of this estimate is that, between seasons, changes in teams' scoring capabilities relative to other teams have a standard deviation of about 3 points. Teams that have ratings within a few points of one another at the end of one season are not unlikely to switch positions in the following season.

3.2.2 Marginal Posterior Distribution of Ratings and HFA parameters

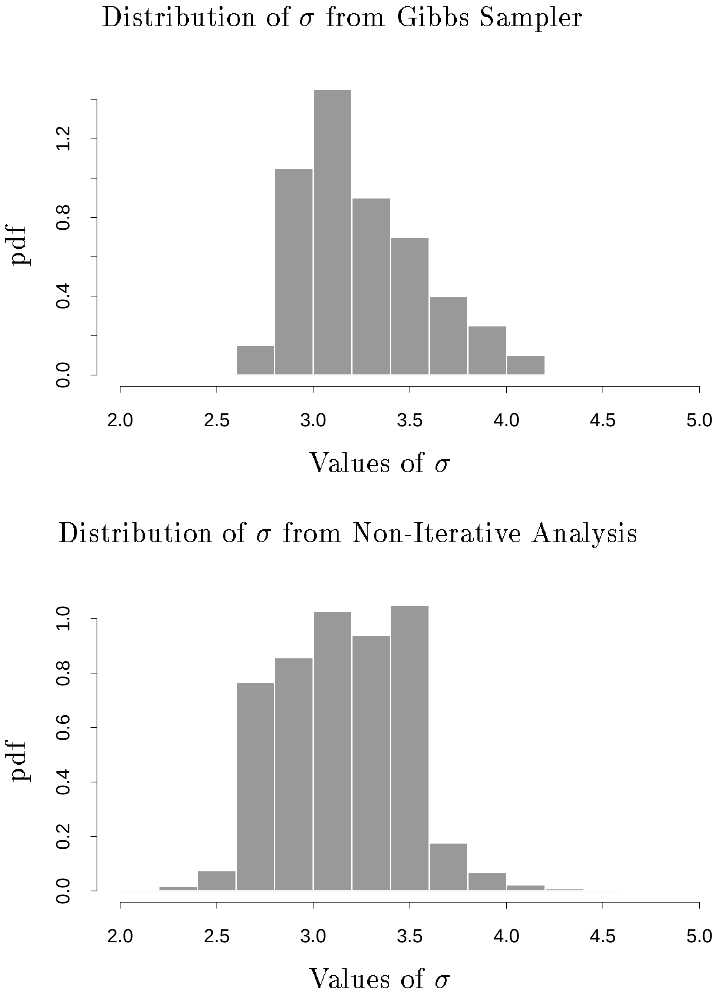
To obtain the marginal posterior distribution of $\boldsymbol{\theta}^{(1991)}$ and $\theta_{(\mathrm{hfa})}$, we used the techniques discussed in Section 2.7. Marginal posterior means and standard errors of $\boldsymbol{\theta}^{(1991)}$ are with $\pi_i = 1/100$ for all $i = 1, \ldots, 100$ in the Gibbs sampler analysis, and with π_i equal to the posterior probabilities given in the bottom histogram in Figure 3.1. We also compute estimated 95% credible intervals by drawing 3000 samples from the posterior distribution of $\boldsymbol{\theta}^{(1991)}$ and reporting the approximate 2.5% and 97.5% percentiles for each parameter. These values for the Gibbs sampler analysis are summarized in Table 3.1 and for the non-iterative analysis in Table 3.2.

Parameter	Mean	Std Dev	95% Credible Interval
Washington Redskins	11.45	4.04	(3.81, 19.54)
San Francisco 49ers	9.47	4.01	(1.01, 17.47)
Houston Oilers	6.15	4.02	(−1.79, 14.08)
New Orleans Saints	5.85	4.02	(−2.04, 13.70)
Buffalo Bills	5.66	4.02	(−2.34, 13.59)
Kansas City Chiefs	5.29	4.02	(−2.88, 13.26)
Philadelphia Eagles	4.02	4.02	(−3.88, 12.05)
New York Giants	2.57	4.02	(−5.55, 10.45)
Denver Broncos	2.47	4.02	(−5.47, 10.15)
Los Angeles Raiders	2.24	4.02	(−5.74, 10.15)
Chicago Bears	2.01	4.02	(−6.16, 10.17)
Seattle Seahawks	1.07	4.02	(−6.58, 9.03)
Atlanta Falcons	0.52	4.02	(−7.35, 8.61)
Detroit Lions	0.11	4.02	(−7.80, 8.07)
Dallas Cowboys	0.08	4.03	(−7.83, 8.18)
Minnesota Vikings	0.02	4.03	(−7.97, 8.43)
Miami Dolphins	−1.28	4.02	(−9.43, 6.61)
San Diego Chargers	−1.51	4.02	(−9.34, 6.69)
Pittsburgh Steelers	−1.82	4.01	(−9.77, 6.46)
Cleveland Browns	−2.84	4.01	(−10.63, 5.30)
New York Jets	−2.87	4.02	(−10.85, 5.42)
Green Bay Packers	−3.77	4.02	(−11.59, 4.06)
Los Angeles Rams	−4.27	4.02	(−12.42, 3.47)
Cincinnati Bengals	−4.89	4.03	(−12.97, 2.80)
Phoenix Cardinals	−6.59	4.02	(−14.59, 1.46)
Tampa Bay Buccaneers	−8.88	4.02	(−16.82, −0.89)
New England Patriots	−8.96	4.02	(−16.86, −0.64)
Indianapolis Colts	−11.29	4.05	(−19.13, −3.40)
Home Field Advantage	2.96	0.29	(2.41, 3.53)

Table 3.1: Distribution of $\theta^{(1991)}$ from Gibbs Sampler Analysis

Parameter	Mean	Std Dev	95% Credible Interval
Washington Redskins	11.33	4.02	(3.41, 19.38)
San Francisco 49ers	9.45	3.98	(1.61, 17.19)
Houston Oilers	6.07	4.00	(−1.89, 13.87)
New Orleans Saints	5.79	3.99	(−2.21, 13.69)
Buffalo Bills	5.66	3.99	(−2.34, 13.72)
Kansas City Chiefs	5.25	4.00	(−2.51, 13.31)
Philadelphia Eagles	4.01	3.99	(−3.64, 12.11)
New York Giants	2.61	4.00	(−5.34, 10.57)
Denver Broncos	2.46	3.99	(−5.53, 10.29)
Los Angeles Raiders	2.24	3.99	(−5.69, 9.98)
Chicago Bears	2.02	3.99	(−5.99, 9.65)
Seattle Seahawks	1.06	3.99	(−6.43, 9.24)
Atlanta Falcons	0.45	3.99	(−7.39, 8.41)
Minnesota Vikings	0.08	4.00	(−7.94, 7.97)
Detroit Lions	0.06	3.99	(−7.99, 7.56)
Dallas Cowboys	0.00	4.01	(−7.91, 7.99)
Miami Dolphins	−1.25	3.99	(−9.18, 6.54)
San Diego Chargers	−1.50	3.99	(−9.20, 6.78)
Pittsburgh Steelers	−1.80	3.98	(−9.95, 6.29)
Cleveland Browns	−2.84	3.98	(−11.01, 4.87)
New York Jets	−2.89	3.99	(−10.67, 4.85)
Green Bay Packers	−3.76	3.99	(−11.92, 3.89)
Los Angeles Rams	−4.19	4.00	(−11.95, 3.80)
Cincinnati Bengals	−4.78	4.01	(−12.74, 3.57)
Phoenix Cardinals	−6.58	3.99	(−14.37, 1.35)
Tampa Bay Buccaneers	−8.85	3.99	(−16.73, −1.06)
New England Patriots	−8.93	3.99	(−17.18, −1.22)
Indianapolis Colts	−11.17	4.03	(−19.68, −3.43)
Home Field Advantage	2.96	0.29	(2.38, 3.53)

Table 3.2: Distribution of $\theta^{(1991)}$ from Non-Iterative Analysis


Distribution of σ from Gibbs Sampler

Distribution of σ from Non-Iterative Analysis

Figure 3.1: Posterior Distribution of σ

Tables 3.1 and 3.2 rank the teams in order of the posterior mean of $\boldsymbol{\theta}^{(1991)}$. Both analyses produce the same rankings, with the exception of the Vikings, the Lions and the Cowboys. In the Gibbs sampler analysis, the Vikings are ranked below the Lions and the Cowboys, while in the non-iterative analysis the Vikings are ranked higher. In both analyses, these three teams have posterior means estimated to be within 0.1 points of each other. the results still appear consistent in spite of the reordering of the teams. The means of the team parameters for each analysis are within 0.2 points of each other and the standard deviations match even more closely. The standard deviations in the Gibbs sampler analysis are consistently larger than those in the non-iterative analysis, but size of the difference suggests that the cause may be attributed extreme values of the posterior distribution not being represented in the small sample obtained in the Gibbs sampler.

The means can be best interpreted as differences. For example, from Table 3.1, the value of the means suggest that we would expect at the end of the 1991 season the Redskins to beat the 49ers by about 2 points, excluding the effect of home field. A more extreme example would be that we would expect the Redskins to outscore the Colts by about 23 points on average.

The standard deviations of the rating parameters are close to 4, indicating substantial variability in the parameter distributions. Posterior correlations among the rating parameters are about 0.6, so that the standard deviation for differences in rating parameters is close to 3.7. The data therefore provide more evidence about differences in parameters than the individual team ratings.

The HFA parameter has a posterior mean of about 2.96 with a small standard error. This can be interpreted as giving the home team approximately a 3-point advantage. The standard error for the HFA parameter is small because it is assumed to remain constant over time, and all of the data is used to estimate the posterior distribution of $\theta_{(\text{hfa})}$, so that unlike the team parameters there is no extra uncertainty incurred over time.

3.2.3 Forecast Distribution of Score Differences

To illustrate the techniques described in Section 2.7, we obtain the forecast distribution of selected game outcomes for the 1992 season. For the first 20 games of the season, we computed the predicted score differences and standard errors. In addition, we simulated 3000 draws from the approximate predictive distribution of $\boldsymbol{y}^{(1992)}$ to obtain 50% prediction intervals for game score differences, and estimates for the probability of the first team in a pair winning. Winning probabilities were computed by counting the fraction of the 3000 simulated game outcomes that produced score differences greater than 0.

Tables 3.3 and 3.4 show the resulting analysis. Again, the comparison between the two analyses lead to very similar results. For all 20 games, the predicted score results are within 0.2 between analyses. The 50% prediction intervals, which are produced by simulation, match reasonably closely suggesting that the different distributions of σ do not have a large impact on the resulting predictions.

It is worth noting that the 50% prediction intervals are quite wide. We also computed 95% intervals (not shown) which spanned about eight touchdowns. Despite the widths, the intervals appear to be honest, as 8 observed score differences fall in the intervals for both the Gibbs sampler analysis and for the non-iterative analysis. The probabilities computed from simulations of the predictive distribution do not seem to predict substantially better than guessing at random in this small sample; from the non-iterative analysis, the log-likelihood (base 10) for the predictive probabilities is -5.91, whereas the corresponding log-likelihood of guessing at random is $\log(.5^{20}) = -6.02$. A more thorough analysis of this model indicates that the model predictions are significantly better than chance, and better than competing models (Stern 1992a).

1992	Predicted	Actual	50% Prediction	Probability
Games	Score Difference	Score Difference	Interval	Win
PHA vs. NO	1.14	2	(−7.88, 11.65)	0.55
BUF vs. LAN	12.89	33	(3.59, 23.14)	0.82
CHI vs. DET	4.86	3	(−4.24, 14.14)	0.64
ATL vs. NYJ	6.35	3	(−2.54, 16.14)	0.69
HOU vs. PIT	10.93	−5	(1.43, 20.20)	0.78
MIN vs. GB	0.84	3	(−8.70, 10.24)	0.51
CLE vs. IND	5.49	−11	(−3.86, 14.93)	0.65
SF vs. NYG	3.94	17	(−5.50, 14.16)	0.61
KC vs. SD	3.84	14	(−6.12, 13.20)	0.60
SEA vs. CIN	8.92	−18	(−0.35, 18.48)	0.74
TB vs. PHX	0.66	16	(−9.11, 10.68)	0.52
DEN vs. LAA	3.19	4	(−6.38, 12.41)	0.60
WAS vs. DAL	8.40	−13	(−1.24, 18.26)	0.72
LAN vs. NE	7.65	14	(−2.54, 17.15)	0.69
WAS vs. ATL	13.89	7	(4.46, 23.19)	0.84
DAL vs. NYG	−5.45	6	(−14.90, 4.06)	0.34
NO vs. CHI	6.80	22	(−2.46, 16.86)	0.70
LAA vs. CIN	4.17	−3	(−5.54, 13.48)	0.62
DET vs. MIN	3.04	14	(−6.71, 12.78)	0.59
KC vs. SEA	7.18	19	(−2.09, 16.97)	0.70

Table 3.3: Forecasts of 1992 Games from Gibbs Sampler Analysis

1992			Predicted	Actual	50% Prediction	Probability
Games			Score Difference	Score Difference	Interval	Win
PHA	vs.	NO	1.19	2	(–7.89, 10.68)	0.53
BUF	vs.	LAN	12.81	33	(3.12, 22.03)	0.81
CHI	vs.	DET	4.92	3	(–4.98, 14.78)	0.64
ATL	vs.	NYJ	6.30	3	(–3.16, 15.95)	0.68
HOU	vs.	PIT	10.84	–5	(1.78, 20.69)	0.78
MIN	vs.	GB	0.88	3	(–8.81, 9.64)	0.51
CLE	vs.	IND	5.36	–11	(–3.94, 15.14)	0.65
SF	vs.	NYG	3.87	17	(–5.82, 13.37)	0.60
KC	vs.	SD	3.79	14	(–5.79, 13.23)	0.61
SEA	vs.	CIN	8.80	–18	(–0.97, 18.24)	0.73
TB	vs.	PHX	0.70	16	(–9.18, 9.87)	0.51
DEN	vs.	LAA	3.19	4	(–6.20, 12.84)	0.59
WAS	vs.	DAL	8.37	–13	(–1.84, 17.87)	0.72
LAN	vs.	NE	7.70	14	(–1.81, 17.60)	0.70
WAS	vs.	ATL	13.84	7	(3.69, 23.45)	0.82
DAL	vs.	NYG	–5.58	6	(–14.94, 3.80)	0.35
NO	vs.	CHI	6.72	22	(–2.56, 16.86)	0.68
LAA	vs.	CIN	4.06	–3	(–5.71, 13.96)	0.61
DET	vs.	MIN	2.95	14	(–6.50, 12.28)	0.58
KC	vs.	SEA	7.15	19	(–2.57, 16.80)	0.70

Table 3.4: Forecasts of 1992 Games from Non-Iterative Analysis

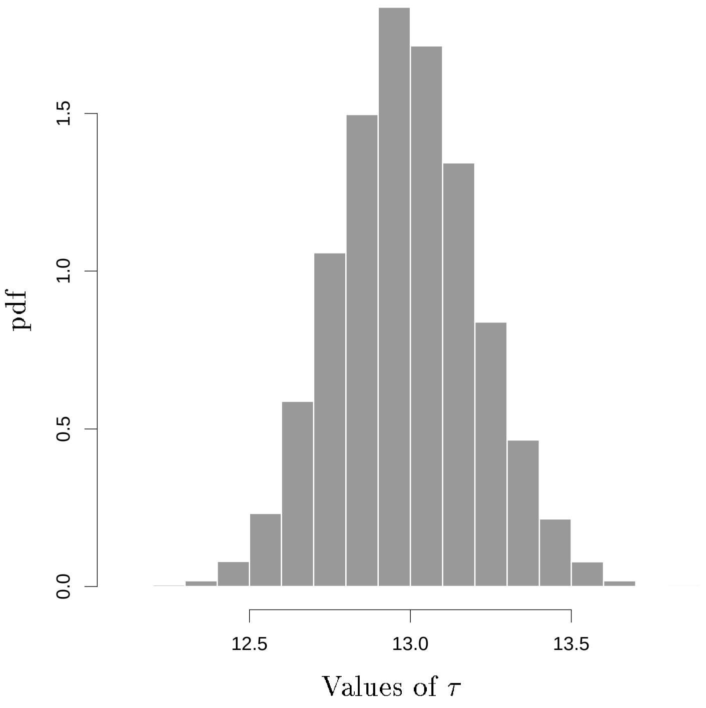


Figure 3.2: Distribution of τ from Non-Iterative Analysis

3.2.4 Distribution of the Observation Variance

The distribution of τ, the observation standard deviation, can be found from the non-iterative analysis by drawing from the posterior distribution of ω, and then subsequently drawing $\phi = 1/\tau^2$, the observation precision, from the gamma distribution conditional on ω. We simulated 10000 draws from the posterior distribution of τ and show the resulting distribution in Figure 3.2. The distribution of τ appears to be symmetric with mean and standard deviation are 13.0 and 0.218 points, respectively. With such a small standard deviation, the parameter τ is well estimated.

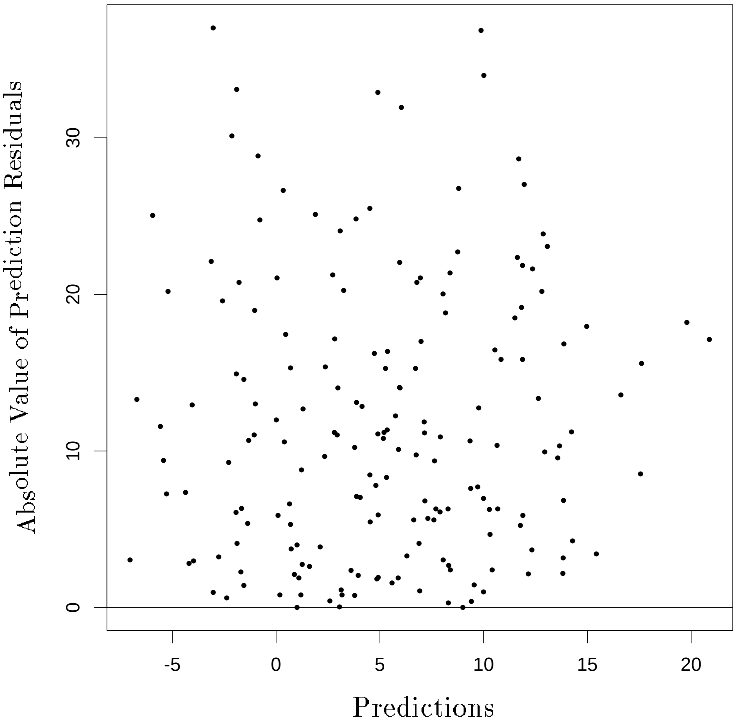


Figure 3.3: Residuals versus Absolute Predicted Score Differences

3.2.5 Diagnostics

We examined the forecast residuals of games played in the 1992 season using the results from the non-iterative analysis. Figure 3.3 shows a plot of residuals against absolute values of predicted score differences for the first 181 games of the 1992 season (those available at the time of the analysis). The ordering of teams is arbitrary, so the sign of the residuals are not meaningful. The plot shows random scatter with no clearly emerging pattern. The plot suggests the assumption of constant variance with respect to predicted values seems appropriate. A plot of the absolute residuals against the sum of team scores for the 181 games is shown in Figure 3.4. We might expect that the residuals would increase in variability with larger sums of scores but Figure 3.4 shows no substantial increase in variability in the residuals.

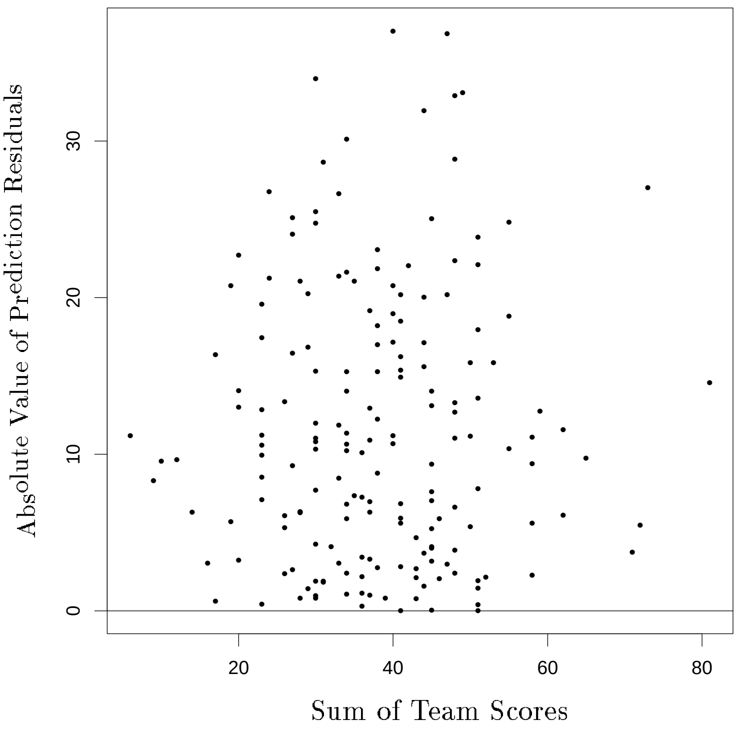


Figure 3.4: Forecast Residuals against Sum of Scores

We checked the assumption of normality by examining the absolute value of residuals from the 1991 game results. A half-normal probability plot of the residuals is shown in Figure 3.5. From this plot, the residuals appear to be slightly light-tailed, although it is difficult to argue that the distribution differs for practical purposes from a normal distribution.

3.3 The Effect of Additional Seasons

We compare the variability of the parameters when different amounts of data are incorporated into the analysis. We consider three analyses; the first using data through 1984,

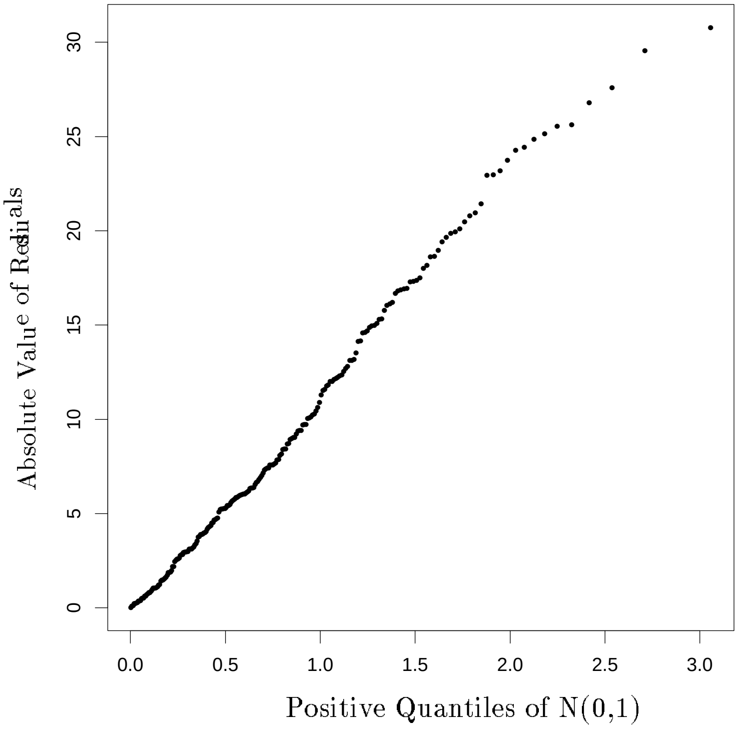


Figure 3.5: Normal Probability Plot of 1991 Residuals

the second using data through 1988, and finally the complete analysis using data through 1991. Figure 3.6 shows the distribution of σ, the innovation standard deviation, computed from the non-iterative analysis in each case. The figure shows clearly that the distribution of σ becomes less variable as a greater amount of data is used in the analysis. The approximate means and standard deviations, respectively, for each of the distributions are 3.56 and 0.601 for data through 1984, 3.17 and 0.386 for data through 1988, and 3.16 and 0.307 for data through 1991.

The variability of predictions using the three different distributions of σ are comparable. Table 3.5 shows the approximate 95% prediction intervals for hypothetical games taking place at the beginning of the 1985, 1989 and 1992 seasons, respectively, using the three posterior distributions of σ shown in Figure 3.6. The teams are those involved in the first 20 games of the 1992 season. We computed the approximate prediction intervals

	Widths of Approximate 95% Prediction Intervals		
1992 Games	Data through 1984	Data through 1988	Data through 1991
PHA vs. NO	54.63	57.11	54.09
BUF vs. LAN	56.54	57.36	55.93
CHI vs. DET	53.52	54.95	56.62
ATL vs. NYJ	55.91	55.26	53.64
HOU vs. PIT	55.76	55.40	55.69
MIN vs. GB	55.40	54.72	55.10
CLE vs. IND	55.67	56.54	55.58
SF vs. NYG	55.27	55.99	53.82
KC vs. SD	55.40	56.46	55.12
SEA vs. CIN	55.43	56.15	55.38
TB vs. PHX	55.64	55.90	55.33
DEN vs. LAA	54.82	56.62	56.47
WAS vs. DAL	55.53	56.39	55.00
LAN vs. NE	53.52	56.16	57.19
WAS vs. ATL	55.95	56.34	57.12
DAL vs. NYG	54.66	55.62	54.58
NO vs. CHI	54.65	56.69	56.24
LAA vs. CIN	55.88	55.52	57.46
DET vs. MIN	55.17	55.99	57.09
KC vs. SEA	54.70	57.62	54.18

Table 3.5: 95% Prediction Interval Widths for Three Sets of Data

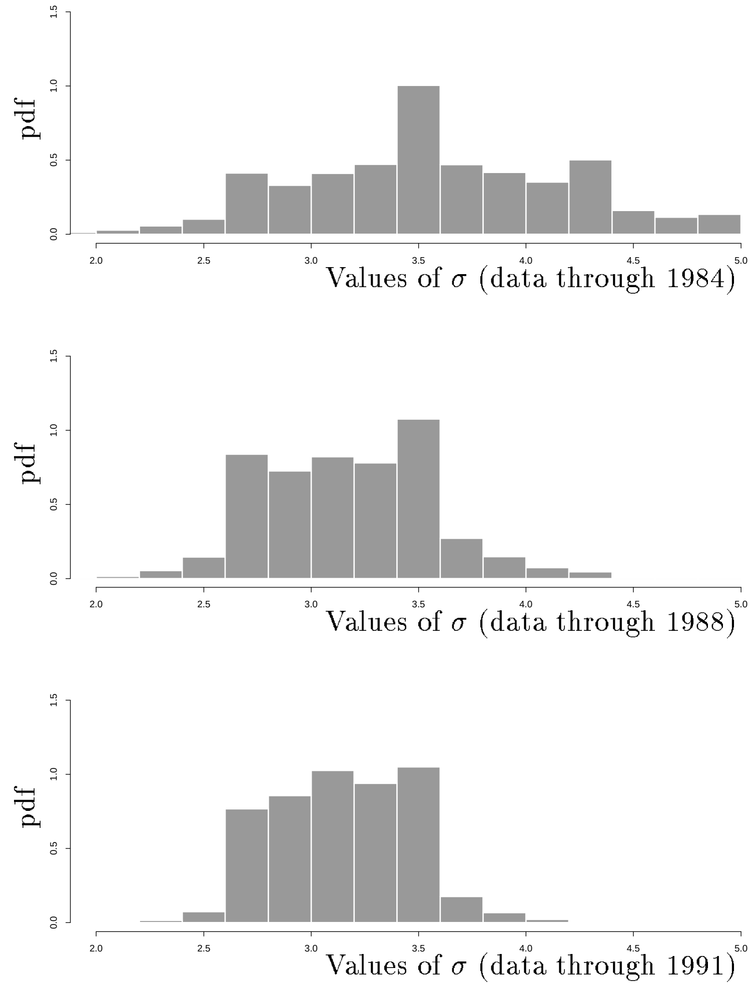


Figure 3.6: Distribution of σ over three sets of data

by generating 3000 random draws of the predictive distribution for score differences for each analysis, and computing the difference between the 2.5% and 97.5% quantiles for each game. The table shows that the prediction intervals do not vary much by year. This suggests that greater precision in σ does not appear to lead to noticeably greater precision of the forecasts.

3.4 Updating the Posterior Distribution for 1992 NFL Game Results

An important aspect of the model is the ability to update the parameter distributions as new data is acquired. We computed parameter updates incorporating the first four weeks of NFL game scores (56 observations) into the analysis using the non-iterative techniques of Section 2.5.2. The distribution of σ changed very little; the mean and standard deviation of σ after incorporating the four weeks of data from 1992 were 3.19 and 0.306, respectively. The mean changed by 0.03 from the end of the 1991 season, and the change in standard deviation was negligible. We also computed the posterior mean and standard deviation for the team parameters and for the home field advantage, shown in Table 3.6. Even early in the 1992 season, some teams appear to have changed appreciably as measured by the mean of $\boldsymbol{\theta}^{(1992)}$. This can be seen by comparing to Table 3.2. For example, the Redskins dropped 2.7 points in mean point scoring ability whereas Buffalo increased by about the same amount. The Dallas Cowboys, the eventual league champions, have improved from 16th to 9th best after just four weeks. The standard deviations are larger than those in Table 3.2 because under the assumptions of the model there is greater uncertainty in the team parameters towards the beginning of the season. This is also reflected in the reported approximate 95% credible intervals, which are wider when the 1992 results from the beginning of the season are incorporated into the analysis. The posterior mean of the home field advantage parameter changes from 2.96 to 2.94.

3.5 Discussion

The analysis of the NFL data using the methodology of Chapter 2 appears to be a reasonable and flexible approach to forecasting NFL game outcomes. For the 1991 season, the team with the highest expected ability as measured by the posterior means was the winner of the superbowl. Furthermore, the model seems to provide honest posterior prediction intervals, the 14 point posterior predictive standard deviation matching the result obtained by Stern (1991) in an analysis of NFL scores.

An issue which may affect the validity of the results is the non-normality of the data. Because the scoring system in football enables certain scores to be more likely to occur than others, treating score differences as continuous may be too approximate. A more accurate approach might attempt to model the response as a discrete variable.

The normal model for football scores may also be inappropriate for describing the behavior of large score differences. Because the object of the game is merely to outscore an opponent, and not maximize the amount by which a team outscores another, there is no strong incentive to "blow out" an opponent.

Other work on modeling professional football scores have incorporated a variety of

Parameter	Mean	Std Dev	95% Credible Interval
San Francisco 49ers	10.60	4.63	(1.52, 19.92)
Washington Redskins	8.65	4.75	(–0.37, 17.87)
Buffalo Bills	8.20	4.56	(–0.77, 17.26)
Kansas City Chiefs	6.83	4.66	(–2.21, 15.89)
New Orleans Saints	6.58	4.65	(–2.28, 15.94)
Philadelphia Eagles	6.57	4.73	(–2.52, 16.16)
Houston Oilers	5.88	4.64	(–3.16, 15.37)
Miami Dolphins	2.77	4.64	(–6.44, 11.75)
Dallas Cowboys	2.23	4.76	(–7.18, 11.67)
New York Giants	2.16	4.74	(–7.18, 11.54)
Minnesota Vikings	1.40	4.57	(–7.32, 10.38)
Denver Broncos	0.88	4.65	(–8.06, 10.13)
Detroit Lions	0.84	4.65	(–8.11, 9.85)
Chicago Bears	0.59	4.57	(–8.31, 9.49)
Atlanta Falcons	0.29	4.56	(–8.27, 9.20)
Pittsburgh Steelers	0.19	4.63	(–9.08, 9.43)
Los Angeles Raiders	–0.93	4.65	(–10.31, 8.40)
Seattle Seahawks	–2.31	4.64	(–11.28, 7.08)
Cleveland Browns	–3.58	4.63	(–12.44, 5.63)
New York Jets	–4.29	4.63	(–13.24, 4.95)
San Diego Chargers	–4.90	4.65	(–14.25, 4.41)
Cincinnati Bengals	–4.96	4.64	(–13.93, 4.22)
Green Bay Packers	–5.00	4.56	(–14.01, 3.88)
Tampa Bay Buccaneers	–5.24	4.57	(–13.98, 3.67)
Los Angeles Rams	–5.25	4.64	(–13.94, 3.81)
Phoenix Cardinals	–7.86	4.74	(–17.06, 1.20)
Indianapolis Colts	–9.47	4.65	(–18.28, –0.45)
New England Patriots	–10.84	4.73	(–20.04, –1.48)
Home Field Advantage	2.94	0.29	(2.38, 3.49)

Table 3.6: Distribution of $\theta^{(1992)}$ after 4 weeks of 1992 NFL Game Scores

approaches. Harville (1980) constructed linear models for football scores where the parameters follow an autoregressive process, and the parameters are estimated by restricted maximum likelihood estimation. Sallas and Harville (1988) propose a Bayesian dynamic linear model for football scores, but unlike our model, they estimate the system variance parameter a priori. Some least squares approaches to predicting football outcomes are developed in Stefani (1977) and Stefani (1980), while maximum likelihood methods for rating football teams are proposed in Thompson (1975). Stern (1992a) compares some of these different methods for rating NFL football teams and find that the Bayesian dynamic model (with σ^2 fixed) performs better than single season or non-dynamic approaches.

Chapter 4

Paired Comparison Models with Indicator Outcomes

In Chapter 2, we considered a series of models for paired comparison experiments where the response was the final score difference of a competition. This chapter treats the paired comparison setting in which the only available information on the result of a comparison is the identity of the winning team or preferred object. The response variables for the models considered here are binary indicator variables rather than continuous variables. In the following sections, we examine the Bradley-Terry paired comparison model along with a Bayesian extension, and demonstrate the construction of a dynamic paired comparison model that allows parameters to change over time. The methodology in this chapter is similar to that of Chapter 2. We show how to analyze the data under the model using both an iterative simulation approach and a non-iterative method. We also consider an extension of the model to accommodate paired comparison experiments in which the outcome may be a tie or an indication of no preference. In such cases, the response variable is trinary rather than binary.

4.1 The Bradley-Terry Model

The paired comparison model that we consider as the basis for the development of the methodology in this chapter is the Bradley-Terry model(Bradley and Terry 1952). Although the model appeared prior to their 1952 paper (see, for example, Zermelo 1929), the model is connected with their names due to the scope of their work on the subject.

Suppose p teams engage in competition. Assume that each team i has an associated parameter π_i. This parameter, the so-called worth parameter, can be interpreted as the relative ability or strength of the competing team. In conventional paired comparison settings, π_i is interpreted as the "merit" of the i-th object that is to be judged. The

Bradley-Terry model represents the probability one team defeats another as

$$p_{ij} = \Pr(i \text{ defeats } j) = \frac{\pi_i}{\pi_i + \pi_j}. \tag{4.1}$$

Note that $p_{ij} + p_{ji} = 1$ so that the model as stated applies only to paired comparisons without ties. Also, since the values of π_i are unique only up to a multiplicative constant in determining the win probabilities, a constraint is usually imposed on the π_i's, typically $\sum_{i=1}^{m} \pi_i = 1$. The simple formulation of the Bradley-Terry model has made it one of the most popular paired comparison models.

The Bradley-Terry model can be derived in a number of ways. One derivation, useful for showing the connection to the models in Chapter 2, assumes that when team i competes, it produces an unobserved score S_i independently of the opposing team with the cumulative distribution function

$$S_i \sim F_i(s) = \exp(-e^{-(s-\log \pi_i)}).$$

Thus team i draws a random score from an extreme-value distribution (Gumbel 1961) with location parameter $\log \pi_i$. It directly follows that the distribution of the difference $S_i - S_j$ follows a logistic distribution with mean $\log \pi_i - \log \pi_j$, that is,

$$S_i - S_j \sim F_{ij}(s) = \frac{1}{1 + e^{-(s-(\log \pi_i - \log \pi_j))}},$$

which in turn implies

$$\Pr(S_i > S_j) = \Pr(S_i - S_j > 0) = 1 - \frac{1}{1 + e^{\log \pi_i - \log \pi_j}} = \frac{\pi_i}{\pi_i + \pi_j}$$

as in (4.1). Alternative motivations for the Bradley-Terry model include Yellott (1977), who derives the Bradley-Terry model from Luce's Choice Axiom (Luce 1959), and Henery (1986) and Joe (1988), who provide a maximum entropy derivation of the Bradley-Terry model.

A comparison can be made here between the above derivation of the Bradley-Terry model and the normal model in of Chapter 2. The normal model postulates that score differences follow a normal distribution, while the Bradley-Terry model postulates that latent score differences follow a standard logistic distribution. One feature of our version of the Bradley-Terry model is the lack of a corresponding parameter to τ^2 in the normal model. This is not a critical issue because the Bradley-Terry model assumes that scores are not observed, so that any observation variance parameter would not be identifiable simultaneously with the worth parameters. Many paired comparison models allow for a latent observation variance parameter, and possibly different variances for different objects. The Bradley-Terry model, for example, can be extended to allow for separate variances for each object. Another popular model is the Thurstone-Mosteller model (Mosteller 1951; Thurstone 1927) which assumes that preference probabilities can be specified as $p_{ij} = \Phi(-(\gamma_i - \gamma_j)/\sqrt{\sigma_i^2 + \sigma_j^2})$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. Stern (1992b) has shown that the choice among popular paired comparison models, including the Bradley-Terry model and the Thurstone-Mosteller model, is somewhat arbitrary in the sense that the models tend to fit paired comparison data equally well or poorly.

In analyzing paired comparison data under the model, suppose that teams i and j compete n_{ij} times, with team i winning y_{ij} times and losing $n_{ij} - y_{ij} = y_{ji}$ times. Then, letting $\boldsymbol{\pi} = (\pi_1, ..., \pi_p)$ be the Bradley-Terry worth parameters, the distribution of $\boldsymbol{y} = (y_{ij},\ i, j = 1, 2, ..., p)$ is

$$f(\boldsymbol{y}|\boldsymbol{\pi}) = \prod_{i<j} \binom{n_{ij}}{y_{ij}} (\frac{\pi_i}{\pi_i + \pi_j})^{y_{ij}} (\frac{\pi_j}{\pi_i + \pi_j})^{y_{ji}}.$$

The likelihood for $\boldsymbol{\pi}$ can be expressed as

$$\text{Lik}(\boldsymbol{\pi}|\boldsymbol{y}) \ \propto \ \frac{\prod_{i=1}^{p} \pi_i^{y_i}}{\prod_{i<j} (\pi_i + \pi_j)^{n_{ij}}},$$

where $y_i = \sum_{j=1}^{p} y_{ij}$, the total number of wins for team i.

Maximum likelihood estimates of π_i can be obtained by the Newton-Raphson algorithm in the usual way, incorporating the restriction on the sum of the π_i's. Ford (1957) shows that the maximum likelihood estimate is unique as long as in every partition of teams into two non-empty subsets, some team in the second has beaten some team in the first. This condition implies that maximum likelihood estimates cannot be found if any team went undefeated, or if any team lost every game. It follows from the likelihood that, given n_{ij}, the y_i are sufficient statistics for the π_i, so that only teams' total scores are needed to perform the estimation. Bühlmann and Huber (1963) show that the Bradley-Terry model is the only linear paired comparison model for which this is true.

4.2 The Bayesian Formulation

The Bayesian model we adopt here is due to Leonard (1977). For a p-team population, we assume that the probability team i defeats team j is given by (4.1). We impose a multivariate normal prior distribution on the $\log \pi_i$. Notice that it is not appropriate to consider a multivariate normal distribution on the π_i because π_i ranges only over values greater than zero, Reparametrizing the model in terms of $\gamma_i = \log \pi_i$, we have

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j} = \frac{e^{\gamma_i}}{e^{\gamma_i} + e^{\gamma_j}},$$

where γ_i is the i-th team's "rating," we can assume a multivariate normal prior on $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)$ with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ and non-singular covariance matrix $\mathbf{C}$. Naturally, under these assumptions, we release the constraint that $\sum_{i=1}^{m} \pi_i = 1$.

If we view γ_i as representing the relative playing strength of team i, imposing a multivariate normal distribution on $\boldsymbol{\gamma}$ has a nice interpretation. Recalling that the Bradley-Terry model can be derived from the assumption that team scores are drawn from extreme-value distributions with location parameters $\log \pi_i$, the multivariate normal prior distribution on $\boldsymbol{\gamma}$ gives information on the location of each individual team's performance distribution. The mean component, μ_i, of the prior distribution represents the location of the distribution describing, on average, team i's performances. The larger the value of μ_i, the greater ability of the team. The diagonal elements, c_{ii}, of the covariance matrix, $\mathbf{C}$, indicate the variability about the means μ_i of teams' rating distributions. The greater the variance of γ_i, the less certainty exists about a team's ability. The off-diagonal elements, c_{ij}, are the covariances between γ_i and γ_j. When c_{ij} approaches zero, the location parameters γ_i and γ_j vary independently of one another, so that locations of ratings of two competing teams are independent. When the correlation between γ_i and γ_j approaches 1, then (γ_i, γ_j) has a degenerate bivariate normal distribution. This suggests that once we know the location of team i's performance distribution, we then know exactly the location of team j's performance distribution. A perfect correlation and equal variances implies that the difference $\gamma_i - \gamma_j$ is constant.

Given a multivariate normal prior distribution on the parameters $\boldsymbol{\gamma}$, and suppressing the conditioning on the prior parameters, we have

$$f(\boldsymbol{\gamma}) \propto \exp(-\frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{\gamma} - \boldsymbol{\mu})).$$

The Bayesian analysis proceeds by computing a posterior distribution for $\boldsymbol{\gamma}$ given the tournament results, $\boldsymbol{y}$. The likelihood after reparametrizing is given by

$$\mathrm{Lik}(\boldsymbol{\gamma}|\boldsymbol{y}) \propto \prod_{i<j} \frac{(e^{\gamma_i})^{y_{ij}}(e^{\gamma_j})^{y_{ji}}}{(e^{\gamma_i} + e^{\gamma_j})^{n_{ij}}}. \tag{4.2}$$

The posterior distribution for $\boldsymbol{\gamma}$ is proportional to the product of the prior and the likelihood, which can be written as

$$\begin{aligned} f(\boldsymbol{\gamma}|\boldsymbol{y}) &\propto f(\boldsymbol{\gamma})\mathrm{Lik}(\boldsymbol{\gamma}|\boldsymbol{y}) \\ &\propto \exp(-\frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{C}^{-1}(\boldsymbol{\gamma} - \boldsymbol{\mu})) \left(\prod_{i<j} \frac{(e^{\gamma_i})^{y_{ij}}(e^{\gamma_j})^{y_{ji}}}{(e^{\gamma_i} + e^{\gamma_j})^{n_{ij}}} \right). \end{aligned}$$

Leonard (1977) suggests approximating the posterior distribution by a multivariate normal distribution, thereby making the parameter updating nearly conjugate. To do this, the posterior mode is obtained by performing the Newton-Raphson algorithm on the log-posterior distribution for $\boldsymbol{\gamma}$. The posterior covariance matrix is estimated by the inverse of the non-singular Hessian matrix used in the last iteration.

Alternatively, if the likelihood is approximately normal, the posterior distribution can be approximated in the following manner. The maximum likelihood estimate, $\hat{\boldsymbol{\gamma}}$, and estimated singular information matrix, $\boldsymbol{R}$, of $\boldsymbol{\gamma}$ from the likelihood in (4.2) can be found using the Newton-Raphson algorithm. We can then approximate the distribution of $\boldsymbol{\gamma}$ as a multivariate normal distribution with mean $\hat{\boldsymbol{\gamma}}$ and singular precision matrix $\boldsymbol{R}$. Then the posterior density for $\boldsymbol{\gamma}$ is the product of a normal prior density and an approximately normal likelihood, and therefore results in an approximately normal posterior distribution. This method for approximating the posterior by a multivariate normal is equivalent to first expanding the log-likelihood in a Taylor Series around $\hat{\boldsymbol{\gamma}}$, dropping the cubic and higher order terms, and combining the remaining expression with the prior distribution to form a multivariate normal density. The approximate posterior distribution for $\boldsymbol{\gamma}$ is attained in the usual manner,

$$f(\boldsymbol{\gamma}|\boldsymbol{y}) \propto \exp(-\frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\mu}')^{\mathsf{T}}\mathbf{C}'^{-1}(\boldsymbol{\gamma} - \boldsymbol{\mu}'))$$

where

$$\begin{aligned} \boldsymbol{\mu}' &= (\mathbf{C}^{-1} + \boldsymbol{R})^{-1}(\mathbf{C}^{-1}\boldsymbol{\mu} + \boldsymbol{R}\hat{\boldsymbol{\gamma}}) \\ \mathbf{C}' &= (\mathbf{C}^{-1} + \boldsymbol{R})^{-1} \end{aligned}$$

are the posterior parameters. As can be seen from the above equation, the posterior mean can be interpreted as a weighted average of the prior mean and the estimated mean from the data. This posterior distribution can then be used as the prior distribution when additional data becomes available.

The implementation of this model is straightforward. Before a tournament or season of competitions begins, a normally distributed prior distribution is assumed on the ratings. The tournament is played, and then using one of the procedures described, we compute posterior estimates of the mean and covariance matrix for ratings and use these updated parameters as the prior parameters for the next tournament. If a team has never competed in tournaments, the model allows for the assignment of a large prior variance to describe the uncertainty in the location of the team's performance distribution.

Other Bayesian formulations of the Bradley-Terry model have been suggested. The first Bayesian approach was developed by Davidson and Solomon (1973). They introduce a conjugate prior for the Bradley-Terry likelihood, and the worth parameters can then be estimated by finding the mode of the posterior distribution. Chen and Smith (1984) consider a Dirichlet prior on the worth parameters. For their model, estimates are found by taking weighted averages of posterior means.

4.3 A Dynamic Bradley-Terry Model

We consider in this section a dynamic extension to the Bradley-Terry model analogous to the dynamic extension of the normal model in Section 2.3. The model in Section 4.2 is not appropriate for the situation in which teams' abilities may change over time because it treats the data as time exchangeable.

Let $\boldsymbol{y}^{(t)} = (y_{ij}^{(t)},\ i,j = 1,2,...,p, i \neq j)$ be the matrix of preference outcomes where $y_{ij}^{(t)}$ is the number of times i defeats j in tournament t, with $t = 1,\ldots,T$. We assume the Bradley-Terry model as before, with

$$p_{ij}^{(t)} = \Pr(i \text{ defeats } j \text{ in tournament } t) = \frac{\exp(\gamma_i^{(t)})}{\exp(\gamma_i^{(t)}) + \exp(\gamma_j^{(t)})}. \tag{4.3}$$

We now assume a probability model for team ratings over time with

$$\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}^{(t-1)} + \boldsymbol{\nu}^{(t)} \tag{4.4}$$

where $\boldsymbol{\nu}^{(t)}$ is the amount team abilities change from tournament $t-1$ to t. As in the normal model, we assume that $\boldsymbol{\nu}^{(t)}$ is stochastically independent of $\boldsymbol{\gamma}^{(t-1)}$, though more general formulations are possible. We assume that, in general,

$$(\boldsymbol{\nu}^{(t)}|\sigma^2) \sim \mathrm{N}(\boldsymbol{\alpha}_t, \sigma^2\mathbf{I}_p), \tag{4.5}$$

where $\boldsymbol{\alpha}_t$ is the mean amount by which teams change between tournaments $t-1$ and t, and σ^2 is the variance of the innovation that occurs between tournaments. We also specify an initial prior distribution on $\boldsymbol{\gamma}^{(1)}$,

$$\boldsymbol{\gamma}^{(1)} \sim \mathrm{N}(\boldsymbol{\mu}^{(1)}, \mathbf{C}^{(1)})$$

and a prior distribution on $\omega = 1/\sigma^2$,

$$f(\omega|a_0, b_0, D_0) \propto \omega^{a_0-1}e^{-b_0\omega} \tag{4.6}$$

where the hyperparameters $\boldsymbol{\mu}^{(1)}$,$\mathbf{C}^{(1)}$,a_0 and b_0 are specified in advance, and can be chosen to represent prior beliefs.

The model in (4.3)–(4.6) show unmistakable similarity to the dynamic normal model of (2.8)–(2.12). The Bradley-Terry model in (4.3) models the outcomes of comparisons at a specific point in time. The system equation of the dynamic model in (4.4), shows how the Bradley-Terry ratings change over time. The parameter $\boldsymbol{\alpha}_t$ in (4.5), which may be a known function of time or an estimable parameter, indicates how teams' abilities evolve over time. As in the normal dynamic model of Chapter 2, $\boldsymbol{\alpha}_t$ may incorporate information about factors that might influence the outcome of comparisons, like age, preparation, and so on. However, as in Chapter 2, we set $\boldsymbol{\alpha}_t = 0$ for all t in developing the dynamic model. This defines a random walk model for the parameter $\boldsymbol{\gamma}^{(t)}$.

4.4 Non-iterative Analysis of the Dynamic Bradley-Terry Model

We consider in this and the next section two possible methods for analyzing data under the dynamic Bradley-Terry model. These two sections parallel Sections 2.5.1 and 2.5.2

in describing a non-iterative approach and a Gibbs sampler approach. The analyses are made more complex because the likelihood for each paired comparison experiment is no longer a normal density.

This section demonstrates the methods used to analyze data under the dynamic Bradley-Terry model using a non-iterative approach. As with the normal dynamic model, the analysis of the joint posterior distribution of parameters is intractable analytically, so we resort to similar techniques employed in Section 2.5.2. The non-iterative approach involves approximating the distribution of ω by a discrete distribution. Furthermore, we make use of the normal approximation to the Bradley-Terry likelihoods at time t, as described in Section 4.2, in order to facilitate the analysis.

To motivate the non-iterative approach, we assume the full dynamic Bradley-Terry model described in Section 4.3. We are primarily interested in the marginal posterior distribution of $(\gamma^{(T)}|D_T)$. The distribution can be expressed as

$$f(\gamma^{(T)}|D_T) = \int f(\gamma^{(T)}|\omega, D_T)\, f(\omega|D_T)\, d\omega. \tag{4.7}$$

The first density in the integrand can be approximated analytically using the methodology described in Section 4.4.1. The second density, however, has no convenient closed-form representation. We consider an approach in Section 4.4.2 that provides a discrete approximation to $f(\omega|D_T)$. In Section 4.4.3 we consider approximations to the integral in (4.7) in order to examine the posterior distribution of the $\gamma^{(T)}$. Techniques for updating the distributions of the parameters to include new data at time $T+1$ are described in Section 4.4.4.

4.4.1 Analysis conditional on ω

We develop in this section a methodology that allows us to approximate the posterior distribution of $\gamma^{(T)}$ conditional on ω by a multivariate normal distribution. Once we find the distribution of ω, perhaps using the techniques of Section 4.4.2, we can then obtain the posterior distribution of $\gamma^{(T)}$ marginal over ω. The analysis here is similar to the analysis of Section 2.4 for normal scores.

The full posterior posterior distribution of all parameters, given ω, can be written as

$$\begin{aligned}
&f(\gamma^{(1)},\ldots,\gamma^{(T)}|\omega, D_T)\\
&\quad\propto \{f(\gamma^{(1)}|\omega, D_0)\} \times \{f(d_1|\gamma^{(1)},\omega, D_0)\}\\
&\qquad\times\{f(\gamma^{(2)}|\gamma^{(1)},\omega, D_1)\} \times \{f(d_2|\gamma^{(1)},\gamma^{(2)},\omega, D_1)\}\\
&\qquad\vdots\\
&\qquad\times\{f(\gamma^{(T)}|\gamma^{(1)},\ldots,\gamma^{(T-1)},\omega, D_{T-1})\} \times \{f(d_T|\gamma^{(1)},\ldots,\gamma^{(T)},\omega, D_{T-1})\},
\end{aligned}$$

or in more detail as

$$\begin{aligned} & f(\boldsymbol{\gamma}^{(1)},\ldots,\boldsymbol{\gamma}^{(T)}|\omega, D_T) \\ \propto \;& \exp(-\frac{1}{2}(\boldsymbol{\gamma}^{(1)}-\boldsymbol{\mu}^{(1)})^{\mathsf{T}}\mathbf{C}^{(1)\,-1}(\boldsymbol{\gamma}^{(1)}-\boldsymbol{\mu}^{(1)})) \\ & \times \prod_{i<j}\frac{\exp(\gamma_i^{(1)}y_{ij}^{(1)})\exp(\gamma_j^{(1)}y_{ji}^{(1)})}{(\exp(\gamma_i^{(1)})+\exp(\gamma_j^{(1)}))^{n_{ij}^{(1)}}} \\ & \times \prod_{t=2}^{T}\left\{\exp(-\frac{\omega}{2}(\boldsymbol{\gamma}^{(t)}-\boldsymbol{\gamma}^{(t-1)})^{\mathsf{T}}(\boldsymbol{\gamma}^{(t)}-\boldsymbol{\gamma}^{(t-1)}))\right. \\ & \qquad \left.\times \prod_{i<j}\frac{\exp(\gamma_i^{(t)}y_{ij}^{(t)})\exp(\gamma_j^{(t)}y_{ji}^{(t)})}{(\exp(\gamma_i^{(t)})+\exp(\gamma_j^{(t)}))^{n_{ij}^{(t)}}}\right\} \end{aligned}$$

Normal Approximation

We now approximate each of the Bradley-Terry likelihoods by a multivariate normal distribution using the technique of Section 4.2. For tournament t, we obtain the maximum likelihood estimate, $\hat{\boldsymbol{\gamma}}^{(t)}$, and the estimated information matrix $\boldsymbol{R}^{(t)}$ of the parameter $\boldsymbol{\gamma}^{(t)}$, and approximate the t-th likelihood by

$$\mathrm{Lik}(\boldsymbol{\gamma}^{(t)}|d_t) \propto \exp(-\frac{1}{2}(\boldsymbol{\gamma}^{(t)}-\hat{\boldsymbol{\gamma}}^{(t)})^{\mathsf{T}}\boldsymbol{R}^{(t)}(\boldsymbol{\gamma}^{(t)}-\hat{\boldsymbol{\gamma}}^{(t)})).$$

We therefore approximate the full posterior distribution of all parameters by

$$\begin{aligned} & f(\boldsymbol{\gamma}^{(1)},\ldots,\boldsymbol{\gamma}^{(T)}|\omega, D_T) \\ \propto \;& \exp(-\frac{1}{2}(\boldsymbol{\gamma}^{(1)}-\boldsymbol{\mu}^{(1)})^{\mathsf{T}}\mathbf{C}^{(1)\,-1}(\boldsymbol{\gamma}^{(1)}-\boldsymbol{\mu}^{(1)})) \\ & \times \exp(-\frac{1}{2}(\boldsymbol{\gamma}^{(1)}-\hat{\boldsymbol{\gamma}}^{(1)})^{\mathsf{T}}\boldsymbol{R}^{(1)}(\boldsymbol{\gamma}^{(1)}-\hat{\boldsymbol{\gamma}}^{(1)})) \\ & \times \prod_{t=2}^{T}\left\{\exp(-\frac{\omega}{2}(\boldsymbol{\gamma}^{(t)}-\boldsymbol{\gamma}^{(t-1)})^{\mathsf{T}}(\boldsymbol{\gamma}^{(t)}-\boldsymbol{\gamma}^{(t-1)}))\right. \\ & \qquad \left.\times \exp(-\frac{1}{2}(\boldsymbol{\gamma}^{(t)}-\hat{\boldsymbol{\gamma}}^{(t)})^{\mathsf{T}}\boldsymbol{R}^{(t)}(\boldsymbol{\gamma}^{(t)}-\hat{\boldsymbol{\gamma}}^{(t)}))\right\}. \end{aligned}$$

Thus the posterior density is approximated by the product of normal densities, and is itself, therefore, an approximately normal density. Note that the information matrices, $\boldsymbol{R}^{(t)}$, are singular due to the constraint used in defining the maximum likelihood estimates. This does not present any problems here.

Updating and Forecasting Recursions

To obtain the marginal distribution of $(\boldsymbol{\gamma}^{(t)}|\omega, D_T)$ conditional on ω, we demonstrate a recursion analogous to that of Section 2.4 for the normal model. This is the Bayesian calculation that updates the distribution of $(\boldsymbol{\gamma}^{(t)}|\omega, D_{t-1})$ to the distribution of $(\boldsymbol{\gamma}^{(t)}|\omega, D_t)$, and then, via a forecast step, obtains the distribution of $(\boldsymbol{\gamma}^{(t+1)}|\omega, D_t)$.

Suppose before observing tournament t the prior distribution on $\boldsymbol{\gamma}^{(t)}$ is

$$(\boldsymbol{\gamma}^{(t)}|\omega, D_{t-1}) \sim \mathrm{N}(\boldsymbol{\mu}^{(t)}, \mathbf{C}^{(t)}).$$

We now observe tournament outcomes, d_t, and update the distribution of $\boldsymbol{\gamma}^{(t)}$ as in Section 4.2 to obtain

$$(\boldsymbol{\gamma}^{(t)}|\omega, D_t) \sim \mathrm{N}(\boldsymbol{\mu}'^{(t)}, \mathbf{C}'^{(t)}),$$

where

$$\begin{aligned}
\boldsymbol{\mu}'^{(t)} &= (\mathbf{C}^{(t)\,-1} + \boldsymbol{R}^{(t)})^{-1}(\mathbf{C}^{(t)\,-1}\boldsymbol{\mu}^{(t)} + \boldsymbol{R}^{(t)}\hat{\boldsymbol{\gamma}}^{(t)}) \\
\mathbf{C}'^{(t)} &= (\mathbf{C}^{(t)\,-1} + \boldsymbol{R}^{(t)})^{-1}.
\end{aligned}$$

These are the updating equations to modify the distribution of parameters with observed data. Now time passes between tournaments t and $t+1$, and to incorporate the extra uncertainty due to the passage of time we have

$$(\boldsymbol{\gamma}^{(t+1)}|\boldsymbol{\gamma}^{(t)}, \omega, D_t) \sim \mathrm{N}(\boldsymbol{\gamma}^{(t)}, \frac{1}{\omega}\mathbf{I}_p).$$

The joint posterior distribution of $\boldsymbol{\gamma}^{(t+1)}$ and $\boldsymbol{\gamma}^{(t)}$ is therefore

$$\begin{aligned}
&f(\boldsymbol{\gamma}^{(t+1)}, \boldsymbol{\gamma}^{(t)}|\omega, D_t) \\
&\quad= f(\boldsymbol{\gamma}^{(t)}|\omega, D_t) \times f(\boldsymbol{\gamma}^{(t+1)}|\boldsymbol{\gamma}^{(t)}, \omega, D_t) \\
&\quad\propto \exp(-\frac{1}{2}(\boldsymbol{\gamma}^{(t)} - \boldsymbol{\mu}'^{(t)})^{\mathsf{T}}\mathbf{C}'^{(t)\,-1}(\boldsymbol{\gamma}^{(t)} - \boldsymbol{\mu}'^{(t)})) \\
&\qquad \times \exp(-\frac{\omega}{2}(\boldsymbol{\gamma}^{(t+1)} - \boldsymbol{\gamma}^{(t)})^{\mathsf{T}}(\boldsymbol{\gamma}^{(t+1)} - \boldsymbol{\gamma}^{(t)})).
\end{aligned}$$

Marginalizing over $\boldsymbol{\gamma}^{(t)}$ yields

$$\begin{aligned}
&f(\boldsymbol{\gamma}^{(t+1)}|\omega, D_t) \\
&\quad= \int f(\boldsymbol{\gamma}^{(t+1)}, \boldsymbol{\gamma}^{(t)}|\omega, D_t)\, d\boldsymbol{\gamma}^{(t)} \\
&\quad\propto \exp(-\frac{1}{2}(\boldsymbol{\gamma}^{(t+1)} - \boldsymbol{\mu}^{(t+1)})^{\mathsf{T}}\mathbf{C}^{(t+1)\,-1}(\boldsymbol{\gamma}^{(t+1)} - \boldsymbol{\mu}^{(t+1)})),
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\mu}^{(t+1)} &= \boldsymbol{\mu}'^{(t)} \\
\mathbf{C}^{(t+1)} &= \mathbf{C}'^{(t)} + \frac{1}{\omega}\mathbf{I}_p. \qquad (4.8)
\end{aligned}$$

These are the forecasting equations to account for the extra uncertainty in the parameters due to the passage of time. The interpretation of (4.8) is that the mean rating of players remains constant over time under the random walk assumption for the innovations, but the variances increase by $\sigma^2 = 1/\omega$. We now have the approximate desired distribution,

$$(\boldsymbol{\gamma}^{(t+1)}|\omega, D_t) \sim \mathrm{N}(\boldsymbol{\mu}^{(t+1)}, \mathbf{C}^{(t+1)}).$$

This defines one complete iteration of the updating and forecasting recursion for the dynamic Bradley-Terry model conditional on ω.

Predictive Distribution of d_{t+1} given ω

Predictive distributions for the preference probability $p_{ij}^{(T+1)}$ at time $T+1$ given ω can be obtained by marginalizing over $\boldsymbol{\gamma}^{(T+1)}$. From the Bayesian updating and forecasting recursions, we have the distribution

$$(\boldsymbol{\gamma}^{(T+1)}|\omega, D_T) \sim \mathrm{N}(\boldsymbol{\mu}^{(T+1)}, \mathbf{C}^{(T+1)}).$$

We also have the probability i is preferred to j at time $T+1$ conditional on $\gamma_i^{(T+1)}$ and $\gamma_j^{(T+1)}$ given the available data,

$$p_{ij}^{(T+1)}(\boldsymbol{\gamma}^{(T+1)}, \omega, D_T) = \frac{\exp(\gamma_i^{(T+1)})}{\exp(\gamma_i^{(T+1)}) + \exp(\gamma_j^{(T+1)})}.$$

Then the expected probability marginalized over $\boldsymbol{\gamma}^{(T+1)}$ is given by

$$\begin{aligned} & p_{ij}^{(T+1)}(\omega, D_T) \\ &= \int p_{ij}^{(T+1)}(\boldsymbol{\gamma}^{(T+1)}, \omega, D_T)\, f(\boldsymbol{\gamma}^{(T+1)}|\omega, D_T)\, d\boldsymbol{\gamma}^{(T+1)} \\ &= \int \frac{\exp(\gamma_i^{(T+1)})}{\exp(\gamma_i^{(T+1)}) + \exp(\gamma_j^{(T+1)})}\, \varphi(\boldsymbol{\gamma}^{(T+1)}|\boldsymbol{\mu}^{(T+1)}, \mathbf{C}^{(T+1)})\, d\boldsymbol{\gamma}^{(T+1)}, \end{aligned}$$

where $\varphi(\boldsymbol{\gamma}^{(T+1)}|\boldsymbol{\mu}^{(T+1)}, \mathbf{C}^{(T+1)})$ is the multivariate normal density function with mean $\boldsymbol{\mu}^{(T+1)}$ and variance $\mathbf{C}^{(T+1)}$. The first factor of the integrand only depends on $\gamma_i^{(T+1)}$ and $\gamma_j^{(T+1)}$, so that the integral reduces to

$$p_{ij}^{(T+1)}(\omega, D_T) = \int \frac{\exp(\gamma_i^{(T+1)})}{\exp(\gamma_i^{(T+1)}) + \exp(\gamma_j^{(T+1)})}\, \varphi(\boldsymbol{\gamma}_{(ij)}^{(T+1)}|\boldsymbol{\mu}_{(ij)}^{(T+1)}, \mathbf{C}_{(ij)}^{(T+1)})\, d\gamma_i^{(T+1)} d\gamma_j^{(T+1)}, \tag{4.9}$$

where $\boldsymbol{\gamma}_{(ij)}^{(T+1)} = (\gamma_i^{(T+1)}, \gamma_j^{(T+1)})^{\mathsf{T}}$, that is, the subset vector of $\boldsymbol{\gamma}^{(T+1)}$ that contains components i and j. The parameters $\boldsymbol{\mu}_{(ij)}^{(T+1)}$ and $\mathbf{C}_{(ij)}^{(T+1)}$ are the corresponding mean vector and variance matrix of $\boldsymbol{\gamma}_{(ij)}^{(T+1)}$. The first factor of the integrand can be expressed as $1/(1 + \exp(-(\gamma_i^{(T+1)} - \gamma_j^{(T+1)})))$, and then a change of variables allows us to express the integral as

$$p_{ij}^{(T+1)}(\omega, D_T) = \int_{-\infty}^{\infty} \frac{1}{2} \left(\frac{1}{1 + \exp(-u)} \right)\, \varphi(u|\mu_u, \mathbf{C}_u)\, du,$$

where

$$\begin{aligned} u &= \gamma_i^{(T+1)} - \gamma_j^{(T+1)} \\ \mu_u &= \mathrm{E}(\gamma_i^{(T+1)} - \gamma_j^{(T+1)}) \\ \mathbf{C}_u &= \mathrm{Var}(\gamma_i^{(T+1)} - \gamma_j^{(T+1)}). \end{aligned}$$

While no closed-form solution exists, numerical methods can be employed to calculate this integral. Gelman and King (1990) obtain the same integral in a different context, and they approximate $\frac{1}{1+\exp(-u)}$ by a third degree polynomial to obtain an approximate value of the integral. An alternative approach is to perform Monte Carlo integration by drawing a large random sample of values u_k from the normal distribution with mean μ_u and variance $\mathbf{C}_u$, and using the sample mean of $1/2(1+\exp(-u_k))$ as an estimate of $p_{ij}^{(T+1)}$.

4.4.2 Analysis to obtain the distribution of $(\omega|D_T)$

In Section 2.5.2, we introduced the method of approximating the prior distribution on ω by a discrete distribution, and showed how to obtain an approximate posterior distribution on $(\omega|D_T)$. Here we apply this method to the analysis of the dynamic Bradley-Terry model. As in the normal model, we assume here that we have obtained a sample from the prior distribution on ω, or that we have selected overdispersed values for ω and chosen probabilities over these values to reflect our prior beliefs.

Suppose, in particular, we have selected a set of m values of ω, $\{\omega_1, \ldots, \omega_m\}$, and assume we have approximated the prior distribution by the probability function

$$f(\omega|D_0) = \begin{cases} \pi_i^{(0)} & \text{if } \omega = \omega_i \\ 0 & \text{if } \omega \notin \{\omega_1, \ldots, \omega_m\} \end{cases}$$

so that $f(\omega|D_0)$ has mass only on the set $\{\omega_1, \ldots, \omega_m\}$. For any season t, we have by Bayes theorem

$$f(\omega|D_t) = \frac{f(d_t|D_{t-1}, \omega)}{f(d_t|D_{t-1})} f(\omega|D_{t-1}).$$

Thus the marginal distribution of ω given data through season t has a simple relationship to the distribution of ω given data through season $t-1$. Continuing the recursion, we have for all t

$$\begin{aligned} f(\omega|D_t) &= \frac{\prod_{j=1}^{t} f(d_j|D_{j-1}, \omega)}{\prod_{j=1}^{t} f(d_j|D_{j-1})} f(\omega|D_0) \\ &\propto f(\omega|D_0) \prod_{j=1}^{t} f(d_j|D_{j-1}, \omega). \end{aligned}$$

Thus the marginal density of $(\omega|D_T)$ is proportional to the product of the prior density of $(\omega|D_0)$ with a reweighting factor, which is the product of conditional probabilities of the form $f(d_t|D_{t-1},\omega)$. A single term in this product can be written as

$$f(d_t|D_{t-1},\omega) = \int f(d_t|\boldsymbol{\gamma}^{(t)},\omega) f(\boldsymbol{\gamma}^{(t)}|\boldsymbol{\mu}^{(t)},\mathbf{C}^{(t)},\omega)d\boldsymbol{\gamma}^{(t)}, \tag{4.10}$$

where $f(d_t|\boldsymbol{\gamma}^{(t)},\omega)$ is the product of Bradley-Terry probabilities at tournament t given the parameters $\boldsymbol{\gamma}^{(t)}$, and $f(\boldsymbol{\gamma}^{(t)}|\boldsymbol{\mu}^{(t)},\mathbf{C}^{(t)},\omega)$ is the multivariate normal density that is obtained from the updating and forecasting recursions of Section 4.4.1. The integral in (4.10) cannot be evaluated in closed form, so we use Laplace's method (see, for example, Thisted 1988) to obtain an approximate evaluation of the integral. To apply Laplace's method, we find the mode, $\boldsymbol{\gamma}^* = \hat{\boldsymbol{\gamma}}^{(t)}$, of $f(d_t|\boldsymbol{\gamma}^{(t)},\omega)$ and the Hessian matrix, $\boldsymbol{R}^{(t)}$, evaluated at the mode. The approximate mode of the entire integrand is therefore $(\boldsymbol{R}^{(t)} + \mathbf{C}^{(t)\,-1})^{-1}(\boldsymbol{R}^{(t)}\hat{\boldsymbol{\gamma}}^{(t)} + \mathbf{C}^{(t)\,-1}\boldsymbol{\mu}^{(t)})$, and the Hessian matrix of the integrand is $\boldsymbol{R}^* = \boldsymbol{R}^{(t)} + \mathbf{C}^{(t)\,-1}$. Laplace's method then obtains

$$f(d_t|D_{T-1},\omega) \approx f(d_t|\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}^*) f(\boldsymbol{\gamma}^*|\boldsymbol{\mu}^{(t)},\mathbf{C}^{(t)},\omega)(2\pi)^{p/2}|\boldsymbol{R}^*|^{-1/2}. \tag{4.11}$$

This approximation is justified by performing a Taylor series expansion of the logarithm of the integrand around its mode, and dropping the cubic and higher order terms. The method works best when the log of the integrand is well approximated by a quadratic form.

The approximate posterior distribution of ω can be obtained by reweighting the prior distribution by a product of terms of the form given in (4.11),

$$\pi_i^{(T)} = \frac{\pi_i^{(0)} \prod_{j=1}^T f(d_j|\omega_i, D_{j-1})}{\sum_{k=1}^m \pi_k^{(0)} \prod_{j=1}^T f(d_j|\omega_k, D_{j-1})}. \tag{4.12}$$

In summary, to calculate the approximate posterior distribution of ω, we begin by choosing m overdispersed values $\omega_1,\ldots,\omega_m$ of ω and set corresponding prior probabilities to be $\pi_1^{(0)},\ldots,\pi_m^{(0)}$. Then for each value in the set $\{\omega_1,\ldots,\omega_m\}$, compute the approximate normal parameters associated with $f(\boldsymbol{\gamma}^{(t)}|\omega_i, D_{t-1})$, for all t, and calculate via Laplace's method $f(d_t|\omega_i, D_{t-1})$ as given by (4.11). The posterior updates, $\pi_i^{(T)}$, for $\pi_i^{(0)}$, are the reweightings given by (4.12).

4.4.3 Marginal and Forecasting Distributions

Inferences on $\boldsymbol{\gamma}^{(t)}$, or predictions of future outcomes, can be obtained by integrating out the nuisance parameter ω. We discuss in this section methods of performing such inferences or predictions.

Marginal Distribution of $(\gamma^{(T)}|D_T)$

From the analysis of the dynamic Bradley-Terry model, we obtain a discretized posterior distribution of ω. Suppose the posterior distribution of ω has positive mass only on the set $\{\omega_1, \ldots, \omega_m\}$ and that

$$f(\omega|D_T) = \begin{cases} \pi_i^{(T)} & \text{if } \omega = \omega_i \\ 0 & \text{if } \omega \notin \{\omega_1, \ldots, \omega_m\} \end{cases}$$

Moments of ω or $\sigma^2 = 1/\omega$ can be calculated relative to this distribution.

Making inferences on the distribution of $(\boldsymbol{\gamma}^{(T)}|D_T)$ requires marginalizing over ω. We can obtain the marginalized posterior density of $\boldsymbol{\gamma}^{(T)}$ as

$$\begin{aligned} f(\boldsymbol{\gamma}^{(T)}|D_T) &= \int f(\boldsymbol{\gamma}^{(T)}|\omega, D_T)\, f(\omega|D_T) d\omega \\ &= \sum_{i=1}^{m} \pi_i^{(T)} f(\boldsymbol{\gamma}^{(T)}|\omega_i, D_T), \end{aligned}$$

which is a mixture of approximately normal densities. Inferences on $(\boldsymbol{\gamma}^{(T)}|D_T)$ can be made, therefore, by drawing a large sample from the mixture of normal distributions and using the sample to obtain empirical confidence intervals. A single draw from the distribution of $(\boldsymbol{\gamma}^{(T)}|D_T)$ is obtained by first drawing ω from the discrete distribution of $(\omega|D_T)$, and subsequently drawing $\boldsymbol{\gamma}^{(T)}$ from the approximately normal distribution of $(\boldsymbol{\gamma}^{(T)}|\omega, D_T)$. This process is repeated until the sample drawn is large enough to make inferences at the desired level.

Forecasts of Game Outcomes

A predictive estimate of the probability i is preferred to j at time $T+1$, given data through time T, is

$$\begin{aligned} & p_{ij}^{(T+1)}(D_T) \\ &= \int \frac{\exp(\gamma_i^{(T+1)})}{\exp(\gamma_i^{(T+1)}) + \exp(\gamma_j^{(T+1)})}\, \varphi(\boldsymbol{\gamma}_{(ij)}^{(T+1)}|\boldsymbol{\mu}_{(ij)}^{(T+1)}, \mathbf{C}_{(ij)}^{(T+1)}) f(\omega|D_T)\, d\gamma_i^{(T+1)} d\gamma_j^{(T+1)} d\omega \\ &= \int p_{ij}^{(T+1)}(\omega, D_T) f(\omega|D_T)\, d\omega, \end{aligned}$$

where $p_{ij}^{(T+1)}(\omega, D_T)$ is defined in (4.9), and computed as described in Section 4.4.1. Therefore the integral above can be computed as a weighted average of $p_{ij}^{(T+1)}(\omega, D_T)$,

$$p_{ij}^{(T+1)}(D_T) = \sum_{\ell=1}^{m} p_{ij}^{(T+1)}(\omega, D_T) \pi_\ell^{(T+1)}$$

This gives a posterior estimate of the preference probability.

4.4.4 Sequential Updating

This section assumes that we have already observed data through time T and have obtained both the conditional posterior distribution of $(\boldsymbol{\gamma}^{(T)}|\omega, D_T)$ and the posterior distribution of $(\omega|D_T)$, and now we need to update the distributions of the parameters to incorporate information available at time $T+1$.

At time T, we have

$$(\boldsymbol{\gamma}^{(t)}|\omega, D_T) \sim \mathrm{N}(\boldsymbol{\mu}'^{(t)}, \mathbf{C}'^{(t)}),$$

and we also have a sample of m values, $\omega_1, \ldots, \omega_m$, for which

$$f(\omega|D_T) = \begin{cases} \pi_i^{(T)} & \text{if } \omega = \omega_i \\ 0 & \text{if } \omega \notin \{\omega_1, \ldots, \omega_m\} \end{cases}$$

Using the methodology of Section 4.4.1, we can update the distribution of $(\boldsymbol{\gamma}^{(T)}|\omega, D_T)$ employing the updating and forecasting equations. To update the distribution of $(\omega|D_T)$ to the distribution of $(\omega|D_{T+1})$, we apply Bayes rule to obtain

$$\begin{aligned} f(\omega|D_{T+1}) &= \frac{f(d_{T+1}|D_T, \omega)}{f(d_{T+1}|D_T)}\, f(\omega|D_T) \\ &\propto f(d_{T+1}|D_T, \omega)\, f(\omega|D_T). \end{aligned}$$

Laplace's method gives

$$f(d_{T+1}|D_T, \omega) \approx f(d_{T+1}|\boldsymbol{\gamma}^{(T)} = \boldsymbol{\gamma}^*) f(\boldsymbol{\gamma}^*|\boldsymbol{\mu}^{(T)}, \mathbf{C}^{(T)})(2\pi)^{p/2}|\boldsymbol{R}^*|^{-1/2},$$

where $\boldsymbol{\gamma}^*$ and $\boldsymbol{R}^*$ are defined in Section 4.4.2. Therefore, to obtain the updated distribution of ω given the new data at time $T+1$, we reweight the distribution of $(\omega|D_T)$ by the marginal likelihood $f(d_{T+1}|\omega, D_T)$ at time $T+1$,

$$\pi_i^{(T+1)} = \frac{\pi_i^{(T)} f(d_{T+1}|\omega_i, D_T)}{\sum_{j=1}^m \pi_j^{(T)} f(d_{T+1}|\omega_j, D_T)}.$$

From this updated distribution of ω, we can make inferences on $(\boldsymbol{\gamma}^{(T+1)}|D_{T+1})$ or compute forecast estimates using the methods of Section 4.4.3.

4.5 Iterative Simulation Analysis of the Dynamic Bradley-Terry Model

A drawback of the non-iterative analysis described in Section 4.4 is that the Bradley-Terry likelihood for each tournament is approximated by a Gaussian density. It is this approximation that allows the marginal posterior distribution of player parameters to be obtained in closed-form in a manner analogous to the normal models of Chapter 2. However, if the tournaments involve a small number of games then the normal approximation may not be satisfactory. In this section, we consider an analysis using the Gibbs sampler to draw from the exact joint posterior distribution of all parameters. A description of the Gibbs sampler is found in Section 2.5.1. In the current problem, a draw from each of the conditional distributions is accomplished via a Metropolis step (Metropolis et al. 1953) at each iteration of the Gibbs sampler. The Gibbs Sampler with a Metropolis step is described in Section 4.5.1. We then describe the implementation of the Gibbs sampler for the dynamic Bradley-Terry model in Section 4.5.2. In Section 4.5.3 we demonstrate methods for obtaining an approximate marginal posterior distribution of $\gamma^{(T)}$ and predictive distributions for future games, as well as methods for updating the parameters to incorporate new data using the technique of sequential imputation (Kong, Liu and Wong 1991).

4.5.1 The Gibbs Sampler in Conjunction with the Metropolis Algorithm

The Gibbs sampler, as described in Section 2.5.1, draws samples iteratively from a sequence of conditional distributions. After sufficiently many draws, continued application of iterative sampling produces samples from the full joint distribution of the parameters. In some cases, even though the conditional densities may be specified exactly, it is not possible to draw a sample from these densities directly. The Metropolis-Hastings algorithm (Hastings 1970) provides a remedy to this problem.

Suppose we would like to draw a random sample from a distribution F with density function $f(x)$, and we know how to draw a sample from a distribution G with density $g(x)$ both defined on the same space $\mathcal{X}$. The Metropolis algorithm is a Monte Carlo Markov chain method that sequentially draws values $x_1, x_2, \ldots$ in a manner described below that eventually produces a random sample from F. Denote $x_0 \in \mathcal{X}$ an arbitrary initial draw, and let x_i be the i-th draw of sequence. To obtain the $(i+1)$-th draw of the sequence, we draw a trial value y from G, and let

$$x_{i+1} = \begin{cases} y & \text{w.p.} \quad \min(1, \frac{w(y)}{w(x_i)}) \\ x_i & \text{w.p.} \quad 1 - \min(1, \frac{w(y)}{w(x_i)}), \end{cases}$$

where $w(\cdot) = f(\cdot)/g(\cdot)$, the importance ratio. Upon convergence, this algorithm will produce a random sample from the target distribution F (Hastings 1970).

The Metropolis-Hastings algorithm can be incorporated into the Gibbs sampler in a variety of ways. The approach we consider is to implement one Metropolis step, that is, one iteration of the Metropolis-Hastings algorithm, at each draw of a conditional distribution in the Gibbs sampler. To be more precise, suppose we have m random variables $Z_1, \ldots, Z_m$, and we are able to specify the densities of the conditional distributions $(Z_i|Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_m)$, $i = 1, \ldots, m$. Without loss of generality, we restrict our attention to the distribution of $(Z_1|Z_2, \ldots, Z_m)$. At iteration q, we want to draw from the exact density

$$Z_1^{(q)} \sim f(Z_1 | Z_2^{(q-1)}, \ldots, Z_m^{(q-1)}).$$

Suppose drawing from $f(\cdot)$ is difficult, but we can draw z from $g(Z_1 | Z_2^{(q-1)}, \ldots, Z_m^{(q-1)})$. Then the Metropolis step within the Gibbs sampler involves assigning

$$Z_1^{(q)} = \begin{cases} z & \text{w.p.} \quad \min(1, \frac{w(z)}{w(Z_1^{(q-1)})}) \\ Z_1^{(q-1)} & \text{w.p.} \quad 1 - \min(1, \frac{w(z)}{w(Z_1^{(q-1)})}), \end{cases}$$

where $w(\cdot) = f(\cdot)/g(\cdot)$. In general, one Metropolis step is performed for every conditional distribution in the Gibbs sampler for which drawing a sample value is difficult. It is possible to perform several Metropolis steps when sampling from a particular conditional distribution, but the tradeoff between the improved convergence properties and increased computational costs is still an open question.

4.5.2 Gibbs Sampler Analysis of the Dynamic Bradley-Terry Model

The goal of our analysis in this section is to describe a method of drawing a random sample of parameter values from the joint posterior distribution $(\gamma^{(1)}, \ldots, \gamma^{(T)}, \omega | D_T)$. We will base our inferences about the parameters on the sample.

Unlike the Gibbs sampler of Section 2.5.1, we cannot rely on the conjugacy of the likelihoods and the prior distributions to simplify our analysis. Consequently, we do not sample alternately between two conditional distributions, but rather among all $T+1$ conditional distributions. We describe now the sampling strategy for alternately drawing from the distributions $(\gamma^{(t)} | \gamma^{(-t)}, \omega, D_T)$, for $t = 1, \ldots, T$, where $\gamma^{(-t)} = (\gamma^{(1)}, \ldots, \gamma^{(t-1)}, \gamma^{(t+1)}, \ldots, \gamma^{(T)})$, and $(\omega | \gamma^{(1)}, \ldots, \gamma^{(T)}, D_T)$.

Steps of the Gibbs Sampler

The Gibbs sampler proceeds as follows:

1. Pick a starting value, ω_c, for the system precision.

2. For $t = 1, \ldots, T$,

   a. Let $\gamma_o^{(t)}$ be the current value of $\gamma^{(t)}$.

   b. Find the conditional density (up to a scalar constant), f, of $\gamma^{(t)}$ given the remaining parameters.

   c. Compute the approximate mean, M_t, and approximate variance, V_t, of the conditional distribution.

   d. Draw a candidate value γ^* from a normal distribution with mean M_t and variance V_t. Let the normal density be given by g.

e. Accept γ^* as the new current value $\gamma_c^{(t)}$ with probability $\min(1, \frac{w(\gamma^*)}{w(\gamma_o^{(t)})})$, and keep $\gamma_o^{(t)}$ as the current value otherwise, where $w(\cdot) = f(\cdot)/g(\cdot)$.

3. Draw ω_c from

$$(\omega|\gamma_c^{(1)}, \ldots, \gamma_c^{(T)}, D_T) \sim \text{Gamma}(a_0+p(T-1)/2, b_0+\frac{1}{2}\sum_{t=2}^{T}(\gamma_c^{(t)}-\gamma_c^{(t-1)})^{\mathsf{T}}(\gamma_c^{(t)}-\gamma_c^{(t-1)})).$$

Repeat steps (2) and (3) until convergence.

Sampling from $(\gamma^{(t)}|\gamma^{(-t)}, \omega, D_T)$

This section describes step 2 of Gibbs sampler procedure outlined above. To draw from the conditional distributions, we first note that the joint distribution of all parameters can be expressed as

$$\begin{aligned}
&f(\gamma^{(1)}, \ldots, \gamma^{(T)}, \omega|D_T) \\
&\quad\propto \ \{f(\omega|D_0)\} \times \{f(\gamma^{(1)}|\omega, D_0)\} \times \{f(d_1|\gamma^{(1)}, \omega, D_0)\} \\
&\qquad \times\{f(\gamma^{(2)}|\gamma^{(1)}, \omega, D_1)\} \times \{f(d_2|\gamma^{(1)}, \gamma^{(2)}, \omega, D_1)\} \\
&\qquad \vdots \\
&\qquad \times\{f(\gamma^{(T)}|\gamma^{(1)}, \ldots, \gamma^{(T-1)}, \omega, D_{T-1})\} \times \{f(d_T|\gamma^{(1)}, \ldots, \gamma^{(T)}, \omega, D_{T-1})\},
\end{aligned} \tag{4.13}$$

where the conditional densities of $\gamma^{(t)}$ are normal densities, and the likelihoods are products of Bradley-Terry probabilities.

The conditional posterior distribution for $\gamma^{(t)}$ given the rest of the parameters has a density proportional to that in (4.13). Let ω_c and $\gamma_c^{(-t)}$ be the current draws of ω and $\gamma^{(-t)}$, and let $\gamma_o^{(t)}$ be the most recent draw of $\gamma^{(t)}$. We consider the case where $1 < t < T$. Neglecting terms constant with respect to $\gamma^{(t)}$ yields

$$\begin{aligned}
&f(\gamma^{(t)}|\gamma_c^{(-t)}, \omega_c, D_T) \\
&\quad\propto \ \{f(\gamma^{(t)}|\gamma_c^{(t-1)}, \omega_c, D_{t-1})\} \times \{f(d_t|\gamma^{(t)}, \omega_c, D_{t-1})\} \times \{f(\gamma_c^{(t+1)}|\gamma^{(t)}, \omega_c, D_t)\}.
\end{aligned} \tag{4.14}$$

The conditional distribution of $\gamma^{(t)}$ is proportional to the product of a normal density with mean $\gamma_c^{(t-1)}$ and variance $\frac{1}{\omega_c}\mathbf{I}$, a Bradley-Terry likelihood, and a normal density, where $\gamma^{(t)}$ has mean $\gamma_c^{(t+1)}$ and variance $\frac{1}{\omega_c}\mathbf{I}$. When $t = 1$, the first term in (4.14) is the normal prior density for $\gamma^{(1)}$, and when $t = T$, the last term in (4.14) vanishes. Let $\hat{\gamma}^{(t)}$ be the mode of the likelihood for tournament t and let $\boldsymbol{R}^{(t)}$ be the information matrix evaluated at the mode. The variance and mean of the distribution defined by the product of these three densities is given by

$$\begin{aligned}
V_t &= \text{Var}(\gamma^{(t)}|\gamma_c^{(-t)}, \omega_c, D_T) = (\boldsymbol{R}^{(t)} + 2\omega_c\mathbf{I})^{-1} \\
M_t &= \text{E}(\gamma^{(t)}|\gamma_c^{(-t)}, \omega_c, D_T) = V_t(\boldsymbol{R}^{(t)}\hat{\gamma}^{(t)} + \omega_c(\gamma_c^{(t-1)} + \gamma_c^{(t+1)})).
\end{aligned} \tag{4.15}$$

Again, when $t = 1$ or $t = T$ these expressions change to reflect the appropriate conditional density. Letting $g(\gamma^{(t)}|\gamma_c^{(-t)}, \omega_c, D_T)$ be the normal density of $\gamma^{(t)}$ with mean M_t and variance V_t, and letting $w(\cdot) = f(\cdot)/g(\cdot)$, we draw γ^* from the multivariate normal distribution g. The Metropolis step gives the current draw

$$\gamma_c^{(t)} = \begin{cases} \gamma^* & \text{w.p.} \quad \min(1, \frac{w(\gamma^*)}{w(\gamma_o^{(t)})}) \\ \gamma_o^{(t)} & \text{w.p.} \quad 1 - \min(1, \frac{w(\gamma^*)}{w(\gamma_o^{(t)})}) \end{cases}$$

This completes step 2 for a single t of the Gibbs sampler.

Sampling from $(\omega|\gamma^{(1)}, \ldots, \gamma^{(T)}, D_T)$

This section discusses step 3 of the Gibbs sampler outlined above. The conditional posterior distribution of ω given the rest of the parameters does not require a Metropolis step. Suppose we have current draws $\gamma_c^{(1)}, \ldots, \gamma_c^{(T)}$, and we would like to draw ω_c from the conditional distribution of $(\omega|\gamma_c^{(1)}, \ldots, \gamma_c^{(T)}, D_T)$. The conditional posterior distribution of ω is proportional to the density in (4.13). Neglecting terms that are constant with respect to ω and substituting in the exact densities yields

$$\begin{aligned} & f(\omega|\gamma_c^{(1)}, \ldots, \gamma_c^{(T)}, D_T) \\ & \quad \propto \quad \omega^{a_0 - 1} e^{-b_0 \omega} \\ & \qquad \times \prod_{t=2}^{T} \left\{ \omega^{p/2} \exp(-\frac{\omega}{2}(\gamma_c^{(t)} - \gamma_c^{(t-1)})^{\mathsf{T}}(\gamma_c^{(t)} - \gamma_c^{(t-1)})) \right\} \\ & \quad \propto \quad \omega^{a_0 - 1} e^{-b_0 \omega} \\ & \qquad \times \omega^{p(T-1)/2} \exp(-\frac{\omega}{2} \sum_{t=2}^{T} (\gamma_c^{(t)} - \gamma_c^{(t-1)})^{\mathsf{T}}(\gamma_c^{(t)} - \gamma_c^{(t-1)})), \end{aligned}$$

so that

$$(\omega|\gamma_c^{(1)}, \ldots, \gamma_c^{(T)}, D_T) \sim \text{Gamma}(a_0 + p(T-1)/2, b_0 + \frac{1}{2} \sum_{t=2}^{T} (\gamma_c^{(t)} - \gamma_c^{(t-1)})^{\mathsf{T}}(\gamma_c^{(t)} - \gamma_c^{(t-1)})). \tag{4.16}$$

At step 3 of the Gibbs sampler, we draw ω_c according to the gamma distribution in (4.16).

4.5.3 Sequential Imputation for Parameter and Predictive Inferences

The result of performing the Gibbs sampler is a large sample of draws from the joint posterior distribution of $(\gamma^{(1)}, \ldots, \gamma^{(T)}, \omega|D_T)$. We show in this section how to make inferences on parameters of interest based on these samples. We also demonstrate using sequential imputation a method for forecasting future performances and for updating the parameter distributions when new data is observed.

Suppose the Gibbs sampler provides m samples of parameter values which can be considered drawn at random from the joint posterior distribution of the parameters. The collection of these samples can be listed as

$$\left\{(\gamma_1^{(1)},\ldots,\gamma_1^{(T)},\omega_1),(\gamma_2^{(1)},\ldots,\gamma_2^{(T)},\omega_2),\ldots,(\gamma_m^{(1)},\ldots,\gamma_m^{(T)},\omega_m)\right\},$$

where the subscripts index samples rather than players. We can approximate the exact joint continuous posterior distribution by a discrete distribution assigning posterior probabilities of $\pi_i^{(T)} = 1/m$ to each of the m samples. In the following discussion, we refer to the posterior probabilities as $\pi_i^{(T)}$ to retain flexibility.

Obtaining approximate inferences about parameters is trivial in this framework. To estimate the marginal mean and variance of the most recent player parameters, $\gamma^{(T)}$, it suffices to find the expectation over $\pi^{(T)}$. Thus the mean and variance for $\gamma^{(T)}$ are given by

$$\begin{aligned}\mathrm{E}(\gamma^{(T)}|D_T) &\approx \sum_{i=1}^{m}\pi_i^{(T)}\gamma_i^{(T)}\\ \mathrm{Var}(\gamma^{(T)}|D_T) &\approx \sum_{i=1}^{m}\pi_i^{(T)}(\gamma_i^{(T)}-\mathrm{E}(\gamma^{(T)}|D_T))(\gamma_i^{(T)}-\mathrm{E}(\gamma^{(T)}|D_T))^{\mathsf{T}},\end{aligned}$$

which are, of course, just the sample mean and sample variance-covariance matrix. Posterior intervals for $\gamma^{(T)}$ can be computed from the empirical distribution as well.

Inferences about $(\sigma^2|D_T)$ can be performed analogously. The marginal posterior distribution on $(\sigma^2|D_T)$ is approximated by the discrete distribution on m values, $\sigma_i^2 = 1/\omega_i$, with probabilities $\pi_i^{(T)} = 1/m$. Moments of the true distribution can be approximated by the moments of the discrete approximating distribution as before. Confidence intervals for $(\sigma^2|D_T)$ can be estimated from the empirical distribution.

To obtain predictive distributions for the team ratings at time $T+1$, we make use of sequential imputation (Kong, Liu and Wong 1991). In the context of the dynamic Bradley-Terry model, we impute m values for $\gamma_i^{(T+1)}$. We have from our Gibbs sampler posterior draws $\gamma_1^{(T)},\ldots,\gamma_m^{(T)}$ of $(\gamma^{(T)}|D_T)$, and $\sigma_1^2,\ldots,\sigma_m^2$ of $(\sigma^2|D_T)$. We can then approximate the distribution of $(\gamma^{(T+1)}|D_T)$ by drawing m sample values from

$$\gamma_i^{(T+1)} \sim \mathrm{N}(\gamma_i^{(T)},\sigma_i^2\mathbf{I}),$$

thereby obtaining m draws from the marginal distribution of $(\gamma^{(T+1)}|D_T)$.

To estimate the prior probability that player j defeats player k at time $T+1$, we need to find the value of the

$$\begin{aligned}p_{jk}^{(T+1)}(D_T) &= \int p_{jk}^{(T+1)}(\gamma^{(T+1)},D_T)f(\gamma^{(T+1)}|D_T)d\gamma^{(T+1)}\\ &= \int \frac{\exp(\gamma_j^{(T+1)})}{\exp(\gamma_j^{(T+1)})+\exp(\gamma_k^{(T+1)})}f(\gamma^{(T+1)}|D_T)d\gamma^{(T+1)}.\end{aligned}$$

From the imputed $\gamma_i^{(T+1)}$, $i = 1, \ldots, m$, this integral can be approximated by

$$\sum_{i=1}^{m} \frac{\exp(\gamma_{ij}^{(T+1)})}{\exp(\gamma_{ij}^{(T+1)}) + \exp(\gamma_{ik}^{(T+1)})} \pi_i^{(T)},$$

where $\gamma_{ij}^{(T+1)}$ is the j-th component of the vector $\gamma_i^{(T+1)}$. Thus for each i, we compute the Bradley-Terry probability, and compute its expectation over the empirical distribution $\pi^{(T)}$ of the m values.

Updating $\pi^{(T)}$ to $\pi^{(T+1)}$ to incorporate tournament data at time $T+1$ requires the application of sequential imputation in a manner similar to that of Section 2.8. From Bayes rule, we have

$$\begin{aligned} f(\gamma^{(1)}, \ldots, \gamma^{(T+1)}, \omega | D_{T+1}) &= \frac{f(d_{T+1} | D_T, \gamma^{(1)}, \ldots, \gamma^{(T+1)}, \omega)}{f(d_{T+1} | D_T)} f(\gamma^{(1)}, \ldots, \gamma^{(T+1)}, \omega | D_T) \\ &\propto f(d_{T+1} | D_T, \gamma^{(1)}, \ldots, \gamma^{(T+1)}, \omega)\, f(\gamma^{(1)}, \ldots, \gamma^{(T+1)}, \omega | D_T) \\ &= f(d_{T+1} | \gamma^{(T+1)})\, f(\gamma^{(1)}, \ldots, \gamma^{(T+1)}, \omega | D_T). \end{aligned}$$

This last equality holds because the distribution of game outcomes at time $T+1$ given $\gamma^{(T+1)}$ is independent of the other parameters and the previous data. Thus the joint posterior distribution given D_{T+1} is a reweighting of the joint posterior distribution by a factor of $f(d_{T+1} | \gamma^{(T+1)})$, which can be computed exactly. To obtain $\pi_i^{(T+1)}$, we reweight by the exact densities, that is,

$$\pi_i^{(T+1)} = \frac{\pi_i^{(T)} f(d_{T+1} | \gamma_i^{(T+1)})}{\sum_{j=1}^{m} \pi_j^{(T)} f(d_{T+1} | \gamma_j^{(T+1)})}. \tag{4.17}$$

From this new set of values of $\pi_i^{(T+1)}$ we can compute posterior credible intervals for parameters of interest, such as $\gamma^{(T+1)}$ or σ^2 conditional on D_{T+1}.

4.5.4 Summary of Iterative Simulation Analysis

A typical analysis of tournament data will therefore proceed in the following manner.

1. We observe T tournaments of game outcomes, and perform a Gibbs sampler analysis to obtain m draws from the joint posterior distribution of $(\gamma^{(1)}, \ldots, \gamma^{(T)}, \omega | D_T)$, and set $\pi_i^{(T)} = 1/m$ for $i = 1, \ldots, m$.

2. We sequentially impute $\gamma_i^{(T+1)}$ for each of the m draws of $(\gamma_i^{(1)}, \ldots, \gamma_i^{(T)}, \omega | D_T)$ by randomly sampling from $\mathrm{N}(\gamma_i^{(T)}, \frac{1}{\omega_i}\mathbf{I})$.

3. We now observe tournament data at time $T+1$, and update the weights associated with our m samples, $\pi_i^{(T)}$, according to (4.17) to obtain $\pi_i^{(T+1)}$.

An interesting aspect of this analysis is that we are always using the same m draws obtained from the Gibbs sampler at time T. A possible problem with this analysis is that continued sequential updating may reveal that the original sample of parameters do not represent the true distribution accurately. For example, after updating the posterior distributions to incorporate several new tournament results, we may find that the values of ω obtained from the original sample of the Gibbs sampler do not include a wide enough range of values. This may be diagnosed by a skewed distribution on any of the parameters after several sequential updates, or alternatively by noting that few Gibbs samples receive increasing weight in imputation. In this case, it may be necessary to reperform the Gibbs sampler in its entirety to obtain a more representative sample of draws from the posterior distribution.

4.6 Ties

In many situations, a paired comparison can result in neither object being preferred. The models considered to this point assume that a comparison results in a binary outcome. Here we develop a model where a third outcome, no preference, is possible.

4.6.1 A Model for Ties

The probability model we consider here is proposed by Davidson (1970). The model postulates

$$\begin{aligned} p_{ij} = \Pr(i \text{ defeats } j) &= \frac{e^{\gamma_i}}{e^{\gamma_i} + e^{\gamma_j} + e^{\lambda+(\gamma_i+\gamma_j)/2}} \\ p_{ji} = \Pr(j \text{ defeats } i) &= \frac{e^{\gamma_j}}{e^{\gamma_i} + e^{\gamma_j} + e^{\lambda+(\gamma_i+\gamma_j)/2}} \\ p_{ij.0} = \Pr(i \text{ ties } j) &= \frac{e^{\lambda+(\gamma_i+\gamma_j)/2}}{e^{\gamma_i} + e^{\gamma_j} + e^{\lambda+(\gamma_i+\gamma_j)/2}}, \end{aligned} \tag{4.18}$$

where the parameter λ is an index of discrimination. Large positive values of λ indicate higher probabilities of a tie. Equation 4.18 assumes that one additional parameter is sufficient for modeling the occurrences of ties. One generalization that has been examined incorporates a separate parameter λ_{ij} for each pair being compared (Singh and Gupta 1975; Singh and Gupta 1978). Another possible generalization is to allow one tie parameter, λ_i, per player. This possibility is discussed in Section 5.7. A further attempt at extending the Bradley-Terry model to incorporate ties refers back to the extreme value score distribution as a motivation for the Bradley-Terry model. Rao and Kupper (1967) proposed a model that postulates a tie is declared when the difference between two scores is less than a certain threshold parameter.

The Bayesian analysis of the Davidson model proceeds in a manner analogous to Section 4.2. We impose a multivariate normal prior distribution on the parameters

$\gamma_1, \ldots, \gamma_p, \lambda$. The analysis approximates the distribution of the parameters by a multivariate normal distribution centered at the maximum likelihood estimates. The approximate multivariate normal posterior distribution is obtained by the usual averaging of the normal prior distribution with the approximate normal likelihood. Again, this procedure obtains the exact posterior distribution only if the likelihood is, in fact, approximately normal.

We propose a dynamic generalization to the Davidson model. Let $\boldsymbol{\gamma}^{(t)} = (\gamma_1^{(t)}, \ldots, \gamma_p^{(t)})$ be the vector of player ratings at time t, and let the tie parameter, λ, be constant over time. Then the probability model for observations at time t is given by

$$\begin{aligned} p_{ij}^{(t)} = \Pr(i \text{ defeats } j) &= \frac{e^{\gamma_i^{(t)}}}{e^{\gamma_i^{(t)}} + e^{\gamma_j^{(t)}} + e^{\lambda + (\gamma_i^{(t)} + \gamma_j^{(t)})/2}} \\ p_{ji}^{(t)} = \Pr(j \text{ defeats } i) &= \frac{e^{\gamma_j^{(t)}}}{e^{\gamma_i^{(t)}} + e^{\gamma_j^{(t)}} + e^{\lambda + (\gamma_i^{(t)} + \gamma_j^{(t)})/2}} \\ p_{ij.0}^{(t)} = \Pr(i \text{ ties } j) &= \frac{e^{\lambda + (\gamma_i^{(t)} + \gamma_j^{(t)})/2}}{e^{\gamma_i^{(t)}} + e^{\gamma_j^{(t)}} + e^{\lambda + (\gamma_i^{(t)} + \gamma_j^{(t)})/2}}. \end{aligned} \tag{4.19}$$

The probability model for the evolution of the parameters $\boldsymbol{\gamma}^{(t)}$ is given by

$$\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}^{(t-1)} + \boldsymbol{\nu}^{(t)}$$

and as in Section 4.3 we let

$$(\boldsymbol{\nu}^{(t)} | \sigma^2) \sim \mathrm{N}(\boldsymbol{\alpha}_t, \sigma^2 \mathbf{I}_p),$$

where $\boldsymbol{\alpha}_t$ is the mean amount by which team abilities change between tournaments $t-1$ and t, and σ^2 is the variance that reflects the increased uncertainty in $\boldsymbol{\gamma}^{(t)}$ with the passage of time.

We place a prior distribution on $(\boldsymbol{\gamma}^{(1)}, \lambda)$,

$$(\boldsymbol{\gamma}^{(1)}, \lambda) \sim \mathrm{N}(\boldsymbol{\mu}^{(1)}, \mathbf{C}^{(1)}),$$

where $\boldsymbol{\mu}^{(1)}$ and $\mathbf{C}^{(1)}$ are $(p+1)$-dimensional. The prior distribution on $\omega = 1/\sigma^2$ is

$$f(\omega) \propto \omega^{a_0 - 1} e^{-b_0 \omega}.$$

The hyperparameters $\boldsymbol{\mu}^{(1)}$, $\mathbf{C}^{(1)}$, a_0, and b_0 are assumed to be specified in advance.

It is worth noting the similarities and differences between the dynamic Bradley-Terry models with and without ties. The model for the evolution of parameters treats the $\boldsymbol{\gamma}^{(t)}$ as time varying in both models, and the amount of change over time is governed by the variance parameter σ^2. The model also assumes that the prior distribution of $\boldsymbol{\gamma}^{(1)}$ is multivariate normal, and both models assume an identical prior distribution for σ^2. A noteworthy difference between the models, besides the allowance for a third outcome of the paired comparison, is that the parameter λ is treated as constant over time. In the next section, we discuss the implications of this difference on the analysis of the model.

4.6.2 Analysis of the Model

The analysis of the dynamic paired comparison model with ties requires only minor changes to the analyses described in Sections 4.4 and 4.5. We describe the required modifications to the non-iterative and Gibbs sampler analyses here.

As in the paired comparison models with binary outcomes, the non-iterative analysis of the model with ties involves approximating the likelihood at each time t by a multivariate normal density. For the Davidson model, Bradley and Gart (1962) show how to obtain the maximum likelihood estimates and the estimated information matrix for each tournament. The likelihood can be approximated by a multivariate normal density by setting the mean to be the vector of maximum likelihood estimates, and the inverse of the variance to be the estimated information matrix. We let $\gamma_*^{(t)}$ be the $(p+1)$-vector $(\gamma^{(t)}, \lambda)$, and let $\hat{\gamma}_*^{(t)}$ be the maximum likelihood estimate of $\gamma_*^{(t)}$ from the data at time t and let $\boldsymbol{R}_*^{(t)}$ be the $(p+1)$-dimensional information matrix.

The analysis of the model conditional on $\sigma^2 = 1/\omega$ requires the respecification of the updating and forecasting calculations. Suppose before observing tournament at time t, the prior distribution on the parameters is

$$(\gamma_*^{(t)}|\sigma^2, D_{t-1}) \sim \mathrm{N}(\boldsymbol{\mu}^{(t)}, \mathbf{C}^{(t)}).$$

Upon observing data at time t, d_t, we can update the distribution above to obtain the posterior distribution

$$(\gamma_*^{(t)}|\sigma^2, D_t) \sim \mathrm{N}(\boldsymbol{\mu}'^{(t)}, \mathbf{C}'^{(t)})$$

where

$$\begin{aligned} \boldsymbol{\mu}'^{(t)} &= (\mathbf{C}^{(t)\,-1} + \boldsymbol{R}_*^{(t)})^{-1}(\mathbf{C}^{(t)\,-1}\boldsymbol{\mu}^{(t)} + \boldsymbol{R}_*^{(t)}\hat{\gamma}_*^{(t)}) \\ \mathbf{C}'^{(t)} &= (\mathbf{C}^{(t)\,-1} + \boldsymbol{R}_*^{(t)})^{-1}. \end{aligned}$$

The forecasting calculation, which computes the distribution of the parameters reflecting the uncertainty due to the passage of time, is given by

$$(\gamma_*^{(t+1)}|\sigma^2, D_t) \sim \mathrm{N}(\boldsymbol{\mu}^{(t+1)}, \mathbf{C}^{(t+1)}),$$

where

$$\begin{aligned} \boldsymbol{\mu}^{(t+1)} &= \boldsymbol{\mu}'^{(t)} \\ \mathbf{C}^{(t+1)} &= \mathbf{C}'^{(t)} + \sigma^2\mathbf{J}_{p,1}. \end{aligned}$$

Here, $\mathbf{J}_{p,1}$ is the $(p+1)$-dimensional matrix consisting of 1 in the first p diagonal elements, and 0 elsewhere. The forecasting formula for the variance-covariance matrix has the interpretation that the variance increases over time for $\gamma^{(t)}$, but the variance of the non-dynamic parameter, λ, does not change due to the passage of time. These forecasting formulas, along with the updating formulas, can be used successively to obtain the marginal distribution of $\gamma_*^{(T)}$ from the prior distribution on the parameters conditional on σ^2.

Obtaining the distribution of $(\omega|D_T)$ is identical to the analysis described in Section 4.4.2. We assume, as before, a discrete prior distribution on $(\omega|D_0)$. To obtain the posterior distribution on $(\omega|D_T)$, we compute the reweighting factors as products of marginal likelihoods which can be approximated using Laplace's method. Once the approximate distribution of $(\omega|D_T)$ is found, posterior moments and confidence intervals can be found using the approximating discrete distribution. Prediction probabilities can be found via Monte Carlo simulation by generating samples from the approximately normal $(\gamma_*^{(T+1)}|D_T)$, and then computing the preference probabilities for each of the generated samples.

The Gibbs sampler methodology for the model with ties is similar to the analysis considered in Section 4.5. The main difference is now that at each iteration of the Gibbs sampler, draws from $T+2$ conditional distributions are required; drawing from the distribution of $(\omega|\gamma^{(1)}, \ldots, \gamma^{(T)}, \lambda, D_T)$, from each of the $(\gamma^{(t)}|\gamma^{(-t)}, \lambda, \omega, D_T)$ for $t = 1, \ldots, T$, and from $(\lambda|\gamma^{(1)}, \ldots, \gamma^{(T)}, \omega, D_T)$. We briefly describe here the sampling strategy for this Gibbs sampler.

Drawing from the distribution of $(\gamma^{(t)}|\gamma^{(-t)}, \lambda, \omega, D_T)$ is similar to Section 4.5.2 with the Bradley-Terry likelihoods replaced by the Davidson likelihoods. In particular, a multivariate normal distribution is constructed from which a candidate vector for $\gamma^{(t)}$ will be sampled. The parameters of this multivariate normal distribution are obtained in a manner identical to the procedure of Section 4.5.2; a single draw is sampled, and the ratio of importance weights is computed by calculating the densities of both the normal distribution from which the draw was sampled, and the density corresponding to the exact conditional posterior distribution (the target distribution). The Metropolis step then either accept the candidate vector $\gamma^{(t)}$ or retains the previous value. In this entire computation, λ is treated as fixed and known.

To draw from the distribution of $(\lambda|\gamma^{(1)}, \ldots, \gamma^{(T)}, \omega, D_T)$, we also make use of a Metropolis step. First, before performing the Gibbs sampler, we construct a normal distribution from which to sample a candidate value of λ. To do this, we make a rough guess at the mean and variance of the marginal posterior distribution of λ by averaging the posterior means and the posterior variances obtained by considering the T tournaments individually. We then inflate this variance by a positive value greater than 1 to ensure that the candidate samples are dispersed relative to the true posterior distribution of λ. This same normal approximation is used to generate candidate values of λ throughout the Gibbs sampler. Sampling λ conditional on the remaining parameters involves drawing a candidate value of λ from the overdispersed normal distribution, and then invoking the Metropolis step by computing the ratio of importance weights and accepting the candidate value with the appropriate probability.

To sample from $(\omega|\gamma^{(1)},\ldots,\gamma^{(T)},\lambda,D_T)$, the conditional distribution is given by

$$(\omega|\gamma^{(1)},\ldots,\gamma^{(T)},\lambda,D_T) \sim \text{Gamma}(a_0+p(T-1)/2, b_0+\frac{1}{2}\sum_{t=1}^{T-1}(\gamma^{(t+1)}-\gamma^{(t)})^{\mathsf{T}}(\gamma^{(t+1)}-\gamma^{(t)})).$$

Note that the distribution of ω does not involve λ, as the tie parameter is assumed to remain constant over time. This completes a single draw from the Gibbs sampler. The algorithm continues until the Gibbs sampler converges, after which a random sample from the marginal posterior distribution of all parameters may be drawn, and techniques presented in Section 4.5.3 may then be employed to find marginal inferences and predictive probabilities. The approach described here can be used to include more than one non-dynamic covariate. In fact, in Chapter 5 we allow for an order effect (p_{ij} depend on the order in which i and j are compared) in addition to allowing for ties.

4.7 Inclusion of Covariates

In both the Bradley-Terry model and the Davidson model, extensions can be made to incorporate covariate information. We assume covariates can be modeled as a linear combination of parameters, and that these parameters are non-dynamic. Generalizing to dynamic parameters is also straightforward, although we do not consider that case here.

Suppose the q-dimensional vector, β, contains the parameters for the q covariates. Let x_k be a data vector corresponding to the k-th individual paired comparison, and suppose the k-th comparison involves players i_k and j_k. The vector x_k is the set of covariates associated with the k-th game. We then consider the following generalizations of the Bradley-Terry and Davidson models for paired comparisons. For the Bradley-Terry model, we let

$$\begin{aligned} p_{i_kj_k} &= \frac{\exp(\gamma_{i_k}+x_k\beta)}{\exp(\gamma_{i_k}+x_k\beta)+\exp(\gamma_{j_k})} \\ &= \frac{\exp(\gamma_{i_k}-\gamma_{j_k}+x_k\beta)}{\exp(\gamma_{i_k}-\gamma_{j_k}+x_k\beta)+1}. \end{aligned}$$

For the Davidson model, we postulate

$$\begin{aligned} p_{i_kj_k} &= \frac{\exp(\gamma_{i_k}+x_k\beta)}{\exp(\gamma_{i_k}+x_k\beta)+\exp(\gamma_{j_k})+\exp(\lambda+(\gamma_{i_k}+\gamma_{j_k})/2)} \\ p_{j_ki_k} &= \frac{\exp(\gamma_{j_k})}{\exp(\gamma_{i_k}+x_k\beta)+\exp(\gamma_{j_k})+\exp(\lambda+(\gamma_{i_k}+\gamma_{j_k})/2)} \\ p_{i_kj_k\cdot 0} &= \frac{\exp(\lambda+(\gamma_{i_k}+\gamma_{j_k})/2)}{\exp(\gamma_{i_k}+x_k\beta)+\exp(\gamma_{j_k})+\exp(\lambda+(\gamma_{i_k}+\gamma_{j_k})/2)}. \end{aligned}$$

These models can be fit by performing maximum likelihood estimation in the usual way. Critchlow and Fligner (1991) demonstrate that these paired comparison models can be reparametrized as generalized linear models; they take advantage of this structure in order to analyze the models using standard statistical packages. Estimates of γ and β along with the singular covariance matrix can be obtained by recognizing and implementing the paired comparison model as a generalized linear model.

The Bayesian extension to these models is analogous to the Bayesian extension of the ordinary Bradley-Terry model in Section 4.2. A multivariate normal prior distribution is placed on all the parameters, and the likelihood is approximated by a normal density in order to carry out an analysis that is approximately conjugate. We can then report the posterior distribution of parameters as approximately normal.

The Bayesian model can be further extended to a dynamic model in analogy with Section 4.3. It is assumed here that the β remain constant over time. The analysis of such a model is identical to the analysis of the dynamic Davidson paired comparison model of Section 4.6.2. Once again the parameters can be partitioned into those that are dynamic (the $\gamma^{(t)}$) and those that are non-dynamic (λ, β). Both the iterative and non-iterative analyses of Section 4.6.2 can be extended trivially by allowing for the greater number of non-dynamic parameters.

Chapter 5

Analysis of Chess Game Outcomes

In this chapter, we apply the methodology developed in Chapter 4 to chess tournament outcomes collected over a period of time. The model developed here is an extension of the Bradley-Terry model with ties that incorporates an order effect. The model also incorporates a dynamic component that specifies the evolution of chess players' abilities over time. We describe the World Cup chess tournament data set that is used in our analysis in Section 5.1. In Section 5.2, the model for chess game outcomes is described, and in Section 5.3 we describe the implementation of the model for the World Cup data. We present the results of the analysis and posterior inferences for the rating parameters in Section 5.4, and also show how the model can be used for predictive inferences. This is followed in Section 5.5 by an examination of model diagnostics using posterior predictive checks (Rubin 1984). We report the Gibbs sampler analysis of simulated data with a greater number of games between players in Section 5.6. Section 5.7 presents possible computational improvements, and suggests modifications of the model for chess game outcomes.

5.1 World Cup Chess

To demonstrate the methodology and techniques developed in Chapter 4, we consider a model for chess performance applied to the chess World Cup of 1988–1989. The World Cup consisted of six tournaments and was comprised of 29 of the world's top chess players. Information on the World Cup tournaments was obtained from Kavalek (1990). Table 5.1 lists the dates and sites of each of the six tournaments, and the number of competitors who participated in each tournament. Of the 29 players, 22 players competed in 4 of the 6 tournaments, 3 players competed in 3 of the tournaments, and 4 players only competed in 1 tournament. The World Cup participants who competed in 3 or more tournaments were

Tournament	Dates			Number of Competitors
Brussels, Belgium	April 1, 1988	–	April 22, 1988	18
Belfort, France	June 14, 1988	–	July 3, 1988	16
Reykjavik, Iceland	October 3, 1988	–	October 24, 1988	18
Barcelona, Spain	March 30, 1989	–	April 20, 1989	17
Rotterdam, Netherlands	June 3, 1989	–	June 24, 1989	16
Skellefteå, Sweden	August 12, 1989	–	September 3, 1989	16

Table 5.1: World Cup Chess Tournaments, 1988–1989

contenders for monetary prizes. Table 5.2 lists the players and indicates the tournaments in which each player competed.

For each game in the World Cup, the data consists of the players involved in the game, the outcome of the game (win, loss or draw), an indication of which player played the white pieces (the player with the white pieces moves first), and the tournament in which the game occurred. Each tournament is a single round-robin (i.e., each player plays every other player exactly once), except the first tournament in which one player withdrew after playing four games. The data consists of a total of 789 games spanning 17 months.

5.2 A Model for Chess Game Outcomes

The probability model that we assume for chess game outcomes is an extension of the Bradley-Terry model. The extension, due to Davidson and Beaver (1977), includes a tie as a possible outcome of a comparison, and also incorporates a parameter for a within-pair order effect. We would like to model an order effect because the player with the first move in a chess game is commonly believed to have an advantage. The probabilities of the three game outcomes for tournament t, given that player i moves first, are specified as

$$\begin{aligned}
p^{(t)}_{ij(i)} = \Pr(i \text{ defeats } j) &= \frac{e^{\gamma^{(t)}_i}}{e^{\gamma^{(t)}_i} + e^{\eta+\gamma^{(t)}_j} + e^{\lambda+(\gamma^{(t)}_i+\gamma^{(t)}_j)/2}} \\
p^{(t)}_{ji(i)} = \Pr(j \text{ defeats } i) &= \frac{e^{\eta+\gamma^{(t)}_j}}{e^{\gamma^{(t)}_i} + e^{\eta+\gamma^{(t)}_j} + e^{\lambda+(\gamma^{(t)}_i+\gamma^{(t)}_j)/2}} \\
p^{(t)}_{ij(i).0} = \Pr(i \text{ ties } j) &= \frac{e^{\lambda+(\gamma^{(t)}_i+\gamma^{(t)}_j)/2}}{e^{\gamma^{(t)}_i} + e^{\eta+\gamma^{(t)}_j} + e^{\lambda+(\gamma^{(t)}_i+\gamma^{(t)}_j)/2}},
\end{aligned} \tag{5.1}$$

where the parenthesized subscript (i) indicates that player i moved first.

The model in (5.1) extends the Davidson model of Section 4.6.1 by including the order effect parameter, η. If η takes on a positive value, then the probability of winning for the the player with the first move is lower than for his opponent, whereas when η takes

Player	Brussels	Belfort	Reykjavik	Barcelona	Rotterdam	Skellefteå
Andersson	X	X	X			X
Belyavsky	X	X	X	X		
Ehlvest		X	X		X	X
Hjartarson		X	X	X	X	
Hubner		X		X		X
Illescas				X		
Karpov	X	X			X	X
Kasparov		X	X	X		X
Korchnoi	X		X	X		X
Ljubojevic	X	X		X	X	
Nikolic	X		X	X		X
Noguieras	X	X		X	X	
Nunn	X		X		X	X
Petursson			X			
Portisch	X		X		X	X
Ribli		X	X	X		X
Salov	X			X	X	X
Sax	X		X		X	X
Seirawan	X			X	X	X
Short		X		X	X	X
Sokolov	X	X	X		X	
Spassky		X	X	X		
Speelman	X	X	X	X		
Tal	X		X			X
Timman	X	X	X		X	
Vaganian	X			X	X	X
Vanderwiel					X	
Winants	X					
Yusupov		X	X	X	X	

Table 5.2: World Cup Chess Participants, 1988–1989

on a negative value, the player with the first move has a higher probability of winning. In chess, we would anticipate that η is negative. When η is 0, the model indicates no advantage or disadvantage to moving first, and the model reduces to the Davidson model.

We assume the evolution of the player strength parameters is governed by

$$\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}^{(t-1)} + \boldsymbol{\nu}^{(t)}$$

with

$$\boldsymbol{\nu}^{(t)} \sim \mathrm{N}(0, k^{(t)}\sigma^2 \mathbf{I}), \tag{5.2}$$

where $\boldsymbol{\gamma}^{(t)}$ and $\boldsymbol{\gamma}^{(t-1)}$ are the vectors of rating parameters for tournaments t and $t-1$, respectively. Because tournaments are not separated by equal amounts of time, we define σ^2 to be the variance per month and let $k^{(t)}$ be the number of months between tournaments t and $t-1$. We assume a prior distribution on the parameters

$$(\boldsymbol{\gamma}^{(1)}, \lambda, \eta) \sim \mathrm{N}(\boldsymbol{\mu}^{(1)}, \mathbf{C}^{(1)}).$$

As before, $\boldsymbol{\nu}^{(t)}$ is the change incurred in the player parameters from tournament to tournament. The factor $k^{(t)}$ in (5.2) is justified by treating the amount of variability from month to month as independent, so that $k^{(t)}$ months passing between tournaments corresponds to summing $k^{(t)}$ independent random variables with mean 0 and variance σ^2. For our analysis, the $k^{(t)}$ are known and observed quantities, not necessarily integers. Notice that the variation per month in player abilities is assumed constant over time, and the same for all players. We assume that the draw parameter λ and the color parameter η remain constant over time.

We choose a vague prior distribution for the parameters of our model to reflect our initial uncertainty. The prior parameters are

$$\begin{aligned} \boldsymbol{\mu}^{(1)} &= (0, \ldots, 0, 1, -.25) \\ \mathbf{C}^{(1)} &= 10 \cdot \mathbf{I}. \end{aligned}$$

These prior parameters assign the same mean to each player's rating parameter, but assign a large variance as an indication of our uncertainty. The prior mean of 1.0 for λ and the prior mean of $-.25$ for η imply that players of equal ability will draw games slightly over 60% of the time and that the player who moves first will win 56% of the games that do not result in draws.

The prior distribution on $\omega = 1/\sigma^2$, the precision per month, is also vague, with improper density

$$f(\omega) = \omega^{-3/2}$$

which corresponds to a uniform prior distribution on σ. The values of $k^{(t)}$, $t = 2, \ldots, 6$, for the World Cup data are 1.7, 3.0, 5.2, 1.4, and 1.6. These are the number of months between successive tournaments, and are treated as fixed in the analysis.

Our model for chess game outcomes has strong connections with the Elo chess rating system (Elo 1978) that is used, in two different forms, by the U.S. Chess Federation and the World Chess Federation, as well as other national chess federations. The Elo system assumes that the expected score of a game played between players i and j, where the score is a random variable taking the value 1 for a win, 0 for a loss, and $\frac{1}{2}$ for a draw, is the Bradley-Terry probability $e^{\gamma_i}/(e^{\gamma_i}+e^{\gamma_j})$. The "Elo rating" of a player, R_i, is related to the rating parameter γ_i of the Bradley-Terry model as $R_i = \frac{400}{\ln 10}\gamma_i$. The model in (5.1) improves on some of the drawbacks of Elo's system in that we directly model the probability of a draw as a third outcome, and we include a parameter for an order effect. Also, although Elo's model recognizes that players' abilities may change over time, the updating algorithm does not distinguish between varying amounts of player inactivity.

Many variations of paired comparison models have been proposed for rating chess players. Batchelder and Bershad (1979) propose an extension of the Thurstone-Mosteller model that incorporates ties and updates players' abilities by taking weighted averages of past performances. Joe (1990) examines the world's best chess players of all time by with a model that splits players' careers into "peak" periods and "off-peak" periods. Henery (1992) analyzes the same data set as Joe using an extension of the Thurstone-Mosteller model that allows for draws, and proposes using the length of a game as a covariate for the analysis of a larger chess game data set. In a more developmental approach, Joe (1991) derives axiomatically a general framework for a rating system, and shows how the Elo system is a special case.

5.3 Model Implementation

We describe in this section the implementation of the non-iterative analysis and the Gibbs sampler analysis in order to obtain posterior inferences of parameters for the model of Section 5.2. We indicate below how different time intervals affect the analyses.

5.3.1 Gibbs Sampler Implementation

The implementation of the Gibbs sampler here parallels that of Section 3.1.2. We ran three parallel Gibbs samplers with overdispersed starting values of ω: $1/5^2$, $1/.1^2$, and $1/.001^2$. We believe that $\sigma = 5$ and $\sigma = .001$ are towards the extremes of the posterior distribution of σ, so that our choice of starting values is overdispersed for the target distribution. The starting values for $\gamma^{(t)}$ were obtained by first setting $\gamma^{(1)}$ to 0, and then adding Gaussian noise with standard deviation $\sqrt{k^{(t)}}\sigma$ to $\gamma^{(t)}$ to obtain $\gamma^{(t+1)}$. The initial values in the Gibbs sampler for the η and λ for all t were 0. Each Gibbs sampler proceeds by drawing in sequence from each of the conditional distributions of $\gamma^{(t)}$, given the remaining parameters, followed by a draw from their conditional distribution of (λ, η), followed by a draw from the conditional distribution of ω. Because the time between tournaments vary, drawing from the conditional posterior distribution of $\gamma^{(t)}$ involves the inter-tournament intervals $k^{(t)}$ via their contribution to the between-tournament variances. This does not cause any difficulty in the Gibbs sampler analysis, as the term ω is replaced by $\omega/k^{(t)}$ in the conditional densities. The conditional posterior distribution of ω is

$$(\omega|\gamma_c^{(1)},\ldots,\gamma_c^{(T)},D_T) \sim \mathrm{Gamma}(a_0+p(T-1)/2, b_0+\frac{1}{2}\sum_{t=2}^{T}\frac{1}{k^{(t)}}(\gamma_c^{(t)}-\gamma_c^{(t-1)})^{\mathsf{T}}(\gamma_c^{(t)}-\gamma_c^{(t-1)})),$$

where γ_c refers to the current draw of the player rating parameters in the Gibbs sampler. This gives the appropriate distribution of the precision parameter when tournaments are separated by known but varying numbers of months.

Convergence of the three series was assessed by computing the potential scale reduction as in Gelman and Rubin (1992a). After 500 iterations of the Gibbs sampler for each series, we obtained a potential scale reduction of 195.33 computed on $\log_{10}\omega$. This value provides strong evidence that the Gibbs sampler has not yet sampled from the target distribution. Figure 5.1 shows the values of $\log_{10}\omega$ at each iteration in the three series, and clearly indicates that many more iterations of the Gibbs sampler are necessary before convergence.

Several possible reasons exist for the slow, or perhaps lack of, convergence of the Gibbs sampler for the World Cup Chess data. First, the Gibbs sampler for the World Cup data alternates among $T + 2 = 8$ conditional distributions, whereas the Gibbs sampler for the NFL data alternates between only two conditional distributions. It is suggested in Liu (1992) that a Gibbs sampler that collapses over parameters (i.e., draws from the conditional distribution of several parameters are made simultaneously rather than in sequence) will converge more quickly than one that does not. Additionally, if the joint posterior distribution has parameters that are highly correlated, as is probably the case with the World Cup data due to the time-dependence of the $\gamma^{(t)}$, the Gibbs sampler may take many iterations to move from arbitrary regions of the parameter space to the high likelihood regions under the target distribution. Another explanation for the slow convergence concerns the implementation of the Metropolis step within the Gibbs sampler. In the Gibbs sampler for the NFL data, every trial draw is accepted because we sample directly from the exact conditional distributions. The Gibbs sampler with the Metropolis step for the World Cup data results in acceptable draws 60.7% of the time. This suggests that without accounting for the dependence among the parameters, the World Cup chess analysis would be expected to require about $1/.607 = 1.65$ times as many iterations as the NFL analysis in order to produce the same movement about the parameter space. A third possible reason for the slow convergence is that the information contained in the data is not strong enough to cause the Gibbs sampler to move quickly to regions of the parameter space of high likelihood. Whereas the NFL game score data give a better indication of differences in team abilities, the World Cup data only indicate a player who wins a game (or whether a game ends in a draw).

Because the Gibbs sampler does not appear to converge within 500 iterations, we do not base our analyses of the World Cup data on the results of the Gibbs sampler. We illustrate the Gibbs sampler in Section 5.6 with simulated tournament data consisting of 10 games per player pair within each tournament.

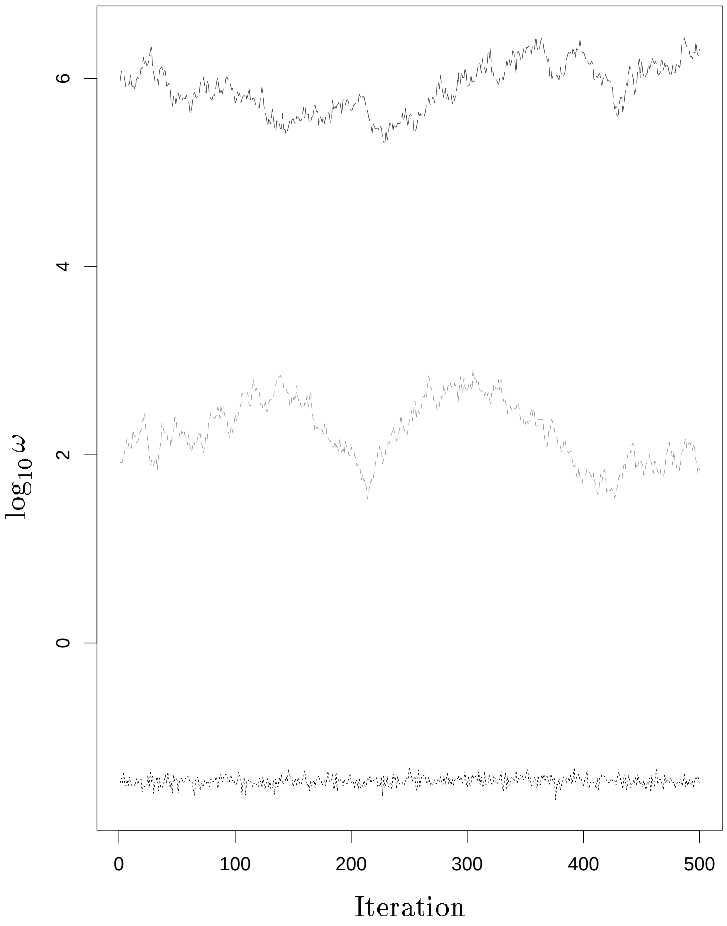


Figure 5.1: Three parallel Gibbs sampler series of $\log_{10}\omega$

5.3.2 Non-iterative Analysis

The non-iterative analysis of Section 4.4 can also be applied to the World Cup model. The distributions of parameters at time t are approximated by multivariate normal distributions centered at the maximum likelihood estimates. The maximum likelihood estimates are obtained by fitting an appropriate GLIM model as described in Critchlow and Fligner (1991). Approximate variance-covariance matrices are obtained according to Palmgren (1981), who describes methodology to obtain asymptotic variances for multinomial regression models. The prior distribution of σ is approximated according to $f(\sigma) = 1/51$ for $\sigma = 1/\sqrt{\omega} \in \{0, \frac{1}{100}, \frac{2}{100}, \ldots, \frac{50}{100}\}$. This set of values is chosen to span the posterior distribution of σ. The inclusion of the time intervals, $k^{(t)}$, requires only minor modifications of the non-iterative analysis. Forecasting between tournaments t and $t+1$ now becomes

$$\mathbf{C}^{(t+1)} = \mathbf{C}'^{(t)} + k^{(t)}\sigma^2\mathbf{J}_{p,2},$$

where $\mathbf{J}_{p,2}$ is the identity matrix with the last two diagonal elements (corresponding to the variance of the change in the draw and order parameters) set to 0. The reweighting factors, $f(d_t|D_{t-1}, \omega_i)$, are computed by the methods of Section 4.4.2, after noting that the forecasting equations involves the addition of $k^{(t)}\sigma^2\mathbf{J}_{p,2}$ rather than $\sigma^2\mathbf{J}_{p,2}$.

5.4 Results of World Cup Analysis

In this section, we show the results of the non-iterative analysis of the World Cup chess data. We examine the posterior distribution of the system standard deviation, σ, and obtain the approximate marginal posterior distributions of $\gamma^{(6)}$, λ, and η. The players might be ranked according to the marginal posterior mean vector, $\boldsymbol{\mu}'^{(T)}$.

5.4.1 Marginal Posterior Distribution of σ

Figure 5.2 displays the approximate posterior distribution of σ, the system standard deviation, from the non-iterative analysis. The figure shows a right-skewed distribution with values of σ near 0 having most support. With the uniform prior, this indicates that the data do not give strong evidence that players' abilities change substantially for the period of the World Cup tournaments. This result makes sense, in light of the long careers of the World Cup competitors, 17 months may not be much time for players' abilities to improve or decline. The approximate posterior mean and standard deviation of σ are 0.115 and 0.0812, respectively. Using the posterior mean of σ as a summary, players' abilities change from month to month by an amount with a standard deviation of roughly 0.115, which corresponds to a 20 point rating change on the Elo scale.

5.4.2 Marginal Posterior Distribution of $\gamma^{(6)}$

For the non-iterative analysis, we compute the posterior distribution of $(\gamma^{(6)}|D_6, \sigma)$ for each $\sigma \in \{0, \frac{1}{100}, \frac{2}{100}, \ldots, \frac{50}{100}\}$. From the previous section, we have the approximate posterior distribution on $(\sigma|D_6)$, so we can compute the approximate posterior means and standard deviations of $\gamma^{(6)}$ marginalized over σ. Table 5.3 displays these values, along with approximate 95% credible intervals for $\gamma^{(6)}$. The 95% credible intervals were computed by generating 3000 values from the marginal distribution of $\gamma^{(6)}$ and computing the 2.5 and 97.5 percentiles empirically. The players are ranked in decreasing order according to the posterior mean of $\gamma^{(6)}$.

The non-iterative analysis indicates that the posterior means range from -3.50 to 1.94, with the current world champion having the highest posterior mean. The standard

Parameter	Mean	Std Dev	95% Credible Interval
Kasparov	1.94	0.78	(0.37, 3.52)
Karpov	1.82	0.78	(0.24, 3.37)
Salov	0.94	0.76	(−0.56, 2.43)
Timman	0.62	0.84	(−0.96, 2.38)
Short	0.57	0.76	(−0.91, 2.09)
Nunn	0.56	0.76	(−0.95, 2.05)
Vanderwiel	0.53	1.03	(−1.46, 2.50)
Ljubojevic	0.51	0.78	(−1.05, 2.04)
Ehlvest	0.46	0.77	(−1.07, 2.01)
Hubner	0.42	0.81	(−1.19, 2.00)
Belyavsky	0.35	0.81	(−1.27, 1.90)
Tal	0.27	0.82	(−1.33, 1.86)
Sokolov	0.23	0.80	(−1.34, 1.81)
Andersson	0.20	0.79	(−1.29, 1.80)
Portisch	0.00	0.76	(−1.54, 1.49)
Seirawan	−0.06	0.76	(−1.51, 1.45)
Vaganian	−0.07	0.78	(−1.61, 1.43)
Ribli	−0.10	0.77	(−1.61, 1.42)
Sax	−0.11	0.77	(−1.62, 1.39)
Speelman	−0.15	0.81	(−1.83, 1.43)
Spassky	−0.25	0.83	(−1.93, 1.33)
Nikolic	−0.25	0.77	(−1.74, 1.24)
Yusupov	−0.28	0.77	(−1.82, 1.21)
Korchnoi	−0.30	0.77	(−1.79, 1.18)
Hjartarson	−0.77	0.78	(−2.29, 0.81)
Noguieras	−0.79	0.78	(−2.30, 0.73)
Petursson	−1.32	1.08	(−3.43, 0.82)
Illescas	−1.47	1.04	(−3.54, 0.60)
Winants	−3.50	1.26	(−6.03, −1.06)
Draw	0.95	0.09	(0.76, 1.14)
Color	−0.48	0.13	(−0.73, −0.23)

Table 5.3: Posterior Distribution of $\gamma^{(6)}$

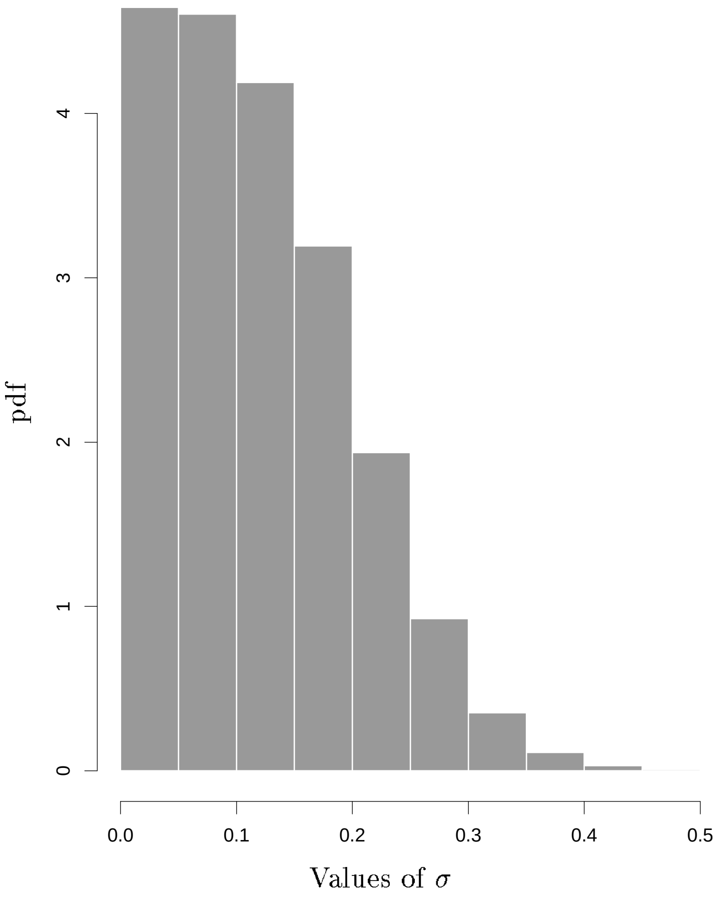


Figure 5.2: Distribution of System Standard Deviation

deviations of the components of $\gamma^{(6)}$ indicate large variability of the posterior distributions. The variation in standard deviations may be explained by the different number of games in which players were involved. For example, Vanderwiel, who played in only one tournament, has a standard deviation of 1.03, whereas players like Kasparov and Karpov who played in four tournaments have a lower standard deviation. The 95% credible intervals show a wide range of values for the player parameters, although the correlations among the $\gamma^{(6)}$ are about 0.5 (not shown on table) so that while the individual standard deviations are large, the standard deviation of the differences in rating parameters are substantially reduced. This makes sense because the data provide information about the difference between the $\gamma_i^{(t)}$ and $\gamma_j^{(t)}$, and not about absolute levels. In fact, if the innovation variance for the model were 0, indicating the rating parameters do not change over time, the posterior correlation among the ratings could be expected to be near 1. The

Player with White	Number of Months after World Cup	Approximate 95% Credible Interval for Probability of Winning	Approximate Median Probability of Kasparov Winning
Kasparov	1	(0.254, 0.572)	0.408
Kasparov	6	(0.218, 0.609)	0.404
Kasparov	12	(0.203, 0.635)	0.407
Kasparov	60	(0.089, 0.807)	0.408
Short	1	(0.166, 0.434)	0.286
Short	6	(0.160, 0.430)	0.282
Short	12	(0.159, 0.433)	0.283
Short	60	(0.124, 0.431)	0.276

Table 5.4: Winning probability estimates for Kasparov against Short

credible intervals for the $\gamma^{(6)}$ appear roughly symmetric around the means, with the 95% intervals spanning about 4 standard deviations, so that a normal distribution may be a reasonable approximation to the marginal posterior distribution of $\gamma^{(6)}$.

The draw and color parameters have posterior means close to their prior parameters. Both the draw and color parameters have small standard deviations, as the model assumes these parameters do not vary over time. The credible interval for the color parameter spanning only negative values indicates a significant advantage to the player with the white pieces. For players of equal rating, the value of $\lambda = .95$ implies that games will be drawn 56.4% of the time, not accounting for an order effect. A value of $\eta = -.48$ implies that a player with the white pieces will win about 61.8% of the games that result in either a win or loss. This corresponds to an 83 point advantage for white on the Elo rating scale.

5.4.3 Forecasting

To demonstrate how the model can be used for forecasting future outcomes, we compute predictive probabilities for games played between Kasparov and Short during future competitions. Table 5.4 shows approximate 95% credible intervals along with the median estimate of the probability that Kasparov will win if the two play 1 month, 6 months, 1 year and 5 years after the World Cup events. The top four lines are computed assuming Kasparov is white and the last four assume Short is white. To compute a given probability, we draw 5000 values from the distribution of $(\omega|D_6)$, and then for each value of ω we compute the approximate prior normal mean and variance of $(\gamma^{(7)}|D_6)$ as $\boldsymbol{\mu}^{(6)}$ and $\mathbf{C}^{(6)} + k^{(7)}\frac{1}{\omega}\mathbf{I}$, respectively. Here, $k^{(7)}$ takes on values 1, 5, 12, and 60 depending on the number of months beyond the completion of the World Cup that we are interested in forecasting.

Table 5.4 shows a large amount of variability in the credible intervals for predicting a Kasparov victory. As more time passes after the World Cup events, the credible intervals become substantially wider. This reflects the increasing uncertainty in game results due to the passage of time. The approximate credible interval when Kasparov plays white after 60 months have passed is exceptionally large, suggesting that the model is a poor forecaster for Kasparov's performance against Short as white. Curiously, however, the increase in the credible interval width when Short plays white is not nearly as not dramatic. This may be due to the reasonable advantage conveyed by playing white.

5.5 Model Adequacy

In this section, we diagnose the adequacy of the model fit in the previous section using the posterior distribution of carefully chosen diagnostic statistics (Rubin 1984). The idea behind posterior predictive model checks is as follows. First, a model is fit to observed data. Given the posterior distribution of the parameters, replications or simulated data sets are generated by drawing "likely" sets of parameters, and then generating simulated data conditional on these parameters. Informative statistics are constructed to test the adequacy of the model against reasonable alternatives. The distribution of the diagnostic statistics over the simulated data can be used as a reference distribution for the value of the diagnostic computed on the observed data. If the statistic for the observed data is an extreme value relative to the distribution of the statistic over the simulated data, then the adequacy of the model is called into question. We examine the adequacy of our model using posterior predictive checks in Section 5.5.1. The posterior checks indicate problems in either the computational precision of the analysis or in the model specification. Accordingly, in Section 5.5.2 we simulate tournament data and perform posterior predictive checks to examine whether the computational approximations used in the analysis are too inaccurate for a moderate sized data set like the World Cup.

5.5.1 Posterior Predictive Checks for the Model

Posterior predictive checks were performed by simulating tournament data according to the posterior distribution of the parameters. To simulate a single sequence of six tournaments, we first draw from the joint posterior distribution of all parameters. This is done by carrying out the following steps.

1. Draw ω from the approximate (discrete) posterior distribution of $(\omega|D_6)$.
2. Draw $(\lambda, \eta|\omega)$ from their normal posterior distribution. Section 4.4.1 describes the procedure for obtaining the approximate normal distribution of the parameters after the last tournament.
3. Conditional on ω and (λ, η), compute the parameters of the approximately normal conditional likelihood of $(\gamma^{(t)}|\lambda, \eta, \omega, d_t)$ for each single tournament at time t. Using

the procedure described in Section 4.6.2 for the analysis of the Gibbs sampler, compute the parameters for the distribution of $(\gamma^{(6)}|\lambda, \eta, \omega, D_6)$, and then draw once from this distribution.

4. For $t = 5, \ldots, 1$, draw $\gamma^{(t)}$ from the approximately normal conditional distribution given the previous draws of $\gamma^{(t+1)}, \ldots, \gamma^{(6)}, \lambda, \eta, \omega$. This is accomplished using the procedure in Section 4.6.2 for drawing from the joint posterior distribution of parameters given ω in the Gibbs sampler.

For each of the six tournaments, we generate simulated game outcomes according to the actual tournament design, conditioned on the parameters drawn from the posterior distribution. The game outcomes are generated according to the outcome probabilities of the extended Davidson-Beaver model according to Section 5.2. This process was performed 500 times to generate a sample of 500 collections of 6 simulated World Cup tournaments. For each collection of 6 tournaments, C_i, $i = 1, \ldots, 500$, we compute the diagnostic statistic $T(C_i)$ relevant for detecting model failures.

We performed two tests of adequacy of the model. The first diagnostic statistic is the maximum proportion of wins per player out of games that result in either a win or loss. An unusually large or small value would indicate that the model does not capture the player to player variability. To check whether a single tie parameter sufficiently describes the variation in drawn games across players, the second diagnostic computes the maximum across players of the proportion of drawn games. Figure 5.3 shows the distribution of each of the two diagnostic statistics for the collection of 500 simulated World Cup matches along with a vertical line representing the statistic computed from the actual data. The first plot shows that the maximum proportion of wins for any player from the actual data is well within the range of values for the simulated tournaments. This suggests that the model parametrization sufficiently captures the strength of the players by a single rating parameter per player varying over time. The second plot, on the other hand, indicates a possible problem with the model. The maximum proportion of draws across players for the actual data is at the right extreme of the distribution for simulated data. The model does not generate data sets with as much variability in draws as the observed data indicates. This implies that a single tie parameter may not be adequately describing the frequency with which players draw games. If a single tie parameter were a sufficient description, the plot shows that the player with the greatest percentage of draws would be expected to draw about 70–75% of the time, whereas the actual data yields a player (Ribli) who draws 86% of his games.

While the preceding analysis suggests that alternative models be explored for describing chess game outcomes for the World Cup tournaments, the problem may stem from difficulties with the computational approximations. The posterior distribution of parameters is obtained by approximating each of the likelihoods by a multivariate normal density, and as each tournament consists of one comparison for each pair, the likelihood may not be adequately approximated by a normal density. Furthermore, in updating parameters to include a new tournament, the posterior distribution of parameters is not approximated

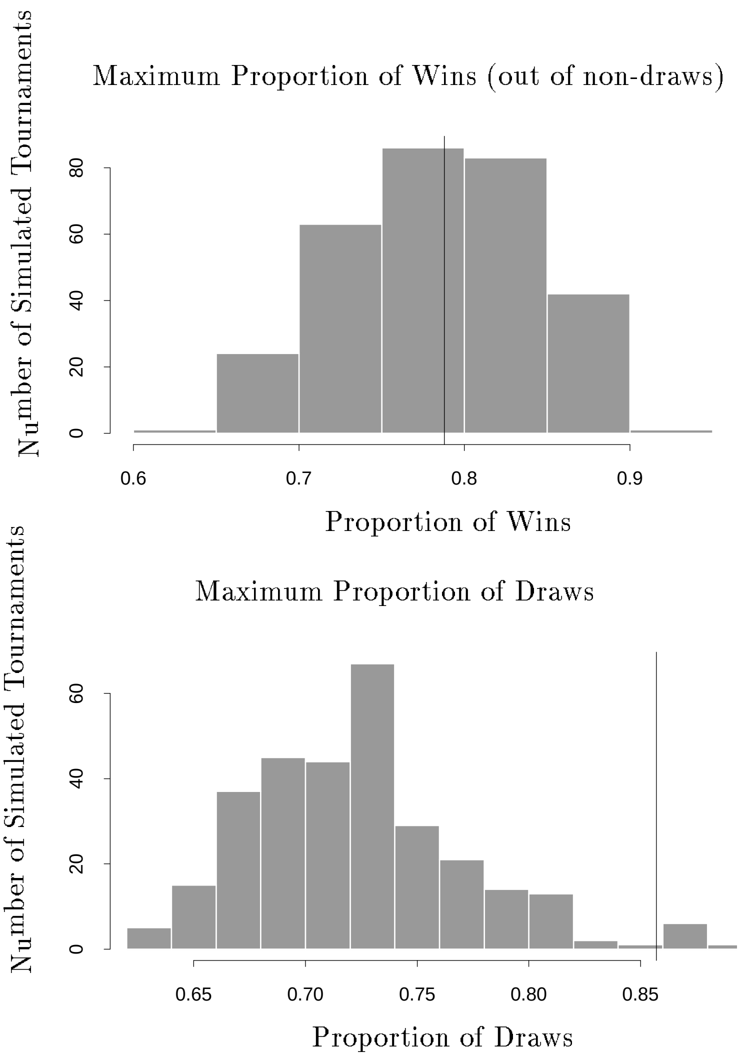


Figure 5.3: Distribution of test statistics for model adequacy

at the exact posterior mode, but at a parameter value that is close to the posterior mode only if the likelihood is approximately normal (see Section 4.2). To illustrate how the normal approximation for the World Cup data may contribute to the inadequacy of the model, Table 5.5 displays the posterior means of $(\gamma^{(6)}, \lambda, \eta)$ in two situations. The first column of means is computed using the approximately conjugate updating of Section 4.4 given $\sigma = 0$ with the prior distribution specified in Section 5.2, except that the prior variances are 100 rather than 10 to allow the data to dominate the distribution of parameters. This column assumes that each of the six likelihoods can be reasonably approximated by a multivariate normal density. The second column of means is computed by treating all six World Cup tournaments as a single large tournament, and finding the posterior means by updating the prior parameters exactly once. Because $\sigma = 0$, the two columns will be similar if the normal approximations are satisfactory. The parameter means in the second

Parameter	Posterior Mean Sequential Updates	Posterior Mean Single Update
Andersson	0.20	0.19
Belyavsky	0.44	0.46
Ehlvest	0.55	0.62
Hjartarson	–0.68	–0.75
Hubner	0.47	0.51
Illescas	–1.54	–1.67
Karpov	1.90	2.10
Kasparov	2.02	2.24
Korchnoi	–0.36	–0.44
Ljubojevic	0.55	0.55
Nikolic	–0.23	–0.27
Noguieras	–0.77	–0.81
Nunn	0.59	0.62
Petursson	–1.39	–1.51
Portisch	–0.01	0.03
Ribli	–0.09	–0.05
Salov	1.00	1.08
Sax	–0.18	–0.17
Seirawan	–0.09	–0.05
Short	0.58	0.64
Sokolov	0.21	0.24
Spassky	–0.22	–0.19
Speelman	–0.08	–0.07
Tal	0.28	0.29
Timman	0.39	0.33
Vaganian	–0.02	–0.09
Vanderwiel	0.57	0.61
Winants	–3.82	–4.08
Yusupov	–0.27	–0.34
Draw	0.95	1.06
Color	–0.47	–0.51

Table 5.5: Posterior means of $(\gamma^{(6)}, \lambda, \eta)$ with $\sigma = 0$

column reflect the data more accurately because the computation only relies on a single normal approximation. While the parameter means for individual parameters are close, some are far enough apart to be of some concern. For example, the posterior mean for λ differs by over 0.1, which, according to Table 5.3, is more than one posterior standard deviation. So the analyses that produce Figure 5.3 may indicate that the computational approximations are too coarse.

5.5.2 Evaluating the Effect of the Normal Approximation

Model monitoring via predictive checks in Section 5.5.1 raise doubts about whether the proposed model for chess game outcomes of Section 5.2 is appropriate. The preceding section suggests that the failure may occur because the normal approximations to the Davidson-Beaver likelihoods are not accurate enough. We provide evidence in this section that when players compete a moderate number of times in each tournament, the normal approximation of the likelihoods does not disturb the predictive checks.

The data used for the analysis in this section was obtained by simulating the World Cup tournaments with the exact same tournament schedule, except players would compete against each other four times (twice with each color) instead of just once. This resulted in a single simulated World Cup consisting of $4 \times 789 = 3156$ game outcomes.

Game outcomes are generated as follows. Strength parameters, $\gamma^{(1)}$, for the first tournament were generated from $\mathrm{N}(0,1)$. For successive tournaments, $\gamma^{(t)}$ was obtained by adding Gaussian noise with a variance of $.25k^{(t)}$, where $k^{(t)}$ is the number of months between tournaments $t-1$ and t, given in Section 5.2. This is equivalent to setting $\sigma = .5$. Also, we set $\lambda = 1$ and $\eta = -.5$. From these parameters, we generated data for the t-th tournament according to the World Cup tournament schedule with four games per pair rather than just one.

Using the non-iterative methodology of Section 4.4, the posterior distribution of the parameters was computed from the simulated data. Predictive checks on the data were performed by generating 150 sets of tournaments consisting of 3156 games each. Figure 5.4 displays the distribution of the diagnostic statistics used in Section 5.5.1 for the simulated data. As in Figure 5.3, the vertical line on the plots represents the initial simulated World Cup data, and the histogram represents the 150 replicated sets of tournaments. In contrast to Figure 5.3, both the observed statistics fall within the body of the simulated reference data. These two predictive checks indicate that it may be appropriate to use the normal approximation of the likelihoods and the approximate parameter updates when the number of games per player pair is as few as four.

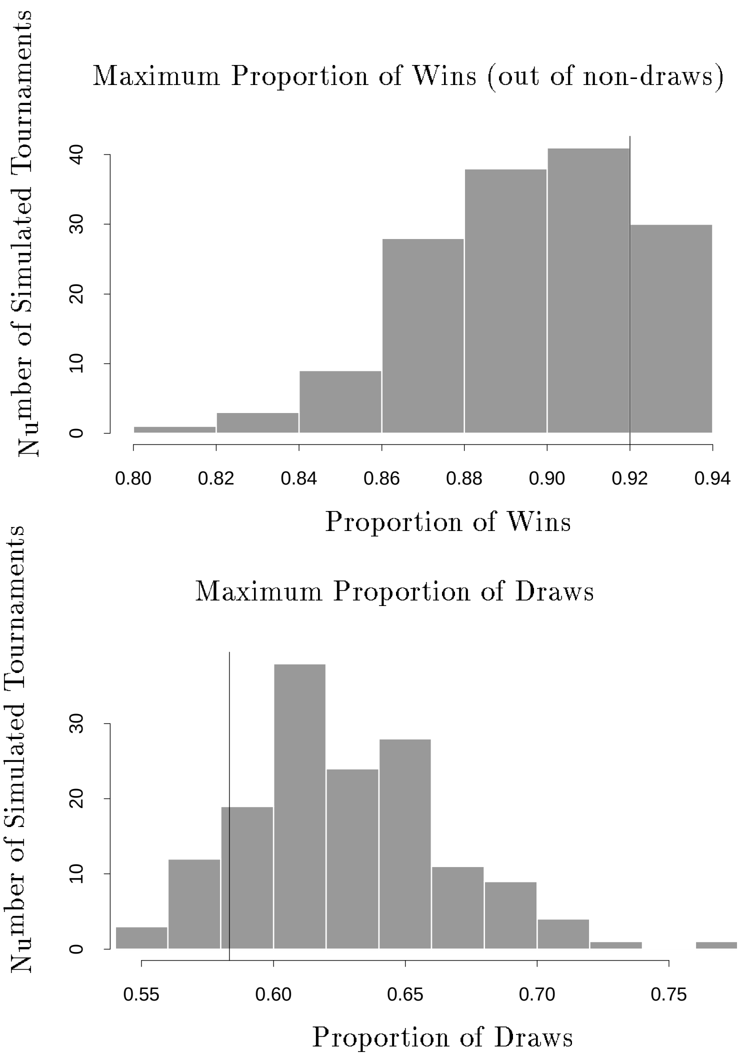


Figure 5.4: Distribution of test statistics for simulated data

5.6 Examination of the Gibbs sampler on Simulated Chess Outcomes

Section 5.3.1 has shown that the Gibbs sampler failed to converge in 500 iterations for the analysis of the World Cup chess data, possibly because the data from single round-robin tournaments does not contain a great deal of information about the innovation variance, σ^2. In this section, we generate simulated tournament data with known innovation variance and perform the Gibbs sampler at dispersed starting values for ω to explore the effect of increasing the amount of data into the analysis.

The data are generated as follows. Seven players are assigned values of $\gamma^{(1)}$ by making random draws from a standard normal distribution. For $t = 2, \ldots, 6$, we obtain $\gamma^{(t)}$ by adding independent random draws from a standard normal distribution to $\gamma^{(t-1)}$. This process results in known values of $\gamma^{(t)}$ for 7 players over 6 equally spaced time periods. Although the $\gamma^{(t)}$ are generated using $\sigma^2 = 1$, due to sampling variability the mean squared difference among the $\gamma^{(t)}$ and $\gamma^{(t-1)}$ is 0.548 for the sample. We assign $\eta = -0.25$ and $\lambda = 1.0$. We then generated game outcomes for ten round-robin tournaments for each t according to the model of Section 5.2 using the ratings $\gamma^{(t)}$ and the values of η and λ. At each t, each competitor played every other ten times; five times as white and five times as black.

The Gibbs sampler with a Metropolis step was performed as described in Section 5.3.1. Four parallel samplers were run with starting values for ω of $1/5^2$, $1/2^2$, $1/.5^2$ and $1/.01^2$. The initial values of $\gamma^{(1)}$ were set to 0, and values of $\gamma^{(t)}$ for $t = 2, \ldots, 6$ were obtained by adding Gaussian noise with variance $1/\omega$ to $\gamma^{(t-1)}$. The Metropolis step accepted new draws at a rate of approximately 82%.

Figure 5.5 shows the value of ω drawn at each iteration for all four samplers. The sampler series corresponding to starting values $\omega = 1/2^2, 1/.5^2, 1/.01^2$, appear to move towards the correct stationary distribution for this parameter. The sampler series starting at $\omega = 1/5^2$ shows a lack of convergence to the correct distribution. To examine whether the difficulty at $\omega = 1/5^2$ is a result of the likelihood of the simulated data not containing sufficient information, the sampler was rerun with 100 replications of the simulated data (1000 comparisons per player pair). In this case the series moved quickly towards the correct stationary distribution.

Several possibilities exist for explaining the problems associated with the Gibbs sampler analysis on the simulated data. The information contained in the simulated data, which consists of 10 game outcomes per player pair, may still not be sufficient to guarantee convergence for the dynamic paired comparison model. With 100 times as many observations, the data appear to have enough of an effect to make the Gibbs sampler converge quickly. A possibility that may explain Figure 5.5 is that the Gibbs sampler is at a minor mode in the posterior distribution of ω. If the distribution is reasonably flat in a neighborhood of this mode, the Gibbs sampler may take many iterations to jump out of this region of the parameter space to find the highest likelihood regions. This phenomenon is discussed in Gelman and Rubin (1992b).

5.7 Discussion

Analyzing paired comparison data using the techniques of Chapter 4 appears to be feasible when the paired comparison model is correct and a reasonable number of games are played between pairs of players. We find, however, that when the number of games in the data set is small relative to the number of parameters, the analysis of game outcomes using the methodology of Chapter 4 may not provide accurate inferences. Diagnostic checks of the proposed model for the World Cup indicate that improvements are needed, either

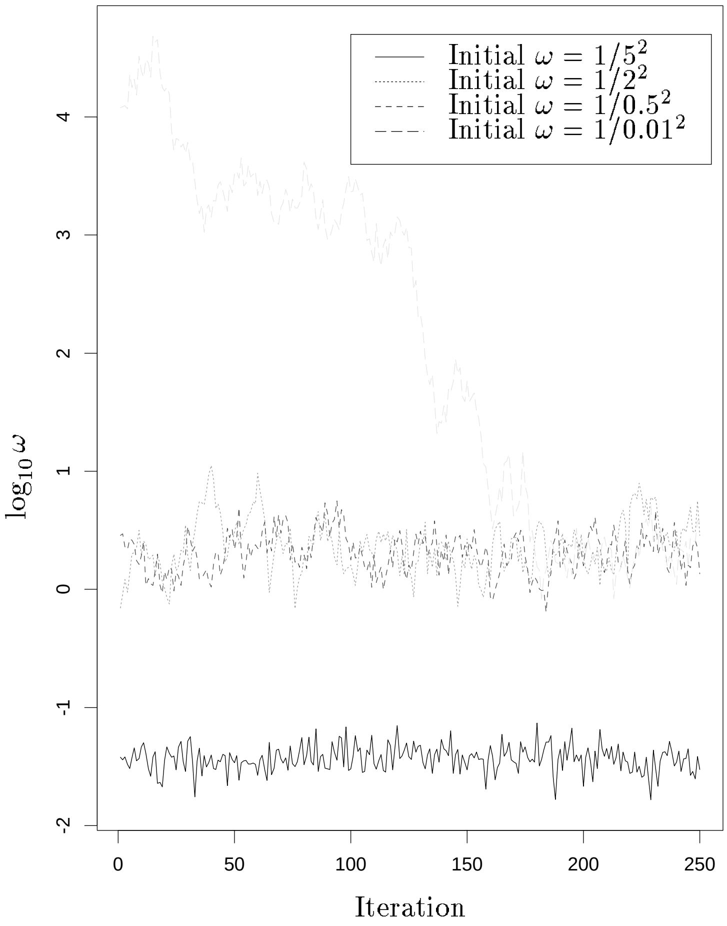


Figure 5.5: Values of $\log_{10}\omega$ for Gibbs sampler on simulated data

in computing the posterior distribution or in the model specification. In this section, we discuss approaches to improving the analysis of the World Cup data set.

One feature of the World Cup data set and of other sports competitions is that within each tournament players compete against each other at most once. For a p-player system, a single round-robin tournament involves $p(p-1)/2$ comparisons and $p+2$ parameters. Portnoy (1988) shows that for exponential families with p parameters and n observations, if p^2/n is not small, the likelihood may not be well approximated by a normal density. In the case of single round-robin tournaments, this ratio can be viewed as large. The updating of parameters described in Section 4.4.1 requires the likelihood to be approximately normal in order to justify the conjugate analysis. Because the likelihoods in the World Cup data set may not be reasonably approximated by normal densities, more accurate methods to fit the model seem necessary.

The computation can be made more accurate for the small sample analysis by computing the exact posterior mode before making the normal approximation as well as computing the Hessian matrix evaluated at the exact mode at each updating. In Section 4.4.1 we just approximate the likelihoods by a normal distribution centered at the maximum likelihood estimate, and then combining this normal approximation with the normal prior distribution. When the likelihood is normal, this gives the exact posterior mean; otherwise, the result is an approximation which is coarse if the likelihood is not well approximated by a normal density. Instead, the exact posterior mode can be found by performing the Newton-Raphson algorithm on the product of the prior density and the exact likelihood. Then the normal approximation can be used directly on the posterior distribution, centered around the posterior mode. The difference in computational cost of using this approach in place of the approach of Section 4.4.1 is that for every ω_i in the discrete space $\{\omega_1, \ldots, \omega_m\}$, the Newton-Raphson algorithm is invoked T times (once for each tournament) to obtain the posterior distribution of parameters after each tournament. Thus, mT Newton-Raphson calculations are performed. The approach taken in Section 4.4.1 requires the maximum likelihood estimates to be found only once for each tournament, resulting in a total of T Newton-Raphson analyses. It should be noted that even finding the correct posterior mode after each update and approximating the likelihoods by normal densities at this true mode may not be an adequate approximation if the likelihood is substantially non-normal. In such a case, accurate inferences may be obtained using iterative simulation such as through the Gibbs sampler.

The problem using the Gibbs sampler is clearly the computational burden. For the World Cup data, the Gibbs sampler did not converge within 500 iterations. The rate of convergence has been shown to be related to the dependence among the parameter distributions (Liu 1992), and with σ small, the dependence of the $\gamma^{(t)}$ across t is large so that the convergence can be expected to be slow. Using the Gibbs sampler for analyzing actual paired comparison data sets, therefore, may not be computationally feasible.

Putting aside computational issues, the probability model for chess game outcomes may be further extended. The most likely misspecification of the model is that a single tie parameter, λ, is not sufficient for describing players' tendencies to draw games. An alternative model replaces λ by $(\lambda_i + \lambda_j)/2$ when players i and j compete, where λ_i is an individual tie parameter for player i. This model may more realistically reflect the notion that some players are more likely to draw games than others. The probability of a draw is now a function of the average of the draw parameters for the two players involved in a game, and the larger the average, the more likely the outcome of the game will be a draw. Glickman (1991) proposes this parametrization for the draw parameters, substituting the average of draw parameters in the context of a model proposed by Rao and Kupper (1967). Joe and White (1992) examine a model for chess players that replaces $\exp((\lambda_i + \lambda_j)/2)$ in the Glickman model by $\exp(\lambda_i) + \exp(\lambda_j)$.

Another class of alternative models involve treating either λ or η (or both) as varying over time. As chess theory develops, the understanding of the game may change such that the frequency of draws among top players may change, and the advantage conferred to white may also change. For the World Cup data set, the playing conditions at certain tournaments may have been such that the players choose to end their games in quick draws. Another possibility is that if a player is not feeling well at a particular tournament, he may attempt to draw games more often than usual. A model with time varying λ and η would require an evolution variance parameter for the time-varying parameters which would need to be estimated.

Chapter 6

Conclusions

This thesis presents paired comparison models in which the merits of the objects being compared may change over time. Chapters 2 and 3 address paired comparison experiments with a continuous outcome variable, while chapters 4 and 5 address paired comparison experiments with indicator outcome variables. The models developed in this thesis have strong connections to the Bayesian dynamic models of West and Harrison (1990). In the models we consider, every object involved in a comparison has an associated rating parameter which measures the relative performance or ability of the object. The observation equation of the model relates the outcome of a comparison to the rating parameter, and the system equation describes the evolution of the rating parameters over time. This thesis demonstrates methods for obtaining posterior inferences for the ratings under the model and forecasts for future comparisons.

Two approaches to the data analysis are examined. One approach bases inferences on samples drawn from the posterior distribution of parameters using the Gibbs sampler. The other approach approximates the prior distribution of the system variance, σ^2, by a discrete distribution on a grid of values. This approximation results in a more tractable analysis when marginalizing over the posterior distribution of σ^2. This second approach proves to be the more successful of the two in practice. The analyses of NFL football game outcomes by the Gibbs sampler and by the non-iterative approach result in similar inferences, suggesting that the discrete approximation to the prior distribution of σ^2 was entirely satisfactory. In the analysis of the World Cup chess game outcomes, the Gibbs sampler does not converge in a computationally feasible amount of time. While the computational approximations for the non-iterative analysis of the World Cup data present some difficulties, these can be improved at only a moderate computational cost.

The approach we present in this paper provides a flexible means to modeling paired comparison data. The framework is quite general and extensions to these models, such as incorporating non-dynamic parameters or specifying alternative tie parameters, can be incorporated without changing the underlying structure of the basic models.

Bibliography

Batchelder, W. H. and Bershad, N. J. (1979), "The statistical analysis of a Thurstonian model for rating chess players," *Journal of Mathematical Psychology*, 19, 39–60.

Besag, J. (1974), "Spatial interactions and the statistical analysis of lattice systems (with discussion)," *J. Roy. Statist. Soc. Ser. B*, 36, 192–235.

Bradley, R. A. and Gart, J. J. (1962), "The asymptotic properties of ML estimators when sampling from associated populations," *Biometrika*, 49, 205–14.

Bradley, R. A. and Terry, M. E. (1952), "The rank analysis of incomplete block designs. 1. The method of paired comparisons," *Biometrika*, 39 , 324–345 .

Bühlmann, H. and Huber, P. J. (1963), "Pairwise comparisons and ranking in tournaments," *Annals of Mathematical Statistics*, 34, 501–510 .

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

Chen, C. and Smith, T. M. (1984), "A Bayes-type estimator for the Bradley-Terry model for paired comparison," *Journal of Statistical Planning and Inference*, 10, 9–14.

Critchlow, D. E. and Fligner, M. A. (1991), "Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM," *Psychometrika*, 56, 517–533.

David, H. A. (1988), *The Method of Paired Comparisons* , Oxford University Press (Oxford, New York).

Davidson, R. R. (1970), "On extending the Bradley-Terry model to accommodate ties in paired comparison experiments," *Journal of the American Statistical Association*, 65, 317–328.

Davidson, R. R. and Beaver, R. J. (1977), "On extending the Bradley-Terry model to incorporate within-pair order effects," *Biometrics*, 33, 693–702.

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

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), "Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)," *J. Roy. Statist. Soc. Ser. B*, 39, 1–38.

Dillon, W. and Goldstein, M. (1984), *Multivariate Analysis, Methods and Applications*, John Wiley & Sons (New York).

Elo, A. E. (1978), *The Rating of Chessplayers, Past and Present*, Arco Publishing, Inc. (New York).

Ford, L. R., Jr. (1957), "Solution of a ranking problem from binary comparisons," *American Math Monthly*, 66, 28–33.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and 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. and Smith, A. F. M. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A. and King, G. (1990), "Estimating the electoral consequences of legislative redistricting," *Journal of the American Statistical Association*, 85, 274–282.

Gelman, A. and Rubin, D. B. (1992a), "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7, 457–511.

——— (1992b), "A single series from the Gibbs sampler provides a false sense of security," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Oxford University Press, 625–632.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Geyer, C. J. (1992), "Practical Markov chain monte carlo," *Statistical Science*, 7, 473–511.

Glickman, M. (1991), "A Bayesian paired comparison model for rating chess players," Technical Report, Department of Statistics, Harvard University.

Gumbel, E. J. (1961), "Sommes et différences de valeurs extrêmes indépendantes," *C. R. Acad. Sci. Paris* , 253, 2838–2839.

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

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

Henery, R. J. (1986), "Interpretation of average ranks," *Biometrika*, 73, 224–227.

——— (1992), "An extension to the Thurstone-Mosteller model for chess," *The Statistician*, 41, 559–567.

Joe, H. (1988), "Majorization, entropy and paired comparisons," *Annals of Statistics*, 16, 915–925.

——— (1990), "Extended use of paired comparison models, with application to chess rankings," *Applied Statistics*, 39, 85–93.

——— (1991), "Rating systems based on paired comparison models," *Statistics & Probability Letters*, 11, 343–347.

Joe, H. and White, R. A. (1992), "Paired comparison models with time-varying abilities for chess," Technical Report, Department of Statistics, University of British Columbia.

Kalman, R. E. (1960), "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, 82, 34–45.

Kalman, R. E. and Bucy, R. S. (1961), "New results in linear filtering and prediction theory," *Journal of Basic Engineering*, 83, 95–108.

Kavalek, L. (1990), *World Cup Chess*, Trafalgar Square Publishing (North Pomfret, Vermont).

Kong, A., Liu, J. and Wong, W. H. (1991), "Sequential imputations and Bayesian missing data problems," Technical report no. 321, Department of Statistics, The University of Chicago, and Department of Statistics, Harvard University.

Leonard, T. (1977), "An alternative Bayesian approach to the Bradley-Terry model for paired comparisons," *Biometrics*, 33, 121–132.

Liu, J. (1992), "The collapsed Gibbs sampler and other issues: with applications to a protein binding problem," Research Report R-426, Department of Statistics, Harvard University .

Luce, R. D. (1959), *Individual Choice Behavior*, Wiley (New York).

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, Chapman & Hall (London).

Meinhold, R. J. and Singpurwalla, N. D. (1983), "Understanding the Kalman filter," *American Statistician*, 37, 123–127.

Meng, X. L. and Rubin, D. B. (1991), "Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm," *Journal of the American Statistical Association*, 86, 899–909.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), "Equations of state calculations by fast computing machines," *Journal of Chemical Physics*, 21, 1087–1092.

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.

Palmgren, J. (1981), "The Fisher information matrix for log linear models arguing conditionally on observed explanatory variables," *Biometrika*, 68, 563–566.

Portnoy, S. (1988), "Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity," *Annals of Statistics*, 16, 356–366.

Raiffa, H. and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Division of Research, Harvard Business School.

Rao, P. V. and 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.

Rosner, B. (1976), "An analysis of professional football scores," in *Management Science in Sports*, eds. R. E. Machol, S. P. Ladany and D. G. Morrison, North-Holland (New York), 67–78.

Rubin, D. B. (1984), "Bayesianly justifiable and relevant frequency calculations for the applied statistician," *Annals of Statistics*, 12, 1151–1172.

Sallas, W. M. and Harville, D. A. (1988), "Noninformative priors and restricted maximum likelihood estimation in the Kalman filter," in *Bayesian Analysis of Time Series and Dynamic Models*, ed. J. C. Spall, Dekker (New York), 477–508.

Singh, J. and Gupta, R. S. (1975), "Derivation of a paired comparison model," *Applied Stat., Proc. of Conf. at Dalhousie U., Halifax, North Holland.*

_______ (1978), "A paired comparison model allowing for ties," *Scandinavian Journal of Statistics*, 5, 65–68.

Stefani, R. T. (1977), "Football and basketball predictions using least squares," *IEEE Trans. on Syst., Mgmt., Cybernetics*, 7, 117–120.

_______ (1980), "Improved least squares football, basketball, and soccer predictions," *IEEE Trans. on Syst., Mgmt., Cybernetics*, 10, 116–123.

Stern, H. (1991), "On the Probability of Winning a Football Game," *American Statistician*, 45, 179–183.

_______ (1992a), "Who's number one? – Rating football teams," *Proceedings of the Section on Statistics in Sports of the American Statistical Association.*

_______ (1992b), "Are all linear paired comparison models empirically equivalent," *Mathematical Social Sciences*, 23, 103–117.

Thisted, R. A. (1988), *Elements of Statistical Computing*, Chapman & Hall (London).

Thompson, M. (1975), "On any given Sunday: Fair competitor orderings with maximum likelihood methods," *Journal of the American Statistical Association*, 70, 536–541.

Thurstone, L. L. (1927), "A law of comparative judgment," *Psychological Review*, 34, 273–286.

West, M. and Harrison, J. (1990), *Bayesian Forecasting and Dynamic Models*, Springer-Verlag (Berlin, New York).

Yellott, J. I. (1977), "The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution," *Journal of Mathematical Psychology*, 15, 109–144.

Zeger, S. L. and Karim, M. R. (1991), "Generalized linear models with random effects: a Gibbs sampling approach," *Journal of the American Statistical Association*, 86, 79–86.

Zermelo, E. (1929), "Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung," *Math. Zeit.*, 29, 436–460.