Journal of Applied Statistics, Vol. 28, No. 6, 2001, 673-689

Dynamic paired comparison models with stochastic variances

MARK E. GLICKMAN, *Boston University, Boston, MA 02215, USA*

ABSTRACT *In paired comparison experiments, the worth or merit of a unit is measured through comparisons against other units. When paired comparison outcomes are collected over time and the merits of the units may be changing, it is often convenient to assume the data follow a non-linear state-space model. Typical paired comparison state-space models that assume a fixed (unknown) autoregressive variance do not account for the possibility of sudden changes in the merits. This is a particular concern, for example, in modeling cognitive ability in human development; cognitive ability not only changes over time, but also can change abruptly. We explore a particular extension of conventional state-space models for paired comparison data that allows the state variance to vary stochastically. Models of this type have recently been developed and applied to modeling financial data, but can be seen to have applicability in modeling paired comparison data. A filtering algorithm is also derived that can be used in place of likelihood-based computations when the number of objects being compared is large. Applications to National Football League game outcomes and chess game outcomes are presented.*

1 Introduction

Paired comparison data arise when objects are compared to elicit a preference or a degree of preference. The literature on paired comparison modeling is vast, spanning fields such as statistics, marketing, psychology and decision sciences. Background on fundamental issues in paired comparison modeling along with examples can be found in David (1988) and Bradley (1984). A common situation is to observe paired comparison data over time where the underlying value or worth of the objects are changing. This might occur, for example, in comparing preferences towards the value of marketed products or services, or in the outcomes of games played between competitors whose abilities may be changing over time.

Correspondence: Mark E. Glickman, Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, MA 02215, USA. E-mail: mg@math.bu.edu.

ISSN 0266-4763 print; 1360-0532 online/01/060673-17
DOI: 10.1080/02664760120059219

Recent works by Glickman (1993, 1999), Fahrmeir & Tutz (1994) and Knorr-Held (2000) have adopted a state-space approach to modeling such data where the underlying merit parameters change as a Gaussian stochastic process. While this approach can often appropriately describe the change in merits, it can be too restrictive if the merits can undergo sudden shifts, or if interventions occur that change the merit of an object quickly. For example, in a marketing context, if a product is reported to be defective or dangerous, it will likely lose merit quickly. When comparing human cognitive skill through games such as chess, younger players may undergo quick increases in ability (Simonton, 1997), which the simpler state-space models cannot capture.

This paper describes an extension of the usual state-space models for paired comparison data that allows for sudden movement in the underlying merit parameters. The extension is closely related to the stochastic volatility model (Jacquier *et al.*, 1994; Capobianco, 1996; Uhlig, 1997) developed in the context of modeling financial time series data. Our model involves letting not only the merit of objects change stochastically, but also letting the variance of the state process change stochastically. In this extension, sudden shifts in merit are reflected through the variance of the change in merits becoming large. In Section 2, the model allowing the variance of the state process to vary stochastically is developed. This is followed in Section 3 by an application of the model to a data set on football game outcomes. A filtering algorithm is then presented in Section 4, extending an algorithm developed in Glickman (1999), which approximates likelihood-based computations using nearly closed form calculations. A situation in which one might want to use such an algorithm is when many objects are being compared, or many time periods are involved, so that exact likelihood-based methods become computationally intractable. This approach is demonstrated in Section 5 on a data set consisting of chess games played among the best players of all time. We provide a summary of the model and consider directions for extensions in Section 6.

2 A stochastic variance paired comparison model

Suppose $K_{ij}^{(t)}$ comparisons are to take place between objects i and j at time t. Time is assumed to be discretized into periods of equal duration, so that t, which indexes a time interval, takes on integer values. We treat data observed within a time period as occurring at the start of the period. Let $Y_{ijk}^{(t)}$ be 1 if i is preferred to j in the kth comparison between the two objects at time t, and 0 if j is preferred. Let $\gamma_i^{(t)}$ and $\gamma_j^{(t)}$ be the merits of the objects at time t. We assume for fixed t that the probability i is preferred to j during the kth comparison, given by

$$\mathrm{P}(Y_{ijk}^{(t)} = 1) = F(\gamma_i^{(t)} - \gamma_j^{(t)}, \theta, \mathbf{x}_i^{(t)}, \mathbf{x}_j^{(t)}) \tag{1}$$

where F is a specified probability function monotonically increasing in $\gamma_i^{(t)} - \gamma_j^{(t)}$, θ is a vector of other model parameters and $\mathbf{x}_i^{(t)}$ and $\mathbf{x}_j^{(t)}$ are covariate information for objects i and j at time t. This model, the linear paired comparison model, assumes that preference probabilities are functions of the merit parameters only through their difference. When there are no other model parameters or covariates, two common special cases of this model include the Bradley-Terry model (Bradley & Terry, 1952) when F is a standard logistic distribution function, and the Thurstone-Mosteller model (Thurstone, 1927; Mosteller, 1951) when F is a Gaussian distribution function. In practice, the choice among paired comparison models can usually be assessed only with a large amount of data (Stern, 1992). The model

specification in equation (1) is sufficiently general to include many extensions of basic linear paired comparison models.

The parameters $\gamma_i^{(t)}$ are assumed to change over time through a stochastic process. Our model assumes for each object i,

$$\gamma_i^{(t+1)} | \gamma_i^{(t)}, \sigma_i^{2^{(t+1)}} \sim \mathrm{N}(\gamma_i^{(t)}, \sigma_i^{2^{(t+1)}}) \tag{2}$$

so that the innovations in merit follow a normal distribution centered at zero, and with a variance that depends on time. Alternative formulations of this component of the model can allow for an autoregressive parameter, as we demonstrate in Section 3, or a multivariate model that imposes linear restrictions. The latter formulation was adopted in Glickman & Stern (1998).

We further assume a model for the change in variance,

$$\log \sigma_i^{2^{(t+1)}} | \sigma_i^{2^{(t)}}, \tau^2 \sim \mathrm{N}(\log \sigma_i^{2^{(t)}}, \tau^2) \tag{3}$$

This aspect of the model allows for the innovations in the process for $\gamma_i^{(t)}$ to have a variance that may be stochastically varying. Because adding the same constant to all the $\gamma_i^{(t)}$ results in an equivalent model specification, an additional assumption is necessary to ensure identifiability. This can be accomplished by assuming

$$\gamma_i^{(0)} | \omega^2 \sim \mathrm{N}(0, \omega^2) \tag{4}$$

with a proper prior density assumed for ω^2, so that the $\gamma_i^{(t)}$ at $t = 0$ may be viewed as drawn from a common distribution centered at zero.

Assuming I objects and T time periods, the likelihood for this model can be written as

$$\begin{aligned} \mathrm{L}(\gamma, \boldsymbol{\sigma}^2, \theta, \tau^2, \omega^2 | \mathbf{y}, \mathbf{x}) = & \left(\prod_{i=1}^{I} \mathrm{N}(\gamma_i^{(0)} | 0, \omega^2) \right) \\ & \times \prod_{t=1}^{T} \left(\prod_{i<j} \prod_{k=1}^{K_{ij}^{(t)}} F(\gamma_i^{(t)} - \gamma_j^{(t)}, \theta, \mathbf{x}_i^{(t)}, \mathbf{x}_j^{(t)})^{y_{ijk}^{(t)}} \right. \\ & \left. \times (1 - F(\gamma_i^{(t)} - \gamma_j^{(t)}, \theta, \mathbf{x}_i^{(t)}, \mathbf{x}_j^{(t)}))^{1 - y_{ijk}^{(t)}} \right) \\ & \times \prod_{t=0}^{T-1} \prod_{i=1}^{I} \mathrm{N}(\gamma_i^{(t+1)} | \gamma_i^{(t)}, \sigma_i^{2^{(t)}}) \mathrm{N}(\log \sigma_i^{2^{(t+1)}} | \log \sigma_i^{2^{(t)}}, \tau^2) \end{aligned} \tag{5}$$

where γ and $\boldsymbol{\sigma}^2$ are the arrays of the $\gamma_i^{(t)}$ and $\sigma_i^{2^{(t)}}$, and $\mathrm{N}(\cdot | \cdot, \cdot)$ is a normal density of the first argument with the given mean and variance.

Together with equation (1), equations (2) and (3) form a state-space model with a stochastic variance for paired comparison data. This model can be viewed as an extension of the more usual constant variance model, where $\sigma_i^{2^{(t)}} = \sigma^2$ is assumed for all i and t, in which case equation (3) is no longer a component of the model. The constant variance model has been analyzed by Glickman (1993, 1999) and Fahrmeir & Tutz (1994). An important limitation of the constant variance model is that it does not account for the possibility of sudden shifts, innovations, or periods of uncertainty in the process for the $\gamma_i^{(t)}$. For example, in modeling the development of human expertise, one might expect bursts of cognitive development that would not be predicted by the constant variance model. In a marketing

context, the worth of a product may change suddenly relative to its competitors, and this change may be poorly described by the constant variance model.

The model for the time-varying variance has close connections with stochastic volatility models from finance (e.g. Jacquier *et al.*, 1994). Recent work in modeling of financial time series data has explored the applicability of stochastic volatility models in which the variance of a portfolio index or stock price is assumed to be changing according to equation (3). A major difference between our model and more conventional stochastic volatility models is that stochastic volatility assumes the variance of observations is changing, whereas in our model the variance of the process governing the merits is changing. Because the observations in our model are the results of preferences, and are therefore binomial, the variance of observations is determined from the mean. Thus, in the paired comparison situation, it would not be meaningful to assume a stochastic process on the observation variance except as it is translated through the process on the merit parameters. It does make sense in our situation to assume that the underlying process on the merits can undergo sudden shift, and this can be captured through a process assumed on the variance of the merit process.

Model fitting can be accomplished in a Bayesian framework through Markov chain Monte Carlo (MCMC) simulation from the posterior distribution. A choice of a prior distribution, convenient for model fitting, would assume a product of independent inverse-Gamma densities for τ^2 and ω^2 with low degrees of freedom to reflect initial uncertainty, and a non-informative density (e.g. normal with large variance, or uniform) on θ. Note that the $\gamma_i^{(t)}$ have distributions that are already specified conditionally. Assuming the functional form of F in equation (1) is tractable (e.g. the Bradley-Terry model, the Thurstone-Mosteller model, or various extensions), then, given τ^2, ω^2 and the $\sigma_i^{2(t)}$ for all i and t, the conditional posterior distribution of the remaining parameters has a form that is common to non-linear state-space models. Recognizing that the Bradley-Terry and the Thurstone-Mosteller models are particular examples of generalized linear models (Critchlow & Fligner, 1991), implementation of the Gibbs sampling steps to simulate the $\gamma_i^{(t)}$ and θ conditional on the $\sigma_i^{2(t)}$ can follow Zeger & Karim (1991), Karim & Zeger (1992), Glickman (1993) and Oh (1997). Sampling the $\sigma_i^{2(t)}$ conditional on the $\gamma_i^{(t)}$, τ^2 and ω^2, a necessary step in MCMC algorithms for fitting stochastic volatility models, is straightforward and can be carried out as in Jacquier *et al.* (1994). When the number of time periods is large, block-sampling strategies using the Metropolis-Hastings algorithm can be employed as described in Shephard & Pitt (1997). The conditional distributions of τ^2 and ω^2 given the remaining parameters are inverse-Gamma, so that sampling for this step can be performed directly.

3 Example: NFL football game outcomes

The model of Section 2 can be applied to a data set consisting of NFL football game outcomes from regular season competition from 1996 to 2000. For the football data, the merit (or strength) $\gamma_i^{(t)}$ of a team can be inferred through game outcomes, which are the results of paired comparisons. The results of NFL football games are individual scores, and the team that wins is the one with the higher score. Specifically, the outcome of the kth comparison between teams i and j at time t, $Y_{ijk}^{(t)}$, is 1 if player i defeats j, and 0 if player j defeats i. Very rarely does the outcome of a football game result in a tie. Only two ties occurred during regular season games between 1996 and 2000, both of which took place in 1997. These

two games were removed from the analysis. Dynamic models for NFL football games have been examined by Glickman & Stern (1998), Harville (1980) and Sallas & Harville (1988), treating the score difference as the outcome of interest. In this discussion, the binary indicator of the winning team is the outcome variable.

The data we examine consist of all decisive game outcomes from 1996 to 1999, and the first 10 weeks of game outcomes from 2000, resulting in a total of 1109 games played among 31 teams. One of the teams, the Cleveland Browns, reformed in 1999, so that only 30 teams were included in our data set for 1996–98. For our analysis, game outcomes are grouped into periods of 1 year. We therefore assume that team strengths remain constant within a regular season, but may vary between seasons. During a season, each team plays 16 games. Because teams competing on their home field are understood to have an advantage (e.g. see Glickman & Stern, 1998), the effect of home field is modeled through an order effect following Davidson & Beaver (1977). In game k played during season t between teams i and j, let

$$x_{ijk}^{(t)} = \begin{cases} 1 \text{ if team } i \text{ plays on its home field} \\ -1 \text{ if team } j \text{ plays on its home field} \end{cases}$$

The model for game outcomes is given by

$$\mathrm{P}(Y_{ijk}^{(t)} = 1) = \frac{\exp(\gamma_i^{(t)} + x_{ijk}^{(t)}\beta)}{\exp(\gamma_i^{(t)} + x_{ijk}^{(t)}\beta) + \exp(\gamma_j^{(t)})} \tag{6}$$

where β is the effect of playing on the home field.

For NFL football games, a tendency exists for team strengths to regress to the mean over time. This happens because strong teams have good players who age and therefore become slightly worse over time. Furthermore, poor teams obtain better chances at selecting strong players during the draft lottery, and therefore tend to improve. Letting ρ denote an autoregression parameter, we assume

$$\gamma_i^{(t+1)} | \gamma_i^{(t)}, \sigma_i^{2^{(t+1)}}, \rho \sim \mathrm{N}(\rho\gamma_i^{(t)}, \sigma_i^{2^{(t+1)}}) \tag{7}$$

so that merit parameters move on average towards zero, assuming the magnitude of ρ is inferred to be less than unity. As before, the model for equation (3) describes the change in $\sigma_i^{2^{(t)}}$ over time. A vague but proper prior is assumed for all model parameters. The reciprocal of the variances is modeled following a Gamma distribution with mean 1 and variance 10.

In addition to fitting the stochastic variance model of equation (3), we also fit a constant variance model assuming

$$\gamma_i^{(t+1)} | \gamma_i^{(t)}, \sigma^2, \rho \sim \mathrm{N}(\rho\gamma_i^{(t)}, \sigma^2) \tag{8}$$

where σ^2 is a single variance governing the change in $\gamma_i^{(t)}$ over time. A vague but proper prior distribution is assumed for σ^2 in our analysis.

Both models were fit by MCMC simulation, with burn-in periods of 20 000 iterations, at which point the model was diagnosed to have reached stationarity through trace plots and informal diagnostics (e.g. Geweke, 1992). Model summaries were computed based on the empirical distribution of simulated parameter values for every 100th draw for 2000 000 iterations beyond the 20 000th iteration, resulting in 2000 draws per parameter. Only every 100th simulated value was saved to conserve disk space, and to reduce the effect of autocorrelation of successive parameter draws. Because the size of the problem is relatively small (31 teams with

TABLE 1. Posterior summaries for regular NFL season 2000 team strengths*

	Constant variance		Stochastic variance	
Team	Posterior mean	Posterior SD	Posterior mean	Posterior SD
Tennessee Titans	1.127	0.5487	1.529	0.8355
Minnesota Vikings	0.824	0.5287	1.072	0.7036
Miami Dolphins	0.7841	0.5199	0.8892	0.6567
Oakland Raiders	0.7585	0.5394	0.988	0.777
Indianapolis Colts	0.581	0.5112	0.7729	0.6636
New York Jets	0.5222	0.5176	0.6793	0.6467
Saint Louis Rams	0.5105	0.5171	0.5209	0.6716
Buffalo Bills	0.4992	0.5113	0.5639	0.5895
New York Giants	0.4724	0.5128	0.5952	0.6581
Tampa Bay Buccaneers	0.3795	0.5132	0.3896	0.5633
Washington Redskins	0.3228	0.4892	0.3448	0.5422
Kansas City Chiefs	0.3093	0.5223	0.3431	0.5929
Detroit Lions	0.2102	0.4853	0.2991	0.563
Baltimore Ravens	0.1549	0.4894	0.1766	0.5888
Pittsburgh Steelers	−0.01381	0.521	0.001656	0.6053
Jacksonville Jaguars	−0.02185	0.5374	−0.3095	0.6729
Philadelphia Eagles	−0.04988	0.5073	0.04097	0.5818
Green Bay Packers	−0.05577	0.5134	−0.1259	0.5863
New Orleans Saints	−0.1552	0.5092	−0.1256	0.6641
Denver Broncos	−0.1787	0.5208	−0.3635	0.6442
Dallas Cowboys	−0.1959	0.5072	−0.3262	0.5812
New England Patriots	−0.3066	0.5081	−0.3637	0.6058
Carolina Panthers	−0.3517	0.5059	−0.5142	0.5987
Seattle Seahawks	−0.3547	0.5086	−0.4364	0.5791
Atlanta Falcons	−0.601	0.517	−0.9391	0.6883
Chicago Bears	−0.6011	0.507	−0.6242	0.6144
San Diego Chargers	−0.7667	0.5215	−1.127	0.7829
Arizona Cardinals	−0.8509	0.5322	−1.189	0.7225
San Francisco 49ers	−0.8653	0.5199	−1.382	0.7459
Cincinnati Bengals	−0.9366	0.5491	−1.232	0.7378
Cleveland Browns	−1.129	0.5391	−1.625	0.7722

*The first two columns display model summaries for the constant variance dynamic model, and the second two columns display summaries for the stochastic variance dynamic model.

five time-varying strength parameters), autocorrelation of the MCMC chain was not a problem.

Posterior model summaries for the strength parameters in the 2000 season for each model are displayed in Table 1. The table ranks teams according to the posterior means under the constant variance model. Posterior means are comparable between models, with slightly greater spread of means in the stochastic variance model. While most of the teams rank similarly in both models according to the posterior means, the stochastic variance model infers that the Jacksonville Jaguars were worse in 2000 than in the constant variance model. This inference is made because the Jaguars in 2000 were having an unusually poor season relative to previous years, and the stochastic variance model downweighted the impact of previous seasons' game outcomes in determining the effect on the 2000 season.

It is noteworthy that the posterior uncertainty of the strength parameters, under the constant variance model, is roughly constant, with slightly greater posterior standard deviations for teams that are relatively strong and relatively weak. In comparison, the posterior standard deviations under the stochastic variance model

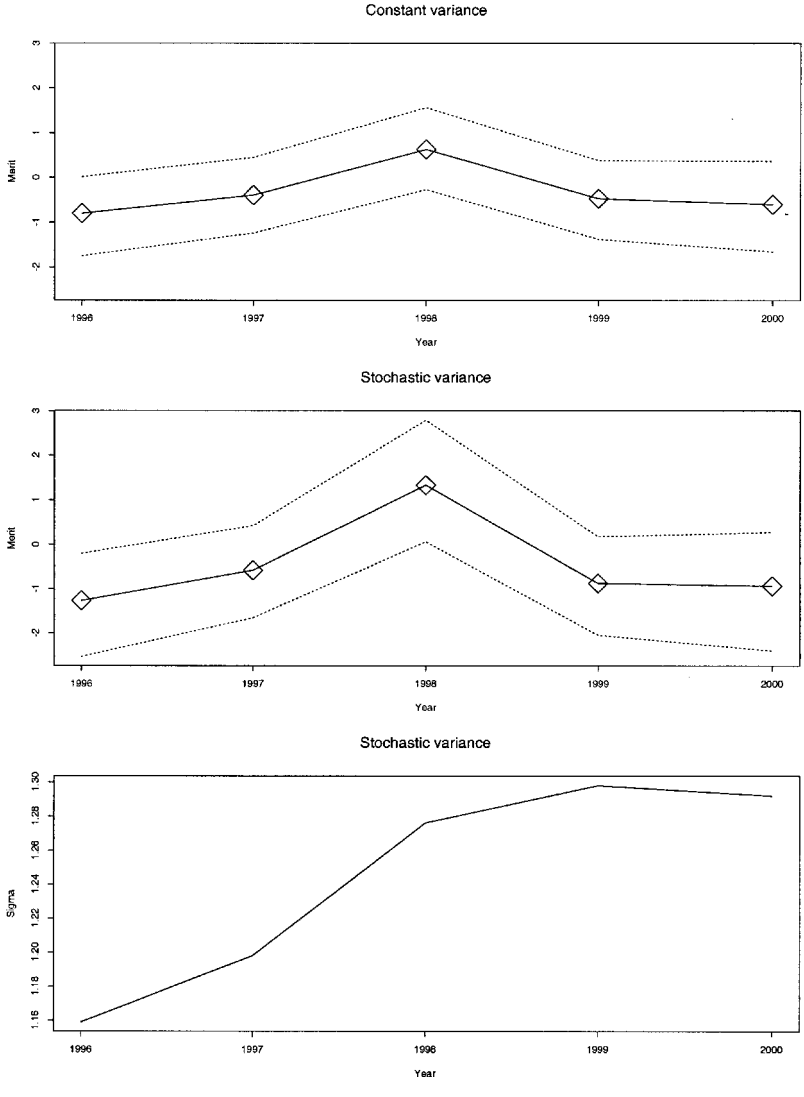

FIG. 1. Model summaries for the Atlanta Falcons. Top: posterior means for the $\gamma_i^{(t)}$ in the constant variance model, with pointwise approximate 95% central posterior intervals. Middle: posterior means for the $\gamma_i^{(t)}$ in the stochastic variance model, with pointwise approximate 95% central posterior intervals. Bottom: posterior means for the $\sigma_i^{2(t)}$.

are larger, and vary more in magnitude. This variation is attributable to somewhat larger inferred changes in strength over time, which results in inferred values for $\sigma_i^{2(t)}$ typically larger than average. For example, the Jaguars' posterior standard deviation under the stochastic variance model is large compared with other teams' posterior standard deviations, reflecting the sudden decline in strength in 2000. Further evidence of the effect of sudden shifts in performance can be seen in Figs 1 and 2. Figure 1 shows model summaries for the Atlanta Falcons, who were a below-average team through most of the 1990s, and then suddenly had a strong performance in 1998 (they played in the Superbowl that season, but lost). The plots indicate that the stochastic variance model allowed a greater change in the

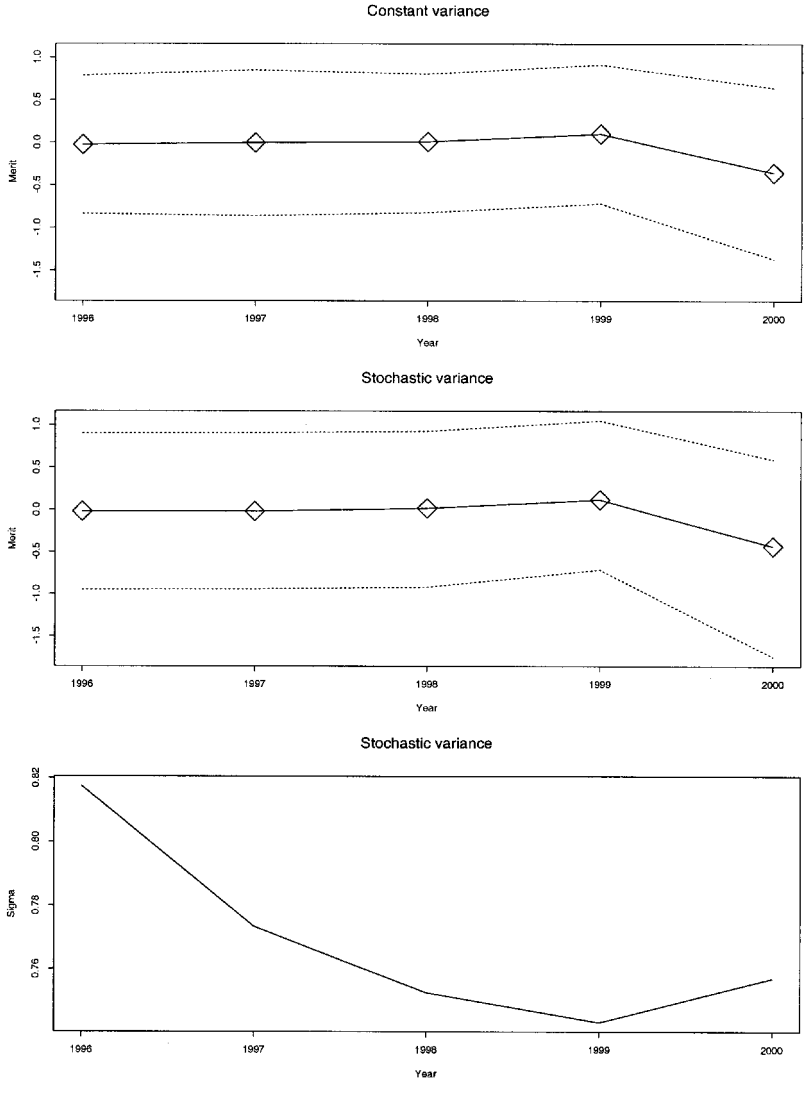

FIG. 2. Model summaries for the Seattle Seahawks. Top: posterior means for the $\gamma_i^{(t)}$ in the constant variance model, with pointwise approximate 95% central posterior intervals. Middle: posterior means for the $\gamma_i^{(t)}$ in the stochastic variance model, with pointwise approximate 95% central posterior intervals. Bottom: posterior means for the $\sigma_i^{2(t)}$.

strength parameter for the Falcons in 1998 compared with the constant variance model. The pointwise posterior standard deviations are also larger than in the constant variance model, reflecting the extra uncertainty in the $\gamma_i^{(t)}$ due to the sudden change in performance. In contrast, Fig. 2, which shows analogous information for the Seattle Seahawks, demonstrates that the stochastic variance model behaves similarly to the constant variance model when a team's performances remain stable. The posterior mean of the $\gamma_i^{(t)}$ and the pointwise posterior standard deviations are similar for both models. Compared to the stochastic variance model summaries for the Falcons, the posterior means of the $\sigma_i^{2(t)}$ for the Seahawks are much lower, which is consistent with the lack of substantial changes in the $\gamma_i^{(t)}$.

Inference for the autoregression parameter, ρ, reveals that substantial regression to the mean occurred over the five seasons. For the constant variance model, a Monte Carlo 95% central posterior interval is $(0.294, 0.803)$. The corresponding interval for the stochastic variance model is $(0.217, 0.752)$. The large posterior variability in ρ indicates a lack of substantial information in the data about this parameter in both models. In either case, the data provide evidence that team strength does shrink, on average, towards the mean over time. Inferences for the home field effect parameter, β, are comparable for both models. For the constant variance model, a 95% central posterior interval for β is $(0.1514, 0.4052)$, and for the stochastic variance model $(0.1602, 0.4399)$. The positive value confirms that an advantage exists for playing on the home field.

4 Analysis for comparing many objects

When many objects are being compared over time, an exact likelihood-based approach (e.g. maximum likelihood, Bayesian analysis) may become computationally intractable. For example, when rating populations of chess players, or competitors in online gaming systems which can attract tens of thousands of players, a likelihood-based analysis would not be possible to perform in real time. Instead, a simple forward filtering algorithm may be preferable. Glickman (1999) develops an approximate Bayesian analysis for the constant variance model in which the $\gamma_i^{(t)}$ are updated sequentially with the acquisition of new data. The result of this analysis is an approximation to the marginal posterior distribution for the most recent merit parameter for any object. Rather than perform an analysis that jointly infers all parameters simultaneously, the approach taken in Glickman (1999) involves updating the merit parameter of an object by integrating out information about other objects through the prior distribution rather than the posterior distribution. While this approach results in a loss of efficiency, there are computational advantages that allow for the derivation of a simple algorithm. The procedure in Glickman (1999) can be extended to the stochastic variance model.

The algorithm to update objects' merit parameters proceeds in the following manner. It is assumed in the algorithm that τ^2 (the variance of the change in $\log \sigma_i^{2(t)}$ over time) and ω^2 (the variance of the $\gamma_i^{(0)}$) have been estimated and fixed in advance. This can be accomplished by fitting the model using an exact likelihood procedure to a set of data of manageable size, and determining the values of τ^2 and ω^2 that maximize the marginal posterior distribution of these two parameters. For the remainder of the development, these parameters are assumed fixed.

(1) At the end of the time period $t-1$, each object's merit parameter, $\gamma_i^{(t-1)}$ has an approximating normal marginal posterior distribution with known mean $\mu_i^{(t-1)}$ and variance $\phi_i^{2(t-1)}$. Also, each object has a current (and known) variance parameter, $\sigma_i^{2(t-1)}$, describing the variability of the time change in merit for that object.
(2) Collect all comparisons during time period t (time periods are assumed to be equally spaced).
(3) For each object individually, perform appropriate calculations (described below) to determine the updating of $\sigma_i^{2(t-1)}$ to $\sigma_i^{2(t)}$, and then the updating of $\mu_i^{(t-1)}$ to $\mu_i^{(t)}$ and $\phi_i^{2(t-1)}$ to $\phi_i^{2(t)}$. These are the new parameters for the distribution of $\gamma_i^{(t)}$.

These calculations are repeated for each time period as data are observed.

To understand the calculations that result in an estimate of $\sigma_i^{2(t)}$, suppose an object in question has a merit $\gamma^{(t-1)}$ that can be summarized at time $t-1$ by the marginal posterior distribution

$$\gamma^{(t-1)} \sim \mathrm{N}(\mu^{(t-1)}, \phi^{2(t-1)}) \tag{9}$$

Assume that, during period t, this object is compared to others indexed by $j = 1, \ldots, m$, with marginal posterior distributions

$$\gamma_j^{(t-1)} \sim \mathrm{N}(\mu_j^{(t-1)}, \phi_j^{2(t-1)}) \tag{10}$$

The stochastic variance model assumes

$$\gamma^{(t)} \sim \mathrm{N}(\gamma^{(t-1)}, \sigma^{2(t)}) \tag{11}$$

$$\log \sigma^{2(t)} \sim \mathrm{N}(\log \sigma^{2(t-1)}, \tau^2) \tag{12}$$

Further, let $\hat{\gamma}^{(t)}$ be the maximum value of $\gamma^{(t)}$ in the likelihood for period t, integrated over the *prior* distribution of the other objects' $\gamma_j^{(t)}$, and let v^2 be the associated asymptotic variance of $\hat{\gamma}^{(t)}$ from the marginalized likelihood. When integrating over the distribution of the other objects' merits, we assume $\log \sigma_j^{2(t)} = \log \sigma_j^{2(t-1)}$ for object j, even though this is only true in expectation. Both $\hat{\gamma}^{(t)}$ and v^2 can be computed using iterative numerical procedures (though an approximation is used in the algorithm that follows). The distribution of $\hat{\gamma}^{(t)}$ can be approximated by

$$\hat{\gamma}^{(t)} \sim \mathrm{N}(\gamma^{(t)}, v^2) \tag{13}$$

Combining equations (9), (11) and (13), integrating out $\gamma^{(t-1)}$ and $\gamma^{(t)}$, and for notational convenience letting $\alpha^{(t-1)} = \log \sigma^{2(t-1)}$ and $\alpha^{(t)} = \log \sigma^{2(t)}$, we have

$$\hat{\gamma}^{(t)} \sim \mathrm{N}(\mu^{(t-1)}, \phi^{2(t-1)} + \exp(\alpha^{(t)}) + v^2) \tag{14}$$

From equation (12), we have

$$\alpha^{(t)} \sim \mathrm{N}(\alpha^{(t-1)}, \tau^2) \tag{15}$$

Noting that all other parameters are known, the approximate marginal posterior density of $\alpha^{(t)}$ is the product of the densities in equations (14) and (15), so that the marginal log-posterior, up to an additive constant, is given by

$$\begin{aligned} \log p(\alpha^{(t)} | \mathbf{y}^{(t)}) = & -\frac{1}{2} \frac{(\alpha^{(t)} - \alpha^{(t-1)})^2}{\tau^2} \\ & -\frac{1}{2} \log(\phi^{2(t-1)} + \exp(\alpha^{(t)}) + v^2) - \frac{1}{2} \frac{(\hat{\gamma}^{(t)} - \mu^{(t-1)})^2}{\phi^{2(t-1)} + \exp(\alpha^{(t)}) + v^2} \end{aligned} \tag{16}$$

where $\mathbf{y}^{(t)}$ denotes the collection of comparison outcomes during time period t. Rather than compute $\hat{\gamma}^{(t)}$ and v^2 numerically, we use approximations derived in Glickman (1999). In particular, we approximate $(\hat{\gamma}^{(t)} - \mu^{(t-1)})$ by a Taylor series expansion through the linear term, and approximate v^2 by the curvature around $\mu^{(t-1)}$ rather than $\hat{\gamma}^{(t)}$. This yields

$$\hat{\gamma}^{(t)} - \mu^{(t-1)} \approx v^2 \sum_{j=1}^{m} \sum_{k=1}^{n_j} g(\phi_j^{2(t-1)}) \{ y_{jk}^{(t)} - \mathrm{E}(y | \mu^{(t-1)}, \mu_j^{(t-1)}, \phi_j^{2(t-1)}) \} \tag{17}$$

with

$$v^2 \approx \left[\sum_{j=1}^{m} n_j g(\phi_j^{2(t-1)})^2 \mathrm{E}(y | \mu^{(t-1)}, \mu_j^{(t-1)}, \phi_j^{2(t-1)}) \{ 1 - \mathrm{E}(y | \mu^{(t-1)}, \mu_j^{(t-1)}, \phi_j^{2(t-1)}) \} \right]^{-1} \tag{18}$$

where $y_{jk}^{(t)}$ is the result of the kth comparison of the object with object j during period t, n_j is the number of times the object is compared to object j and

$$g(\phi^2) = \frac{1}{\sqrt{1 + 3\phi^2/\pi^2}}$$

$$\mathrm{E}(y \mid \mu, \mu_j, \phi_j^2) = \frac{1}{1 + \exp(-g(\phi_j^2)\ (\mu - \mu_j))}$$

The algorithm proceeds by estimating $\alpha^{(t)}$ (and therefore $\sigma^{2(t)}$) by maximizing over equation (16). This can be accomplished using a numerical algorithm. For example, because the first term in (16) is maximized by $\alpha^{(t)} = \alpha^{(t-1)}$, and the second two terms in (16) are maximized by setting

$$(\hat{\gamma}^{(t)} - \mu^{(t-1)})^2 = \phi^{2(t-1)} + \exp(\alpha^{(t)}) + v^2$$

so that $\alpha^{(t)} = \log((\hat{\gamma}^{(t)} - \mu^{(t-1)})^2 - \phi^{2(t-1)})$ assuming

$$(\hat{\gamma}^{(t)} - \mu^{(t-1)})^2 - \phi^{2(t-1)} - v^2 > 0 \tag{19}$$

the Newton-Raphson algorithm will converge to the maximum quickly if the initial value of $\alpha^{(t)}$ is selected to be between $\alpha^{(t-1)}$ and $\log((\hat{\gamma}^{(t)} - \mu^{(t-1)})^2 - \phi^{2(t-1)})$. If equation (19) is false, then the second two terms reach their supremum when $\exp(\alpha^{(t)})$ is set to zero. In this situation, because these latter two terms are bounded above as $\alpha^{(t)} \to -\infty$, the first term dominates, and convergence of the Newton-Raphson algorithm is quick when choosing an initial value of $\alpha^{(t)}$ less than $\alpha^{(t-1)}$.

Once $\sigma^{2(t)}$ is estimated, we set $\phi'^{2(t)} = \phi^{2(t-1)} + \sigma^{2(t)}$, which is the prior variance for $\gamma^{(t)}$ accounting for the passage of time from period $t-1$ to t. Now the algorithm of Glickman (1999) may be applied directly to obtain $\mu^{(t)}$ and $\phi^{2(t)}$, the marginal posterior mean and variance of $\gamma^{(t)}$, the merit parameter for the object in question. These are given by

$$\mu^{(t)} = \mu^{(t-1)} + \frac{1}{1/\phi'^{2(t)} + 1/v^2} \sum_{j=1}^{m} \sum_{k=1}^{n_j} g(\phi_j^{2(t-1)})\ \{y_{jk}^{(t)} - \mathrm{E}(y \mid \mu^{(t-1)}, \mu_j^{t-1}, \phi_j^{2(t-1)})\} \tag{20}$$

$$\phi^{2(t)} = \left(\frac{1}{\phi'^{2(t)}} + \frac{1}{v^2}\right)^{-1} \tag{21}$$

Details of the derivations appear in Glickman (1999).

To assess the accuracy of the approximation algorithm, simulated data under varying parameter values were generated, and nominal coverage was compared with the results of simulations. We performed a total of 32 simulation sets, which are summarized in Table 2. We considered two values (0, 0.5) for the prior mean, μ, four different numbers of comparisons, m, (10, 50, 200, 1000) and four different values of the standard deviation for the change in $\log \sigma^{2(t)}$, τ (0, 0.3, 0.7, 1.2). These values were chosen to span a plausible range of values that might be expected in practice. For each fixed combination of μ, m and τ, data were generated in the following manner: values of $\phi^{(t-1)}$ and $\sigma^{(t-1)}$ were fixed at 0.173 and 0.0576, respectively, so that a 95% prior interval around $\gamma^{(t-1)}$ had length 2/3, and that the standard deviation of the $\gamma^{(t-1)}$ was three times larger than the standard deviation in the change of the $\gamma^{(t-1)}$ over time. A value of $\sigma^{2(t)}$ was simulated conditional on $\sigma^{2(t-1)}$ and τ, and then a value of $\gamma^{(t)}$ was simulated given $\mu^{(t-1)}$, $\phi^{(t-1)}$ and $\sigma^{2(t)}$. For

TABLE 2. Results of the approximating algorithm on simulated data*

Prior mean	Number of comparisons	$\tau = 0$	$\tau = 0.3$	$\tau = 0.7$	$\tau = 1.2$
$\mu = 0$	$m = 10$	(0.5210, 0.9460)	(0.4928, 0.9476)	(0.4892, 0.9450)	(0.4886, 0.9414)
	$m = 50$	(0.4942, 0.9498)	(0.5026, 0.9562)	(0.5042, 0.9504)	(0.4794, 0.9414)
	$m = 200$	(0.5074, 0.9500)	(0.4890, 0.9514)	(0.4822, 0.9450)	(0.4840, 0.9466)
	$m = 1000$	(0.4896, 0.9484)	(0.4922, 0.9482)	(0.4894, 0.9402)	(0.4908, 0.9448)
$\mu = 0.5$	$m = 10$	(0.5048, 0.9524)	(0.4912, 0.9504)	(0.4968, 0.9440)	(0.4932, 0.9384)
	$m = 50$	(0.5082, 0.9486)	(0.5044, 0.9460)	(0.5060, 0.9484)	(0.4924, 0.9398)
	$m = 200$	(0.4984, 0.9466)	(0.5034, 0.9540)	(0.4924, 0.9472)	(0.4930, 0.9436)
	$m = 1000$	(0.4988, 0.9498)	(0.4918, 0.9472)	(0.4914, 0.9446)	(0.4868, 0.9460)

*For each analysis, 5000 replications were simulated and nominal 50 and 95% central posterior intervals were constructed. The pairs of values in the parentheses for each analysis consist of the proportion of 5000 replications in which the simulated value γ was contained in the nominal 50 and 95% intervals, respectively.

the m other objects involved in the comparisons, the collection of $\mu_j^{(t-1)}$ was simulated from a normal distribution with mean zero and standard deviation 0.173. The $\phi_j^{(t-1)}$ and $\sigma_j^{(t)}$ were generated from scaled χ^2 distributions on 20 degrees of freedom with means of 0.173 and 0.0576, respectively. Values of $\gamma_j^{(t)}$ were generated by the same process as $\gamma^{(t)}$. Finally, the outcome of comparisons were generated from the Bradley-Terry model given the generated values of the $\gamma^{(t)}$ and the $\gamma_j^{(t)}$. The algorithm was then applied (ignoring parameter values at time t) to determine the parameters $\mu^{(t)}$ and $\phi^{(t)}$ of the approximating normal posterior to $\gamma^{(t)}$. Approximate 50 and 95% normal central posterior intervals for $\gamma^{(t)}$ were calculated as $\mu^{(t)} \pm z\phi^{(t)}$ with $z = 0.6745$ or 1.96. It was noted whether the generated value of $\gamma^{(t)}$ was contained in this interval. This process was repeated 5000 times, and the fraction of times in which the true parameter value was contained in the intervals is summarized in Table 2.

The results of the simulations demonstrate that the algorithm produces close to nominal coverage under varying conditions. Table 2 reveals that the nominal 50 and 95% posterior intervals contain roughly 0.5 and 0.95 of the generating value of $\gamma^{(t)}$. The accuracy of the approximation algorithm does not seem to change by varying μ. However, there appears to be a small loss of efficiency when m becomes larger and when τ is at its largest. In these cases, the actual coverage is consistently smaller than nominal coverage, indicating that the posterior intervals are not wide enough. The discrepancy does not seem large enough to be of great practical concern.

5 Example: best chess players of all time

The filtering algorithm of Section 4 can be applied to a data set consisting of all known results of chess games from 1857 to 1991 played among 88 of the world's all-time best chess players. For chess data, the merit (or strength), $\gamma_i^{(t)}$, of a player can be inferred through game outcomes, which are the results of paired comparisons. The outcome of the kth comparison between competitors i and j at time t, $Y_{ijk}^{(t)}$, is 1 if player i defeats j and 0 if player j defeats i. The data set, which consists of 15 664 outcomes of games played among 1367 pairs of players, was compiled by Prof. Nathan Divinsky. Not all $\binom{88}{2} = 3828$ pairs of players competed against

each other due to non-overlapping chess careers. A detailed account of the data appears in Keene & Divinsky (1989). Several models of chess playing strength have been fit by these data, including the models of Elo (1978), Joe (1990), Henery (1992) and Glickman (1999).

For our analysis, we group game outcomes into periods of 1 year. We therefore act as if all of the games were played simultaneously at the beginning of each year, with innovations in merit and in variance changing over the remainder of the year. The data consist of 135 periods, though some years (e.g. 1859, 1874 and 1875) contain no game outcomes. For years in which no games were recorded, there is no likelihood contribution from data in equation (5), though the terms for the change in $\gamma_i^{(t)}$ and $\sigma_i^{2(t)}$ still appear.

Unlike the football game data, one aspect of chess outcome data for which the model must account is the existence of ties. Not only does this third possible paired comparison outcome occur in chess, but it occurs frequently. Several extensions to common paired comparisons models have addressed ties as a third outcome, including the extensions by Davidson (1970) and Rao & Kupper (1967) to the Bradley-Terry model, and by Greenberg (1965) for the Thurstone-Mosteller model. Instead of treating a tie as a third outcome to the model, we adopt an approach which acts as if ties are not really observed, but that they are viewed as half the contribution of a win and a loss. In other words, we assume two ties contain the same information about players' strengths as a win followed by a loss (or vice versa). Thus, if p_{ij} is the probability that i defeats j, the contribution to the likelihood of a tie would be $\sqrt{p_{ij}(1-p_{ij})}$. This approach to ties in paired comparisons can also be found in Glickman (1999).

We carried out the filtering algorithm described in Section 4 on these data, and also carried out the filtering algorithm in Glickman (1999) as a comparison. Pilot MCMC algorithms were run to estimate initial parameter values by their approximate posterior means. For the filtering algorithm described in Section 4, τ was set to 0.66, the initial σ for each player was set to 0.05, and the prior distribution of each player's strength, γ, was assumed Gaussian with mean and standard deviation of 0 and 0.15, respectively. For the constant variance filtering algorithm, σ was set at 0.01054, and prior distribution for each player's strength, γ, was assumed Gaussian with mean and standard deviation of 0 and 0.2027, respectively.

The two models result in comparable inferences. In general, the filtered estimates of the $\gamma_i^{(t)}$ have large overlap across the two models. The typical trend for an individual's $\gamma_i^{(t)}$ over time is low early in the player's career, a peak in the middle and then a slow decline towards the end. This finding is consistent with previous analyses of this data set (Joe, 1990; Glickman, 1999).

For the constant variance filtering results, a typical pattern is that the uncertainty about the $\gamma_i^{(t)}$ tends to decrease and stabilize as the player's career progresses. The stochastic variance filtering algorithm keeps the estimated posterior standard deviation of the $\gamma_i^{(t)}$ roughly constant. Also, the changes in the mean $\gamma_i^{(t)}$ tend to be smoother in the constant variance model, and the corresponding changes in the stochastic variance model are often more jagged.

Inferences about two players' careers help to illustrate the differences between the two algorithms. In Fig. 3, filtering results about Max Euwe are displayed, and in Fig. 4 results about Robert Fischer are shown. Both players were world champions for short periods (Euwe in 1935-37 and Fischer 1972-74). Euwe continued playing long after his world championship reign, while Fischer quit

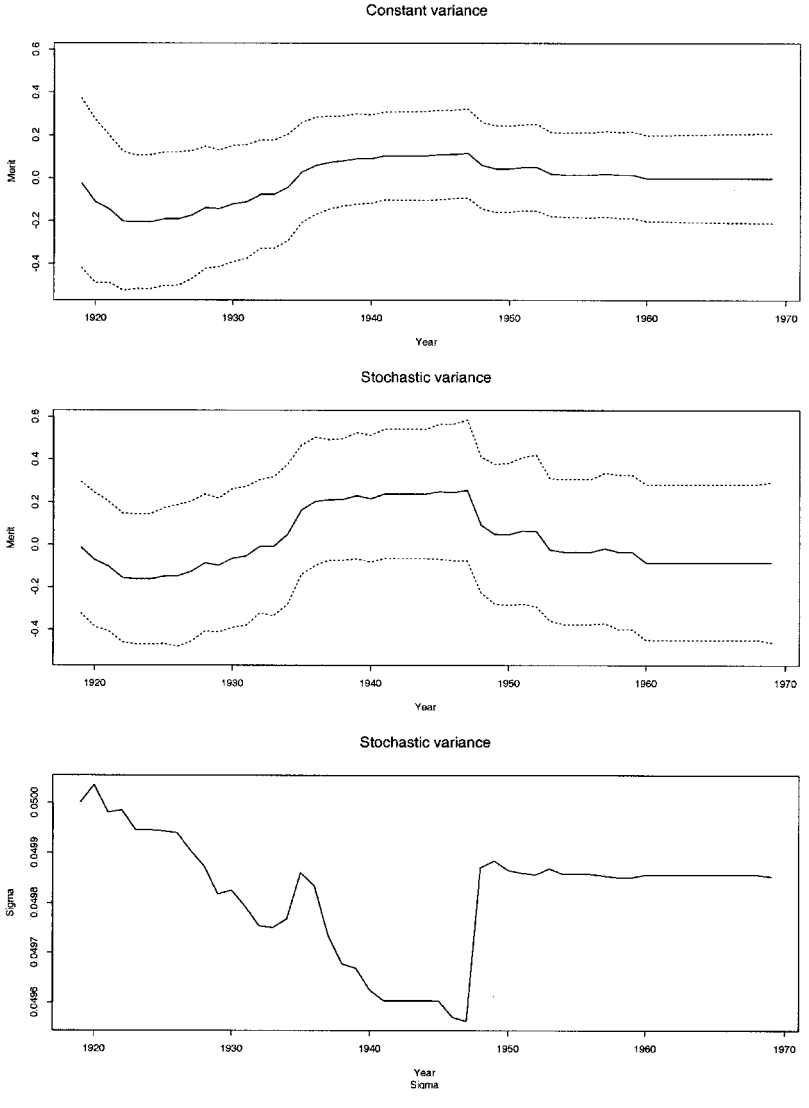

FIG. 3. Model summaries for Max Euwe. Top: posterior means for the $\gamma_i^{(t)}$ in the constant variance filtering procedure, with pointwise approximate 95% central posterior intervals. Middle: posterior means for the $\gamma_i^{(t)}$ in the stochastic variance filtering procedure, with pointwise approximate 95% central posterior intervals. Bottom: estimates for the $\sigma_i^{2(t)}$.

professional chess arguably at the peak of his career. The plots show that the estimates from the stochastic variance filtering algorithm change more abruptly than in the constant variance model. This is reflected in changes in the $\sigma_i^{(t)}$ in the stochastic variance algorithm. When Euwe had poor results in the late 1940s, the stochastic variance algorithm had $\sigma_i^{(t)}$ experience a corresponding increase to reflect the gain in uncertainty in Euwe's true strength. Similarly, Fischer's phenomenal results in the world championship cycle in the early 1970s is shown in Fig. 4 by an increase in the mean strength, but also in an increase in the value of $\sigma_i^{(t)}$.

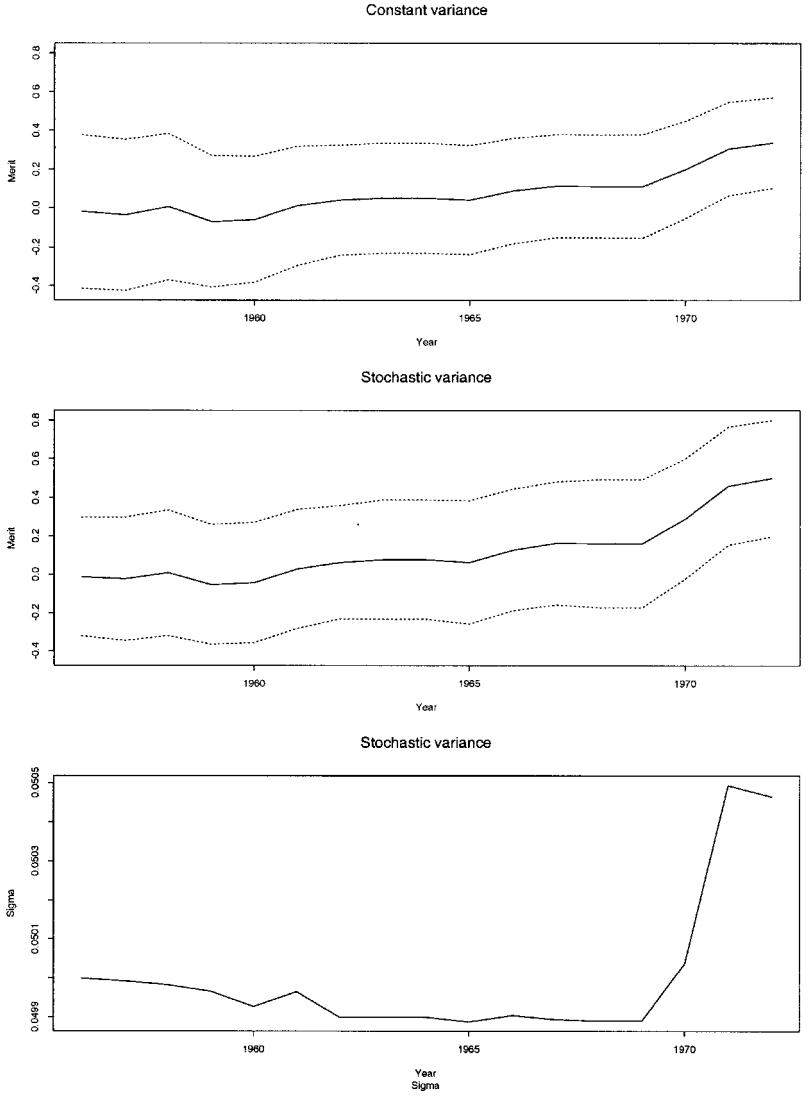

FIG. 4. Model summaries for Robert Fischer. Top: posterior means for the $\gamma_i^{(t)}$ in the constant variance filtering procedure, with pointwise approximate 95% central posterior intervals. Middle: posterior means for the $\gamma_i^{(t)}$ in the stochastic variance filtering procedure, with pointwise approximate 95% central posterior intervals. Bottom: estimates for the $\sigma_i^{2(t)}$.

6 Discussion

The dynamic paired comparison model presented in this paper extends previous work by allowing the variance of the state process to change stochastically. Considering such models increases the flexibility to describe phenomena where the underlying characteristics may undergo sudden shifts or changes in paradigm. Fitting the stochastic variance model for paired comparison data can be carried out using standard Bayesian computational machinery. For situations where many objects are being compared over time, in which case a full likelihood analysis may be too computationally intensive, this paper demonstrates an approximating algorithm that can be carried out with far less of a computational burden.

Aside from the increased flexibility, the stochastic variance model has an important benefit over the constant variance model when, for example, it is suspected that interventions may affect the process for the merits. Under the constant variance model, the variance in an object's merit prior to observing data at time t must decrease after the data are observed. This is a direct consequence of equation (21), where the posterior variance must be less than the prior variance. However, under the stochastic variance model, the variance may *increase* after data are observed. This reflects the increased uncertainty about the merit parameter after observing unusual data.

A variety of extensions can be incorporated into the stochastic variance model. One extension is to incorporate covariate information in the change in $\sigma_i^{2(t)}$ over time. For example, in the context of human cognitive development, older people may stabilize in merit so that the change in $\sigma_i^{2(t)}$ may be assumed to be negatively related with age. Another extension involves using a different distribution other than normal for describing the change in $\gamma_i^{(t)}$, and log-normal for describing the change in $\sigma_i^{2(t)}$ over time. For example, in comparing certain types of financial products over time, it may be more reasonable to assume models describing a greater probability for small increases in $\gamma_i^{(t)}$ but occasional large decreases with small probability. In each case, standard Bayesian tools can still be invoked to fit such model extensions.

REFERENCES

Bradley, R. A. (1984) Paired comparisons: some basic procedures and examples. In: P. R. Kirshnajah & P. K. Sen (Eds) *Handbook of Statistics* 4, pp. 299-326 (Amsterdam, Elsevier).

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

Capobianco, E. (1996) State-space stochastic volatility models: a review of estimation algorithms, *Applied Stochastic Models Data Analysis*, 12, pp. 265-279.

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

David, H. A. (1988) *The Method of Paired Comparisons*, 2nd edn (London, Chapman and Hall).

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

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

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

Fahrmeir, L. & Tutz, G. (1994) Dynamic stochastic models for time-dependent ordered paired comparison systems, *Journal of the American Statistical Association*, 89, pp. 1438-1449.

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith (Eds) *Bayesian Statistics 4* (Oxford, Clarendon Press).

Glickman, M. E. (1993) Paired comparison models with time-varying parameters, PhD Dissertation, Harvard University Department of Statistics, Cambridge, USA.

Glickman, M. E. (1999) Parameter estimation in large dynamic paired comparison experiments, *Applied Statistics*, 48, pp. 377-394.

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

Greenberg, M. G. (1965) A modification of Thurstone's law of comparative judgment to accommodate a judgment category of "equal" or "no difference", *Psychology Bulletin*, 64, pp. 108-112.

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

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

JACQUIER, E., POLSON, N. & ROSSI, P. (1994) Bayesian analysis of stochastic volatility models (with discussion), *Journal of Business Economics and Statistics*, 12, pp. 371-389.

JOE, H. (1990) Extended use of paired comparison models, with application to chess rankings, *Applied Statistics*, 39, pp. 85-93.

KARIM, M. & ZEGER, S. L. (1992) Generalized linear models with random effects; Salamander mating revisited, *Biometrics*, 48, pp. 631-644.

KEENE, R. & DIVINSKY, N. (1989) *Warriors of the Mind: A Quest for the Supreme Genius of the Chess Board* (Brighton, Hardinge Simpole).

KNORR-HELD, L. (2000) Dynamic rating of sports teams, *The Statistician* (in press).

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

OH, M. (1997) A Gibbs sampling approach to Bayesian analysis of generalized linear models for binary data, *Computational Statistics*, 12, pp. 431-445.

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

SALLAS, W. M. & HARVILLE, D. A. (1988) Noninformative priors and restricted maximum likelihood estimation in the Kalman filter. In: J. C. SPALL (Ed.) *Bayesian Analysis of Time Series and Dynamic Models*, pp. 477-508 (New York, Marcel Dekker).

SHEPHARD, N. & PITT, M. K. (1997) Likelihood analysis of non-Gaussian measurement time series, *Biometrika*, 84, pp. 653-667.

SIMONTON, D. K. (1997) Creative productivity: a predictive and explanatory model of career trajectories and landmarks, *Psychological Review*, 104, pp. 66-89.

STERN, H. (1992) Are all linear paired comparison models empirically equivalent?, *Mathematical Social Sciences*, 23, pp. 103-117.

THURSTONE, L. (1927) A law of comparative judgment, *Psychological Review*, 34, pp. 273-286.

UHLIG, H. (1997) Bayesian vector autoregressions with stochastic volatility, *Econometrica*, 65, pp. 59-73.

ZEGER, S. L. & KARIM, M. (1991) Generalized linear models with random effects: a Gibbs sampling approach, *Journal of the American Statistical Association*, 86, pp. 79-86.