The Statistician (2000)
49, *Part* 2, *pp.* 261–276

Dynamic rating of sports teams

Leonhard Knorr-Held

Ludwig–Maximilians University, Munich, Germany

[Received November 1997. Final revision November 1999]

Summary. We consider the problem of dynamically rating sports teams on the basis of categorical outcomes of paired comparisons such as win, draw and loss in football. Our modelling framework is the cumulative link model for ordered responses, where latent parameters represent the strength of each team. A dynamic extension of this model is proposed with close connections to nonparametric smoothing methods. As a consequence, recent results have more influence in estimating current abilities than results in the past. We highlight the importance of using a specific constrained random walk prior for time-changing abilities which guarantees an equal treatment of all teams. Estimation is done with an extended Kalman filter and smoother algorithm. An additional hyperparameter which determines the temporal dynamic of the latent team abilities is chosen on the basis of the optimal one-step-ahead predictive power. Alternative estimation methods are also considered. We apply our method to the results from the German football league *Bundesliga* 1996–1997 and to the results from the American National Basketball Association 1996–1997.

Keywords: Constrained random walk prior; Cumulative link model; Dynamic model; Invariance of estimators; Ordinal response; Paired comparisons; Rating

1. Introduction

When sports teams compete in pairs, they collect scores or goals within a game. Typically, the winning team, the team with more scores at the end of the game, is rewarded with a certain number of points whereas the losing team remains empty handed. Often there is also the possibility of a draw, where both teams have the same score and therefore are rewarded with the same number of points. For example, in football a winning team receives 3 points and a draw is rewarded with 1 point for each team. In general, with or without a draw, results from those paired comparisons are essentially given in ordered categories and those categories determine the standing in the league.

Many approaches to rating sports teams use the difference in scores as the response variable within standard linear model methodology. Other approaches, especially in football, use log-linear Poisson models for the number of goals of both teams. However, it is clear that a rating system should reward a team for winning *per se* and not for running up the score (Harville, 1977). This has led many researchers to propose robust versions for rating sports teams on the basis of the difference in scores. For example, Harville (1977, 1980) proposed to truncate the difference at some predefined cutpoint, whereas Bassett (1997) used L_1-norm regression. Taking Harville's argument to the limit, we use only result categories as the basis to rate teams within a regression model for ordinal categorical data. Thus the main goal here is rating and not prediction, where the score of each team and many other factors should be considered as possible predictors such as the

Address for correspondence: Leonhard Knorr-Held, Institute of Statistics, Ludwig–Maximilians University, Ludwigstrasse 33, D-80539 Munich, Germany.
E-mail: leo@stat.uni-muenchen.de

 0039–0526/00/49261

percentage of ball possession of each team or the number of spectators. Nevertheless, we use the predictive power of our formulation for model fitting.

Let y_{ij} be the result from a paired comparison of team i and team j, where team i is the home team and team j is the visiting team. For example, results from football matches are given in three categories, say

$$y_{ij} = \begin{cases} 1 & \text{if the home team } i \text{ wins,} \\ 2 & \text{for a draw and} \\ 3 & \text{if the visiting team } j \text{ wins.} \end{cases}$$

Cumulative link models are a popular framework to analyse paired comparisons. These models assign a latent ability α_i to each team i, representing the strength of the team. The model is constructed in such a way that the difference in ability of the competing teams $\alpha_i - \alpha_j$ affects the probabilities of the results y_{ij} through a response function F; see for example Agresti (1992).

An important special case is the cumulative logistic model, which boils down to the model of Bradley and Terry (1952) for two possible outcome categories (win or loss). An extension to more than two categories is discussed in Tutz (1986). We give a short review of the classical non-dynamic cumulative link model in Section 2.

When paired comparisons are observed over time, often the teams' performances vary over time because of injuries of important players, changes of the coach or for other reasons. Fahrmeir and Tutz (1994a) introduced dynamic models for longitudinal paired comparison data where abilities are allowed to vary smoothly over time. These models can be seen as Bayesian dynamic models with specific smoothing priors and have close connections with nonparametric smoothing methods, since no functional form is specified for the temporal development of the now time-changing abilities. An extended version of the Kalman filter algorithm for categorical data is used to estimate unknown parameters. Similar dynamic versions of the Bradley–Terry model were proposed independently in Glickman (1993), who used Markov chain Monte Carlo (MCMC) methods for a full Bayesian analysis.

In Section 3 we introduce a dynamic model which is based on the approaches above but has certain amendments. One crucial point is that our model treats all teams symmetrically, which is guaranteed by a specific singular multivariate Gaussian distribution as a smoothing prior for the temporal development of the abilities of the teams. Additional threshold parameters, which represent a possibly existing home advantage, are assumed as time constant and team independent. Estimation and prediction are outlined in Section 4. Posterior mode estimators of the abilities are calculated with the extended Kalman filter and smoother (EKFS) algorithm of Fahrmeir (1992). A hyperparameter, determining the temporal smoothness of the abilities of each team, can be chosen by maximizing one-step-ahead prediction criteria, natural by-products of the EKFS algorithm. Such an approach has an optimality feature, which is understandable, at least in principle, to the public. Alternative ways to estimate the smoothing parameter are an EM type of algorithm and a fully Bayesian analysis by MCMC methods, which are also considered. The corresponding software is available from the author on request.

We apply our method to data from the German football league in 1996–1997 in Section 5. A comparison with estimates by MCMC simulation suggests that inference by the EKFS gives reliable estimates. Furthermore, we analyse a larger data set from the American National Basketball Association, season 1996–1997. Here we have found interesting and pronounced temporal trends in the estimated strengths of several teams. Section 6 concludes with possible modifications and generalizations of our model and with some other comments.

2. Cumulative link model

Let n denote the number of teams in the league, and let R denote the number of categories in the ordinal response scale. A cumulative link model for a comparison y_{ij} of a home team i with a visiting team j is defined by

$$\Pr(y_{ij} \leqslant r) = F(\theta_r + \alpha_i - \alpha_j), \qquad r = 1, \ldots, R-1, \tag{1}$$

where F is a distribution function, $\theta_1 < \ldots < \theta_{R-1}$ are so-called threshold parameters and α_i is the latent ability of team i. For notational convenience we furthermore introduce $\theta_0 = -\infty$ and $\theta_R = \infty$ so that the probability of observing a result $y_{ij} = r$ can be written as

$$\Pr(y_{ij} = r) = F(\theta_r + \alpha_i - \alpha_j) - F(\theta_{r-1} + \alpha_i - \alpha_j), \qquad r = 1, \ldots, R.$$

The threshold parameters can represent a home advantage, which is an important factor for nearly all kinds of sports. For illustration consider a match where both teams have the same ability $\alpha_i = \alpha_j$. The probabilities $\Pr(y_{ij} = r) = F(\theta_r) - F(\theta_{r-1})$ depend now only on the thresholds θ_r and θ_{r-1}. For example, in the case of three categories, the larger θ_1 is, the larger is the probability $\Pr(y_{ij} = 1) = F(\theta_1)$ that the home team wins. Note that the home advantage is assumed to be the same for all teams.

The estimation of $\theta = (\theta_1, \ldots, \theta_{R-1})'$ and $\alpha = (\alpha_1, \ldots, \alpha_n)'$ is done by maximizing the likelihood, the product of individual contributions $\Pr(y_{ij} = r)$ of all matches. Computation can be done conveniently with standard software. Note that the model is unidentifiable because only differences of abilities enter the likelihood. Adding a constant to α_i, $i = 1, \ldots, n$, will not change the likelihood. It is therefore necessary to impose an additional constraint so that the level of abilities is specified. Usual constraints are $\Sigma_{i=1}^{n} \alpha_i = 0$ or $\alpha_n = 0$, say. Invariance of the maximum likelihood (ML) estimator (e.g. Cox and Hinkley (1974)) ensures that estimated abilities are equivalent, whatever constraint was used. For example, the ML estimator $\hat{\alpha}_1, \ldots, \hat{\alpha}_{n-1}$ under the constraint $\alpha_n = 0$ can be used to calculate the ML estimator $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n$ under the constraint $\Sigma_{i=1}^{n} \alpha_i = 0$ by

$$\begin{aligned} \tilde{\alpha}_i &= \hat{\alpha}_i - \frac{1}{n}\sum_{j=1}^{n-1} \hat{\alpha}_j, \qquad i = 1, \ldots, n-1, \\ \tilde{\alpha}_n &= -\frac{1}{n}\sum_{j=1}^{n-1} \hat{\alpha}_j. \end{aligned} \tag{2}$$

In Section 3 we shall show that this invariance no longer holds for dynamic cumulative link models.

The estimated abilities α_i, $i = 1, \ldots, n$, are the basis for rating teams. This approach has certain advantages compared with the standard rating based on adding points in football or counting the number of wins in basketball. In particular, because abilities are estimated simultaneously, the approach automatically adjusts for the strength of the corresponding opponents and for the home advantage. Furthermore, future games can be predicted.

The ML estimator might not exist for some constellations of data, owing to the discrete nature of the data. For example, a team which wins or loses all its matches will have a positive or negative infinite estimated ability respectively. It is therefore advisable to check before the analysis that all teams did not win or lose all their matches. Singularities will also arise if teams can be partitioned into two subsets in which none of the teams in one subset competes against any other team in the other subset. Comparisons within a league, however, are usually scheduled in a way that every team competes against every other team in the league, so this type of singularity will not arise.

3. Dynamic cumulative link model

3.1. The basic model

Suppose now that paired comparisons y_{tij} are observed over time t. We consider time $t = 1, \ldots, T$ as discrete valued and equally spaced, such as days, weeks or months. Our starting-point is to allow in model (1) for time-dependent abilities α_{ti}, $t = 1, \ldots, T$, $i = 1, \ldots, n$:

$$\Pr(y_{tij} \leqslant r) = F(\theta_r + \alpha_{ti} - \alpha_{tj}), \qquad r = 1, \ldots, R-1. \tag{3}$$

This specification allows us to rate sports teams dynamically by estimating time-dependent abilities. We assume Gaussian first-order random walks

$$\alpha_{ti} \sim N(\alpha_{t-1,i}, \sigma^2) \tag{4}$$

as smoothing priors for the abilities of each team i. This is a common assumption for the dynamic modelling of paired comparisons; see Glickman (1993, 1999) and Fahrmeir and Tutz (1994a). The corresponding prior distributions neither impose stationarity nor assume a specific parametric form; in fact model (4) is related to semiparametric and nonparametric smoothing methods as reviewed by Fahrmeir and Knorr-Held (2000). This paper also indicates how to generalize the prior to observations which are not made on a regular time grid.

Model (4) implies that recent matches have more weight for estimating current abilities than results way back in time, which is a natural assumption. The parameter σ^2 determines the weights and hence the loss of memory rate of the random walks. For the limiting case $\sigma^2 = 0$ the model reduces to the classical non-dynamic model (1).

3.2. Ensuring exchangeability by a constrained random walk prior

The crucial point is how to impose a smoothing prior on the abilities $\alpha_t = (\alpha_{t1}, \ldots, \alpha_{tn})'$ without destroying an exchangeable treatment of the teams. This problem occurs, since, as in the time-independent case, we must impose an additional restriction on α_t, $t = 1, \ldots, T$, to assure identifiability. It is, however, not as straightforward as in the time-independent case, because the posterior mode estimator in dynamic models is *not* invariant with respect to the identifiability constraint. We therefore propose a construction based on a specific multivariate singular Gaussian distribution for α_t, which ensures that, marginally, all components of α_t follow a first-order random walk (4), but where the sum $\Sigma_{i=1}^{n} \alpha_{ti}$ is 0 for each time point t. Harvey (1989), pages 41–44, has described a related approach, where seasonal dummy variables sum to 0.

More formally, we assume that α_t follows a constrained multivariate Gaussian random walk

$$\alpha_t = \alpha_{t-1} + u_t, \qquad u_t \sim N(0, Q), \qquad t = 1, \ldots, T, \tag{5}$$

with independent disturbances u_t, $t = 1, \ldots, T$, and initial value α_0 fulfilling $\mathbf{1}'\alpha_0 = 0$. Here $\mathbf{1}$ denotes the vector $(1, 1, \ldots, 1)'$. We now specify a specific singular dispersion matrix Q of rank $n-1$, which ensures that $\mathbf{1}'u_t = 0$ and hence $\mathbf{1}'\alpha_t = 0$ for $t = 1, \ldots, T$. In general there are many such matrices Q but—apart from a proportionality constant—there is only one which treats components of u_t as exchangeable. A detailed discussion can be found in Knorr-Held (1997). For the case where all components of u_t have the same variance σ^2, Q is given by

$$Q = \sigma^2 \left(I - \frac{1}{n} \mathbf{1}\mathbf{1}' \right).$$

Here I denotes the identity matrix. The exchangeable treatment is easily seen, because all non-diagonal entries, and hence all covariances between components of u_t, are equal. Furthermore, all diagonal entries are equal so, marginally and ignoring a multiplicative factor $(n-1)/n$ for the

variance, every component of α_t follows a regular univariate random walk (4). Note that our model implies that components of u_t are *a priori* negatively correlated.

It is easily seen that the sum of components of u_t is 0, because $\mathbf{1}'u_t$ has variance

$$\mathbf{1}'Q\mathbf{1} = \sigma^2\mathbf{1}'\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{1} = 0$$

and is therefore equal to $E(\mathbf{1}'u_t) = 0$ with probability 1. For the more general case with individual variances σ_i^2 for each team, Q is given by $Q = M\Sigma M$ with

$$M = I - \frac{1}{n}\mathbf{1}\mathbf{1}'$$

and $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$.

It is computationally convenient to consider a linear transformation of u_t, say Lu_t, where the first $n-1$ components are Gaussian with *regular* dispersion matrix and the last component is 0 with probability 1. For example,

$$L = \begin{pmatrix} I & -\mathbf{1} \\ \mathbf{1}' & 1 \end{pmatrix}$$

in such a transformation matrix. The first $n-1$ components of Lu_t now have dispersion $P = \sigma^2(I + \mathbf{1}\mathbf{1}')$. We can now perform inference for these $n-1$ *a priori* positively correlated components. The transformation $M(Lu_t)$ with M as defined above, which corresponds to equation (2), retransforms Lu_t to u_t. The whole approach can also be used in the general case where each component has its own variance σ_i^2, $i = 1, \ldots, n$; here $P = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_{n-1}^2) + \sigma_n^2\mathbf{1}\mathbf{1}'$.

Fahrmeir and Tutz (1994a) assumed *independent* Gaussian first-order random walk priors for all except one team, say

$$\alpha_{ti} = \alpha_{t-1,i} + u_{ti} \qquad u_{ti} \sim N(0, \sigma_i^2), \qquad i = 1, \ldots, n-1, \qquad t = 1, \ldots, T, \tag{6}$$

and put the ability of the last team to 0 for each time point t. They also used a similar strategy for other smoothing priors such as the local linear trend model. After estimation, they recentred the estimates to mean 0 for every time point t by using the transformation matrix M from above. However, this approach does not treat teams symmetrically because of the prior independence of the random walks: a change of the reference team will lead to different ratings. This can be best seen for the case where all random walks have the same variance σ^2: the recentred increments u_{ti}, $i = 1, \ldots, n-1$, now have variance $\sigma^2\{1 - 2/n + (n-1)/n^2\}$ whereas increments of the reference team have variance $\sigma^2(n-1)/n^2$, which is smaller for $n > 2$. The estimated abilities of the reference team will show less temporal variation than all the others. For example, for $n = 18$ as in the football example in Section 5.1, the standard deviation of the recentred u_{ti} is 0.970σ, $i = 1, \ldots, n-1$, whereas the standard deviation of u_{tn} is 0.229σ. This difference increases as n increases. In the more general case with team-specific variances σ_i^2, the variance σ_n^2 of team n does not even occur in the model specification (6) and therefore cannot be estimated from the data.

Model (6) can easily be modified to achieve a symmetric treatment of the teams. The replacement of the independent random walk priors for the $n-1$ components with the correlated multivariate random walk with covariance matrix P will give a model which is equivalent to the constrained random walk model (5), as outlined above.

Recently, Glickman and Stern (1998) have used a related, but slightly different, approach for fixing the overall level of time-dependent abilities. They proposed that not α_t but the mean of $\alpha_t|\alpha_{t-1}, \sigma^2$ is centred at 0:

$$\alpha_t|\alpha_{t-1}, \sigma^2 \sim N(M\alpha_{t-1}, \sigma^2 I) \tag{7}$$

with M as defined above. Thus the sum of components of α_t, $\mathbf{1}'\alpha_t$, has expectation 0 and variance $n\sigma^2$ whereas in our model $\mathbf{1}'\alpha_t$ has both expectation and variance equal to 0.

There is an important difference between the Glickman–Stern and our approach. Both are equally valid to predict future outcomes of games, as only differences in team abilities enter the likelihood, implied by equation (3). Also, for a given time t, teams can be ranked or rated on the basis of the Glickman–Stern estimates as well as on our estimates. However, the Glickman–Stern model is not readily usable for judging the temporal development of a given team. For example, suppose that team i has no match scheduled at time t. We would then expect the team's ability to be the same as at time $t-1$, no matter what the abilities of the other teams are at time $t-1$, and this is exactly what our formulation implies. However, in the Glickman–Stern model, the expected ability of this team is

$$\alpha_{ti} = \alpha_{t-1,i} - \frac{1}{n}\sum_{j=1}^{n} \alpha_{t-1,j}.$$

Hence, the expected ability might rise or drop, although the team has not performed in any games at all. Strictly speaking, estimated abilities in the Glickman–Stern model can only be interpreted for a given time t, but not for a given team i. Our model has the advantage that here estimated abilities are valid quantities to assess the performance of a specific team over time. Note that the primary focus in Glickman and Stern (1998) is prediction, and, for this, model (7) is equally well suited.

4. Estimation and prediction

4.1. The extended Kalman filter and smoother

For the moment, consider σ^2 as fixed. We use the EKFS algorithm of Fahrmeir (1992) and Fahrmeir and Tutz (1994a) to estimate time-dependent abilities. A detailed description can be found in Fahrmeir and Tutz (1994b), chapter 8. Threshold parameters θ are estimated by ML given the current (smoothed) estimates of α. Both steps (EKFS estimation of α for fixed θ and ML estimation of θ for fixed α) are alternated until convergence, which takes only a few seconds on a standard workstation or a Pentium processor in both of our applications.

The EKFS algorithm starts with an initial value α_0, a preseason estimate of the abilities, and then subsequently estimates α_t, $t = 1, \ldots, T$, on the basis of all games played until time t. This is called the filtering step and gives filtered estimates $\hat{\alpha}_t$. In a second step, the smoothing step, filtered estimates $\hat{\alpha}_t$, $t = T-1, \ldots, 0$, are smoothed on the basis of all games played until time T. The smoothed estimate of α_0 is used as a new initial value in the next iteration. The algorithm requires an initial value for the prior dispersion Q_0, say, of α_0. In our applications we used $Q_0 = M$, which is weakly informative but avoids numerical problems with more diffuse priors. The final estimates of α_0 are virtually identical whatever starting value for α_0 was used.

This algorithm can be derived as an approximate posterior mode estimator; see for example Fahrmeir and Tutz (1994b). Alternatively, we could use *iterative* EKFS (Fahrmeir and Wagenpfeil, 1997), which gives the exact posterior mode. The computation time increases, however, because an additional level of iteration is required. Furthermore, differences between estimates by non-iterative and iterative EKFS are typically small. See the references above for more details on the properties of posterior mode estimators.

The estimates of α_T and θ can be used to predict future matches. For example, suppose that team j is scheduled to visit team i at the next round ($t = T+1$) to play a match. The probabilities

of the outcomes $\Pr(y_{T+1,ij} = r)$, $r = 1, \ldots, R$, can be estimated by model (3):

$$\Pr(y_{T+1,ij} = r) = F(\hat{\theta}_r + \hat{\alpha}_{Ti} - \hat{\alpha}_{Tj}) - F(\hat{\theta}_{r-1} + \hat{\alpha}_{Ti} - \hat{\alpha}_{Tj}), \tag{8}$$

because the first-order random walk assumption for α_t gives $\hat{\alpha}_T$ as the predicted ability at time $T + 1$. Note that the filtered and smoothed estimates of α_T coincide. More general one-step-ahead predictions are used later to assess the quality of the prediction and to estimate the smoothing parameter σ^2.

4.2. *Estimating the smoothing parameter by maximizing the predictive power*

In the following, we propose to estimate the smoothing parameter σ^2 on the basis of one-step-ahead prediction. Alternative ways are outlined afterwards.

Above it was sketched how filtered estimate $\hat{\alpha}_T$ can be used to predict future games. Filtered estimates $\hat{\alpha}_t$ are also available for $t = 0, \ldots, T - 1$ from the EKFS, so we can perform a retrospective one-step-ahead prediction to assess the model fit. This is done by subsequently predicting outcomes at time $t + 1$ based on filtered estimates $\hat{\alpha}_t$ and comparing the predicted probabilities with the actual observed result by some criterion.

Let N be the total number of paired comparisons over the whole time period. We suppress the dependence on time t and opponents i and j and denote the predicted probabilities $\Pr(y_k = s)$ of outcomes $s = 1, \ldots, R$ by $\hat{p}_k(s)$, $k = 1, \ldots, N$. These predictions are calculated on the basis of filtered estimates $\hat{\alpha}_t$, similar to equation (8), given only game information before period $t + 1$.

A comparison of the actual observed result $y_k = r$ with the predicted probabilities can be done by various criteria. We have implemented the following four concordancy measures:

$$C_1 = \sum_{k=1}^{N} \mathbf{1}\left[\arg\max_{s=1,\ldots,R}\{\hat{p}_k(s)\} = r\right];$$

$$C_2 = \frac{1}{N}\sum_{k=1}^{N} \log\{\hat{p}_k(r)\};$$

$$C_3 = -\frac{1}{N}\sum_{k=1}^{N}\left[\{1 - \hat{p}_k(r)\}^2 + \sum_{s \neq r} \hat{p}_k(s)^2\right];$$

$$C_4 = \frac{1}{N}\sum_{k=1}^{N} \hat{p}_k(r).$$

Criterion C_1 is the number of correctly predicted games, where those outcomes are predicted, which have the highest predictive probability. Criterion C_2 is a log-likelihood criterion and is used also by Glickman (1999) in a similar approach to select hyperparameters. Measure C_3 is based on quadratic loss whereas C_4 is equivalent to the corresponding measure based on absolute loss. Similar criteria are used in discriminant analysis and nonparametric regression for estimating smoothing parameters by cross-validation; see for example Fahrmeir and Tutz (1994b), page 174. We perform a separate estimation of α and θ, as outlined in Section 4.1, for each of a large number of values of σ^2 between 0 and 1. The smoothing parameter σ^2 will then be chosen on the basis of maximal predictive concordancy with respect to the corresponding criterion.

When T is small, often some or all of the criteria are maximized for the limiting case $\sigma^2 = 0$. For larger T, there will often be evidencé for time-changing abilities and (at least some of) the criteria are maximized for truly positive values of σ^2. Note that the '0–1' criterion C_1 has the slightly unattractive feature that it is not continuous as a function of σ^2 so that optimal values of

the smoothing parameter σ^2 are typically within a certain interval. All the other criteria are continuous functions of σ^2.

4.3. Alternative estimation methods

Alternatively, an EM type of algorithm can be implemented to estimate σ^2; see for example Harvey (1989) or Fahrmeir and Tutz (1994b). A disadvantage of this method is rather slow convergence. We have, nevertheless, implemented an EM type of algorithm for σ^2 and report the corresponding estimates in our applications for comparison with the estimates based on predictive concordancy.

For a fully Bayesian analysis, MCMC methods can be used to make simultaneous inference about all the unknown parameters. Such methods require the specification of a prior distribution for the threshold parameters θ and the variance σ^2. Whereas, for the former, improper priors (uniform on the whole real line) can be chosen, for the latter, proper priors must be used, to avoid problems with improper posteriors. Computationally convenient are inverse gamma priors, say $\sigma^2 \sim \text{IG}(a, b)$ with fixed values of a and b. Typical 'weakly informative' choices are $a = 1$ and b small, say 0.001, 0.01 or 0.1 (see for example Besag *et al.* (1995)).

An advantage of MCMC sampling is that the uncertainty about the estimated parameters θ and σ^2 is incorporated in the estimation of α and that samples from predictive distributions can be generated. However, standard MCMC methods do not give *filtered* estimates, an issue that will be further discussed in Section 6. We have also implemented an analysis of dynamic paired comparison models by MCMC simulation to assess the accuracy of our algorithm. The updating of components of θ was done by Gaussian Metropolis proposals (see for example Smith and Roberts (1993)), whereas the updating of abilities was done by (multivariate) conditional prior proposals (Knorr-Held, 1999) for each vector α_t. For inference by MCMC methods, the initial value α_0 can be omitted in the formulation of the model.

5. Applications

In the following we illustrate our method with two applications. We use the dynamic cumulative link model (3) with the logistic response function $F(x) = 1/\{1 + \exp(-x)\}$ together with the constrained random walk prior (5). We have also tried the extreme minimal value distribution function $F(x) = 1 - \exp\{-\exp(x)\}$, which, however, did not fit the data as well as the logistic model in terms of predictability. For simplicity of presentation we do not display (approximate) pointwise credible intervals, which are available from the EKFS.

5.1. The German football league 1996–1997

In the 1996–1997 season of the German football league, the *Bundesliga*, $n = 18$ teams competed for the German championship. The teams met each other twice within the season giving each team a home advantage once. In total, $N = 306$ matches were performed between August 16th, 1996, and May 31st, 1997. We have categorized the timescale in $T = 42$ calendar weeks which gives roughly one match per team and per time point. Note that there was a winter break between December 8th, 1996, and February 13th, 1997, where no matches took place. As noted in Section 1, possible outcomes are given in $R = 3$ categories: win ($y = 1$), draw ($y = 2$) and loss ($y = 3$) of the home team. On the basis of these results, points are assigned to the teams (3 for a win; 1 for a draw) which determine the standings in the league table. Table 1 gives the final standing for the 1996–1997 season.

We have estimated the smoothing parameter σ^2 based on all four prediction criteria. Criterion C_4, the absolute loss criterion, and criterion C_1 agree very closely in fitting the optimal model: C_4

Table 1. Final ranking for the 1996–1997 season

Rank	*Team*	*Total points*	*Points at home*	*Points away*
1	Bayern München	71	43	28
2	Bayer Leverkusen	69	46	23
3	Borussia Dortmund	63	40	23
4	VfB Stuttgart	61	38	23
5	VfL Bochum	53	38	15
6	Karlsruher SC	49	29	20
7	1860 München	49	30	19
8	Werder Bremen	48	35	13
9	MSV Duisburg	45	23	22
10	1. FC Köln	44	28	16
11	Borussia Mönchengladbach	43	32	11
12	Schalke 04	43	22	21
13	Hamburger SV	41	28	13
14	Arminia Bielefeld	40	23	17
15	Hansa Rostock	40	22	18
16	Fortuna Düsseldorf	33	20	13
17	SC Freiburg	29	22	7
18	St Pauli	27	19	8

is maximized for $\sigma^2 = 0.0171$ with a value of 0.4351, compared with 0.4329 for $\sigma^2 = 0$. Criterion C_1 is maximized around $\sigma^2 = 0.02$ with 160 correctly predicted games (157 for $\sigma^2 = 0$). The EM-type estimate is slightly lower with $\hat{\sigma}^2 = 0.0114$. Convergence was very slow. This indicates that there is not much information on temporal variation of abilities in the data. In fact, the other two criteria C_2 and C_3 are maximized for the limiting case $\sigma^2 = 0$. The reason might be that they both give relatively more weight to small values of $\hat{p}(k)$ than do C_1 and C_4. Small estimated probabilities are more likely for large values of σ^2 where estimated abilities have more temporal variation and cause estimated probabilities to be more extreme.

Figs 1 and 2 show filtered and smoothed estimated abilities of all 18 teams for $\sigma^2 = 0.0171$. The smoothed estimates show rather different patterns for the various teams and demonstrate the advantages of our nonparametric dynamic model. Note also that within the winter break ($t = 18, \ldots, 26$) filtered estimates are horizontal lines because of the prior model.

Later champions Bayern München had quite a time constant performance with smoothed abilities between 1.1 and 1.2. Other teams, however, had a substantial time-dependent performance. For example, Borussia Mönchengladbach had a rather poor performance before the winter break whereas its estimated ability at the end of the season was even slightly above average with a value of 0.06. Interestingly, this change of performance coincided with the dismissal of the head coach shortly before Christmas. SC Freiburg showed a similar dynamic but, in contrast with Borussia Mönchengladbach, the better performance towards the end of the season did not pay off: the team was relegated from the first to the second division.

The estimates of the threshold parameters are $\hat{\theta} = (0.050, 1.33)$ and reflect the strong home advantage that was already apparent in the raw data: 51.1% of all matches were won by the home team and only 26.1% by the visiting team.

For comparison, we have analysed this data set also by MCMC simulation. Table 2 gives posterior mean estimates of σ^2 and θ for $a = 1.0$ and various values of b. The estimates have a strong sensitivity, especially $\hat{\sigma}^2$ with respect to the prior for σ^2. Consequently, the estimated abilities differ very much, with the degree of smoothness determined by $\hat{\sigma}^2$. Unfortunately, there are no clear guidelines about how to choose the prior for σ^2. For $a = 1$ and $b = 0.4$, which comes

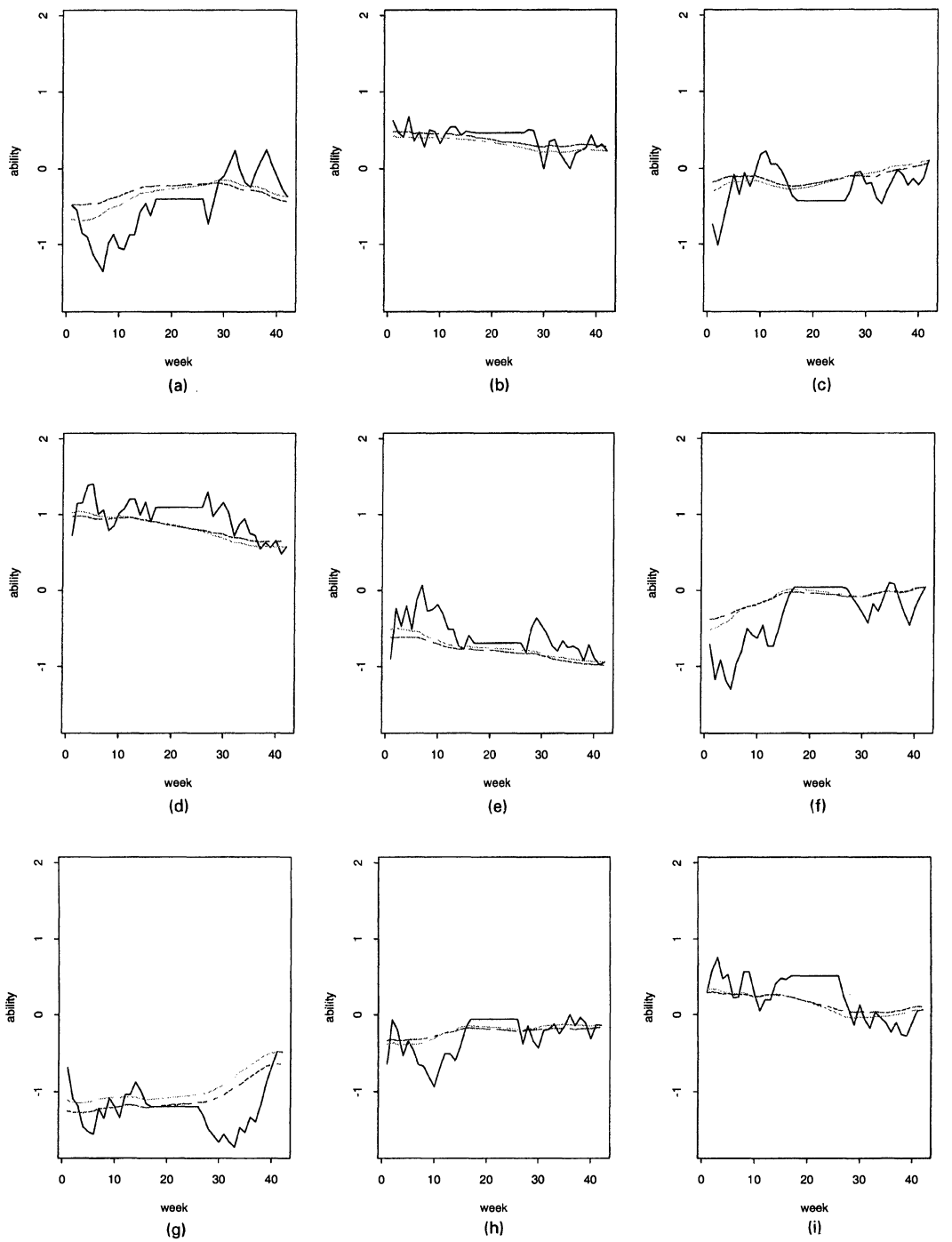

Fig. 1. Football data (———, filtered time-changing abilities; ·····, smoothed time-changing abilities; - - - - -, MCMC estimates): (a) Arminia Bielefeld; (b) VfL Bochum; (c) Werder Bremen; (d) Borussia Dortmund; (e) Fortuna Düsseldorf; (f) MSV Duisburg; (g) SC Freiburg; (h) Hamburger SV; (i) Karlsruher SC

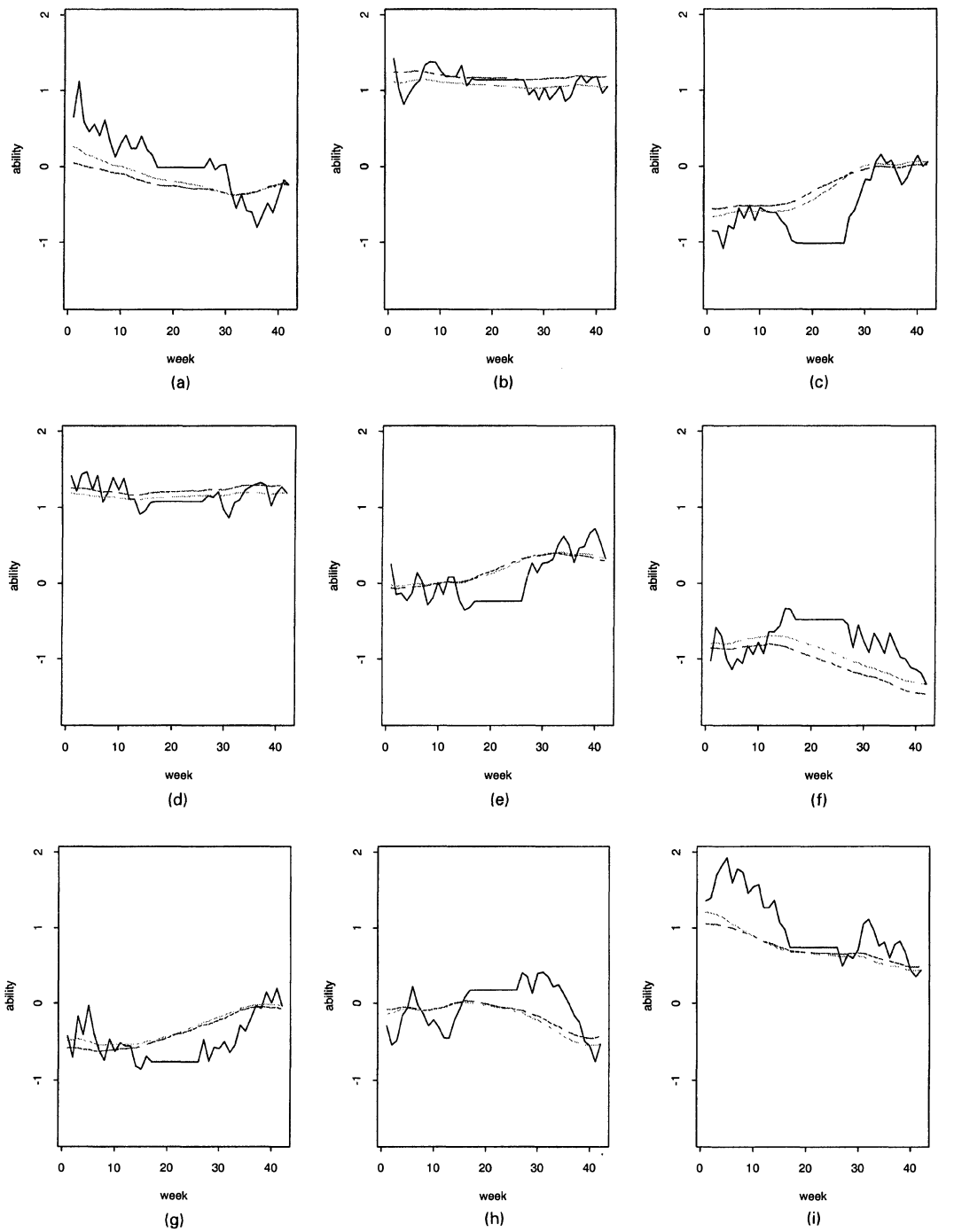

Fig. 2. Football data (———, filtered time-changing abilities; ·····, smoothed time-changing abilities; - - - - -, MCMC estimates): (a) I. FC Köln; (b) Bayer Leverkusen; (c) Borussia Mönchengladbach; (d) Bayern München; (e) 1860 München; (f) St Pauli; (g) Hansa Rostock; (h) Schalke 04; (i) VfB Stuttgart

closest to the C_4-estimate $\hat{\sigma}^2 = 0.0171$, we have calculated posterior mean estimates of the ability parameters, and these are also displayed in Figs 1 and 2. From these pictures it can be seen that the MCMC estimates are quite similar to the smoothed estimates from the EKFS. Hence, the Kalman filter algorithm gives quite reliable results here. The small differences, which seem to depend on the absolute value of $\hat{\alpha}_{it}$, may be caused by the approximateness of the EKFS algorithm, the slightly different model formulation for the MCMC algorithm with all the parameters stochastic and without α_0, or by skewed posterior distributions where the posterior means and modes do not coincide. In a second MCMC analysis, we fixed $\sigma^2 = 0.017$ and the small differences between means and modes decreased slightly.

5.2. 1996–1997 season of the National Basketball Association

In the American National Basketball Association $n = 29$ teams performed paired comparisons in the 1996–1997 season. We analyse $N = 1189$ games excluding results from the play-off matches. These games took place between November 1st, 1996, and April 20th, 1997, which gives a total of $T = 171$ calendar days. Note that in basketball there is not the possibility of a draw because of the overtime rule. Games can only end with $R = 2$ categories.

Our model fit criteria show behaviour similar to that for the football data: C_1 and C_4 again agree very closely. Criterion C_1 is optimal around $\sigma^2 = 0.005$ with 842 correctly predicted games (837 for $\sigma^2 = 0$). C_4 has an optimal value of 0.6366 for $\sigma^2 = 0.00595$, compared with 0.6334 for $\sigma^2 = 0$. The EM-type estimate is also quite close with $\hat{\sigma}^2 = 0.00425$. The log-likelihood criterion C_2 and the quadratic loss function criterion C_3, however, again prefer the non-dynamic model $\hat{\sigma}^2 = 0$.

The following results are based on the C_4-optimal value $\sigma^2 = 0.00595$. The threshold parameter was estimated by $\hat{\theta} = 0.37$, reflecting a substantial home advantage. Filtered and smoothed estimated abilities of selected teams are shown in Fig. 3. Interestingly the Chicago Bulls, the later champions, showed a steadily declining performance. They might have not played with their full force towards the end of the season, having already qualified for the play-off matches. Other teams such as the Houston Rockets, the Phoenix Suns or Utah Jazz have a very remarkable dynamic which would have been overlooked by a parametric model, where, for example, abilities are assumed to develop linearly or quadratically in time.

6. Concluding remarks

This paper has discussed the application of dynamic cumulative link models for rating sports teams. Our prior model ensures a symmetric treatment of all teams, assuming a multivariate singular Gaussian random walk prior with exchangeable components. Similar priors can also be used if the response variable is the difference in scores between the home and the visiting team, which is the more traditional approach to rating and prediction, because estimation can be done

Table 2. Parameter estimates (posterior means) by MCMC sampling for various hyperprior specifications

a	b	$\hat{\sigma}^2$	$\hat{\theta}_1$	$\hat{\theta}_2$
1	0.01	0.0005	0.068	1.24
1	0.1	0.0049	0.072	1.26
1	0.4	0.0183	0.076	1.30
1	1	0.0424	0.080	1.35

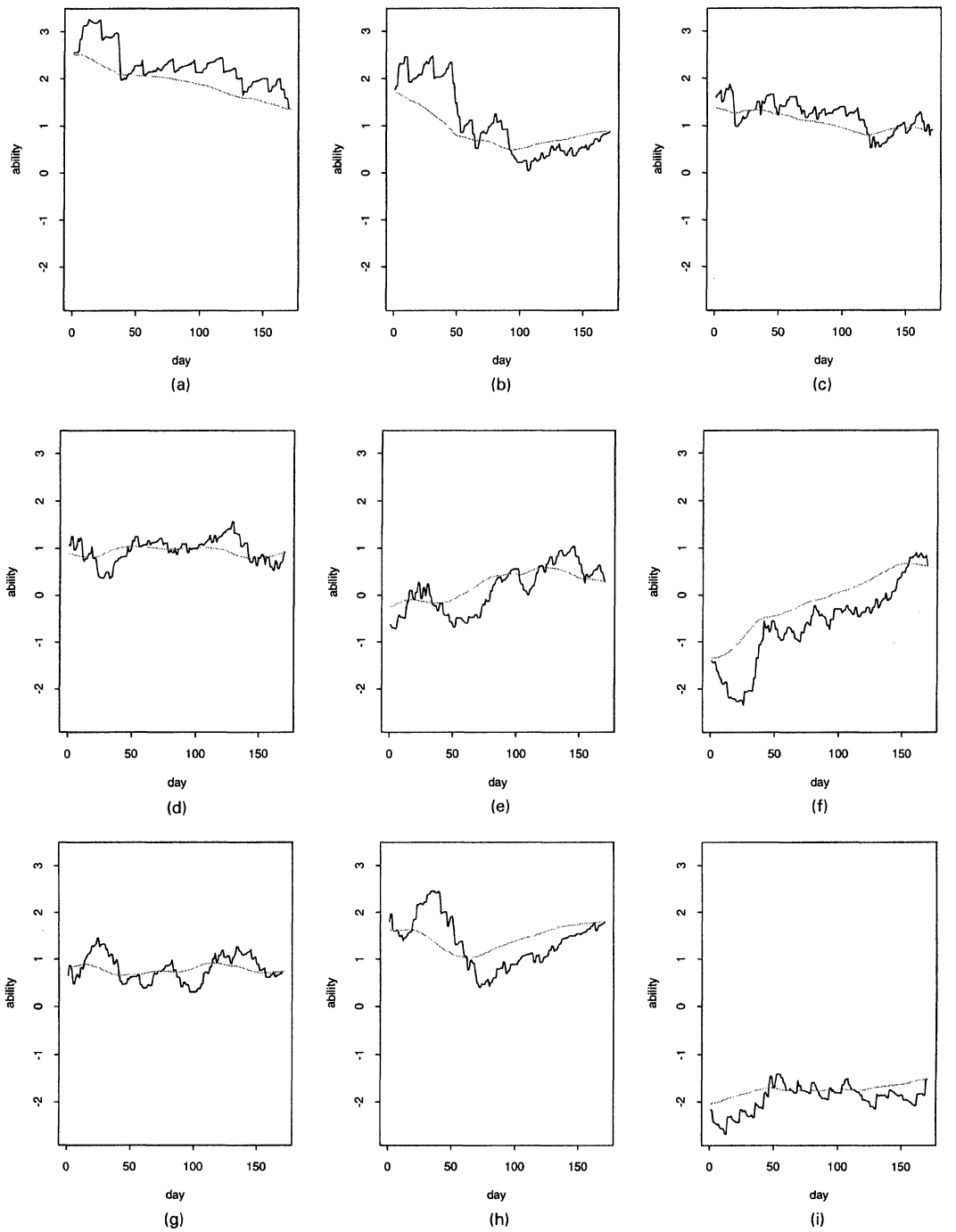

Fig. 3. Basketball data (———, filtered time-changing abilities; ·····, smoothed time-changing abilities): (a) Chicago Bulls; (b) Houston Rockets; (c) Miami Heat; (d) New York Knicks; (e) Orlando Magic; (f) Phoenix Suns; (g) Seattle Supersonics; (h) Utah Jazz; (i) Vancouver Grizzlies

within the standard linear model; see for example Harville (1977, 1980) or Harville and Smith (1994). A dynamic approach for modelling the difference in scores within the state space model is proposed in Sallas and Harville (1988). There are also areas outside dynamic models for paired comparisons where constrained random walk priors are potentially useful. For example, the dynamic modelling of categorical covariate effects with constrained random walk priors is used in Knorr-Held and Besag (1998) for space–time modelling of disease risk data.

The estimation of the smoothing parameter turned out to be difficult. The reason seems to be that the type of categorical data does not provide much information about the temporal variation in teams' abilities. Among the four different measures, both the 0–1 and the absolute loss function criterion showed good performance whereas the other two choices have been disappointing. As an alternative, an EM type of algorithm can be used, which, however, has rather poor convergence properties. Fully Bayesian estimation by MCMC sampling was very sensitive with respect to hyperprior specifications.

A referee suggested considering (independent) stationary autoregressive processes as a prior model for the abilities of each team $i = 1, \ldots, n$:

$$\alpha_{ti} \sim N\{\mu_i + \phi(\alpha_{t-1,i} - \mu_i), \sigma^2\},$$

$$\alpha_{0i} \sim N\left(\mu_i, \frac{\sigma^2}{1-\phi^2}\right).$$

The parameter μ_i could be interpreted as the long-term average ability of team i. Such a model has the advantage that, *a priori*, the abilities of the teams will not in the long run tend towards ∞ or $-\infty$. Estimation can be done by MCMC sampling. Also, identifiability can easily be achieved by fixing the mean ability of a particular team, say the nth, at 0, i.e. $\mu_n = 0$, and it does not matter here which team's mean performance is fixed. However, such a model is less parsimonious and therefore problems of sensitivity with respect to now two hyperparameters σ^2 and ϕ are likely to increase. For data over a rather short time period, like in both of our applications, it seems that our approach will be sufficiently flexible and the long-run behaviour properties of the prior model are not as crucial from an applied point of view.

There are several possible generalizations of our model. As already noted in Section 2, each team can have assigned a team-specific smoothing parameter σ_i^2. This might be appropriate if there are substantially more data than in both of our applications, where information about the temporal variation, summarized in σ^2, seemed to be rather small. Furthermore, it will be very difficult, for moderate to large sizes of n, to find the optimal values of $\sigma_1^2, \ldots, \sigma_n^2$ on the basis of one-step-ahead prediction. Similar comments apply to more enhanced smoothing priors such as the second-order random walk, where independent first differences $\alpha_t - \alpha_{t-1}$ are replaced by second differences $\alpha_t - 2\alpha_{t-1} + \alpha_{t-2}$, or the local linear trend model (Harvey, 1989). On the basis of our experience with these priors, we believe that they have limited use as long as the rating of sports teams within one season is considered as in our two applications. They might be useful if considerably more data are observed over a longer period.

Throughout this paper we have used constant threshold parameters. This assumption can be relaxed, allowing for team-specific or time-dependent threshold parameters. Harville and Smith (1994) analysed college basketball results with various models and found team-to-team differences in the home advantage to be relatively small. Knorr-Held (1997) analysed the *Bundesliga* (1995–1996 season) by MCMC methods with additional team-specific random effects for the thresholds, but there was not much evidence for home advantage heterogeneity here either. Fahrmeir and Tutz (1994a) allowed for time-dependent threshold parameters in an analysis of the *Bundesliga* but found threshold estimates to be stable even over a period of 23 years. We have

therefore used the simple model with constant threshold parameters. If necessary, it would be no problem to extend our approach to more general models.

For the application considered, we prefer the EKFS for statistical inference rather than MCMC methods. Although MCMC sampling is a more elaborate approach, it seems to be much easier to obtain *filtered* estimates by the EKFS. These filtered estimates can be used to assess the predictive power and to choose the smoothing parameter. Furthermore, in a full Bayesian analysis, there often seems to be strong sensitivity with respect to hyperprior specifications. However, recently several interesting proposals have been made to obtain filtered estimates by *dynamic* MCMC methods, e.g. Gordon *et al.* (1993), Berzuini *et al.* (1997) and Pitt and Shephard (1999). Applications of these approaches to dynamic ordered paired comparison systems are, because of the high dimensionality of the model, beyond the scope of this paper.

Acknowledgements

The author expresses thanks to Ludwig Fahrmeir, the Associate Editor and two referees for several suggestions which have improved the paper.

References

Agresti, A. (1992) Analysis of ordinal paired comparison data. *Appl. Statist.*, **41**, 287–297.
Bassett, Jr, G. W. (1997) Robust sports ratings based on least absolute errors. *Am. Statistn*, **51**, 99–105.
Berzuini, C., Best, N. G., Gilks, W. R. and Larizza, C. (1997) Dynamic conditional independence models and Markov chain Monte Carlo methods. *J. Am. Statist. Ass.*, **92**, 1403–1412.
Besag, J. E., Green, P. J., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems (with discussion). *Statist. Sci.*, **10**, 3–66.
Bradley, R. A. and Terry, M. E. (1952) Rank analysis of incomplete block designs: I, The method of pair comparisons. *Biometrika*, **39**, 324–345.
Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics.* London: Chapman and Hall.
Fahrmeir, L. (1992) Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *J. Am. Statist. Ass.*, **87**, 501–509.
Fahrmeir, L. and Knorr-Held, L. (2000) Dynamic and semiparametric models. In *Smoothing and Regression: Approaches, Computation and Application* (ed. M. Schimek), ch. 18. New York: Wiley. To be published.
Fahrmeir, L. and Tutz, G. (1994a) Dynamic stochastic models for time-dependent ordered paired comparison systems. *J. Am. Statist. Ass.*, **89**, 1438–1449.
———(1994b) *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer.
Fahrmeir, L. and Wagenpfeil, S. (1997) Penalized likelihood estimation and iterative Kalman smoothing for non-Gaussian dynamic regression models. *Comput. Statist. Data Anal.*, **24**, 295–320.
Glickman, M. E. (1993) Paired comparison models with time-varying parameters. *PhD Dissertation*. Department of Statistics, Harvard University, Cambridge.
———(1999) Parameter estimation in large dynamic paired comparison experiments. *Appl. Statist.*, **48**, 377–394.
Glickman, M. E. and Stern, H. S. (1998) A state-space model for National Football League scores. *J. Am. Statist. Ass.*, **93**, 25–35.
Gordon, N. J., Salmond, D. J. and Smith, A. F. M. (1993) A novel approach to non-linear and non-Gaussian Bayesian state estimation. *IEE Proc.* F, **140**, 107–133.
Harvey, A. C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.
Harville, D. A. (1977) The use of linear-model methodology to rate high school or college football teams. *J. Am. Statist. Ass.*, **72**, 278–289.
———(1980) Predictions for National Football League games via linear-model methodology. *J. Am. Statist. Ass.*, **75**, 516–524.
Harville, D. A. and Smith, M. H. (1994) The home-court advantage: how large is it, and does it vary from team to team? *Am. Statistn*, **48**, 22–28.
Knorr-Held, L. (1997) *Hierarchical Modelling of Discrete Longitudinal Data; Applications of Markov Chain Monte Carlo*. Munich: Utz.
———(1999) Conditional prior proposals in dynamic models. *Scand. J. Statist.*, **26**, 129–144.
Knorr-Held, L. and Besag, J. (1998) Modelling risk from a disease in time and space. *Statist. Med.*, **17**, 2045–2060.
Pitt, M. K. and Shephard, N. (1999) Filtering via simulation: auxiliary particle filters. *J. Am. Statist. Ass.*, **94**, 590–599.

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

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3–23.

Tutz, G. (1986) Bradley–Terry–Luce models with an ordered response. *J. Math. Psychol.*, **30**, 306–316.