A state-space model for National Football League scores

Mark E. Glickman
Department of Mathematics
Boston University
111 Cummington Street
Boston, MA 02215

Hal S. Stern
Department of Statistics
Snedecor Hall
Iowa State University
Ames, IA 50011

Abstract

This paper develops a predictive model for National Football League (NFL) game scores using data from the period 1988–1993. The parameters of primary interest, measures of team strength, are expected to vary over time. Our model accounts for this source of variability by modeling football outcomes using a state-space model that assumes team strength parameters follow a first-order autoregressive process. Two sources of variation in team strengths are addressed in our model; week-to-week changes in team strength due to injuries and other random factors, and season-to-season changes resulting from changes in personnel and other longer-term factors. Our model also incorporates a home-field advantage while allowing for the possibility that the magnitude of the advantage may vary across teams. The aim of the analysis is to obtain plausible inferences concerning team strengths and other model parameters, and to predict future game outcomes. Iterative simulation is used to obtain samples from the joint posterior distribution of all model parameters. Our model appears to outperform the Las Vegas "betting line" on a small test set consisting of the last 110 games of the 1993 regular NFL season.

Keywords: Bayesian diagnostics, dynamic models, Kalman filter, Markov chain Monte Carlo, predictive inference.

1 Introduction

Prediction problems in many settings, e.g., finance, political elections, and in this article football, are complicated by the presence of several sources of variation for which a predictive model must account. For National Football League (NFL) games, team abilities may vary from year to year due to changes in personnel and overall strategy. In addition, team abilities may vary within a season due to injuries, team psychology, and promotion/demotion of players. Team performance may also vary depending on the site of a game. This paper describes an approach to modeling NFL scores using a normal linear state-space model that accounts for these important sources of variability.

The state-space framework for modeling a system over time incorporates two different random processes. The distribution of the data at each point in time is specified conditional on a set of time-indexed parameters. A second process describes the evolution of the parameters over time. For many specific state-space models, including the model developed in this paper, posterior inferences about parameters cannot be obtained analytically. We therefore use Markov chain Monte Carlo (MCMC) methods, namely Gibbs sampling (Geman and Geman 1984; Gelfand and Smith 1990) as a computational tool for studying the posterior distribution of the parameters of our model. Pre-MCMC approaches to the analysis of linear state-space models include Harrison and Stevens (1976) and West and Harrison (1990). More recent work on MCMC methods includes Glickman (1993), Carter and Kohn (1994) and Fruhwirth-Schnatter (1994) who develop efficient procedures for fitting normal linear state-space models. Carlin et al. (1992), Shephard (1994) and de Jong and Shephard (1995) are only a few of the recent papers in the growing literature on MCMC approaches to non-linear and non-Gaussian state-space models.

The Las Vegas "point spread" or "betting line" of a game, provided by Las Vegas oddsmakers, can be viewed as the "experts'" prior predictive estimate of the difference in game scores. A number of recent papers have examined the point spread as a predictor of game outcomes, including Zuber et al. (1985), Amoako-Adu et al. (1985), and Stern (1991). Stern, in particular, shows that modeling the score difference of a game to have a mean equal to the point spread is empirically justifiable. We demonstrate that our model performs at least as well as the Las Vegas line for predicting game outcomes for the latter half of the 1993 season.

Other work on modeling NFL football outcomes (Thompson 1975; Stefani 1977; Stefani 1980; Stern 1992) has not incorporated the stochastic nature of team strengths. Our model is closely related to one examined by Harville (1977, 1980) and Sallas and Harville (1988), though the analysis that we perform differs in a number of ways. First, we create prediction inferences by sampling from the joint posterior distribution of all model parameters rather than fixing some parameters at point estimates prior to prediction. Our model also describes a richer structure in the data, accounting for the possibility of shrinkage towards the mean of team strengths over time. Finally, the analysis presented here incorporates model checking and sensitivity analysis aimed at assessing the propriety of the state-space model.

2 A model for football game outcomes

Let $y_{ii'}$ denote the outcome of a football game between team i and team i' where teams are indexed by the integers from 1 to p. For our data set, $p = 28$. We take $y_{ii'}$ to be the difference between the score of team i and the score of team i'. The NFL game outcomes can be modeled as approximately normally distributed with a mean that depends on the relative strength of the teams involved in the game and the site of the game. We assume that at week j of season k, the strength or ability of team i can be summarized by a parameter $\theta_{(k,j)i}$. We let $\boldsymbol{\theta}_{(k,j)}$ denote the vector of p team-ability parameters for week j of season k. An additional set of parameters, α_i, $i = 1, \ldots, p$, measure the magnitude of team i's advantage when playing at its home stadium rather than at a neutral site. These home-field advantage (HFA) parameters are assumed to be independent of time, but may vary across teams. We let $\boldsymbol{\alpha}$ denote the vector of p HFA parameters. The mean outcome for a game between team i and team i' played at the site of team i during week j of season k is assumed to be

$$\theta_{(k,j)i} - \theta_{(k,j)i'} + \alpha_i.$$

We can express the distribution for the outcomes of all $n_{(k,j)}$ games played during week j of season k as

$$\boldsymbol{y}_{(k,j)} \,|\, \tilde{\boldsymbol{X}}_{(k,j)}, \tilde{\boldsymbol{\theta}}_{(k,j)}, \phi \;\sim\; \mathrm{N}(\tilde{\boldsymbol{X}}_{(k,j)}\tilde{\boldsymbol{\theta}}_{(k,j)}, \phi^{-1}\mathbf{I}_{n_{(k,j)}}),$$

where $\boldsymbol{y}_{(k,j)}$ is the vector of game outcomes, $\tilde{\boldsymbol{X}}_{(k,j)}$ is the $n_{(k,j)} \times 2p$ design matrix for week j of season k (described in detail below), $\tilde{\boldsymbol{\theta}}_{(k,j)} = (\boldsymbol{\theta}_{(k,j)}, \boldsymbol{\alpha})$ is the vector of p team-ability parameters and p HFA parameters, and ϕ is the regression precision of game outcomes. We let $\tau^2 = \phi^{-1}$ denote the variance of game outcomes conditional on the mean. The row of the matrix $\tilde{\boldsymbol{X}}_{(k,j)}$ for a game between team i and team i' has the value 1 in the i-th column (corresponding to the first team involved in the game), -1 in the i'-th column (the second team), and 1 in the $(p+i)$-th column (corresponding to the home-field advantage) if the first team played on its home field. If the game were played at the site of the second team (team i') then home field would be indicated by a -1 in the $(p+i')$-th column. Essentially, each row has entries 1 and -1 to indicate the participants and then a single entry in the column corresponding to the home team's home-field advantage parameter (1 if it is the first team at home or -1 if it is the second team). The designation of one team as the first team and the other as the second team is arbitrary and does not affect the interpretation of the model, nor does it affect inferences.

We take K to be the number of seasons of available data. For our particular data set there are $K = 6$ seasons. We let g_k, for $k = 1, \ldots, K$, denote the total number of weeks of data available in season k. Data for the entire season are available for season k, $k = 1, \ldots, 5$, with g_k varying from 16 to 18. We take $g_6 = 10$, using the data from the remainder of the sixth season to perform predictive inference. Additional details about the structure of the data are provided in Section 4.

Our model incorporates two sources of variation related to the evolution of team ability over time. The evolution of strength parameters between the last week of season k and the first week of season $k+1$ is assumed to be governed by

$$\boldsymbol{\theta}_{(k+1,1)} \,|\, \beta_s, \boldsymbol{\theta}_{(k,g_k)}, \phi, \omega_s \;\sim\; \mathrm{N}(\beta_s \boldsymbol{G} \boldsymbol{\theta}_{(k,g_k)}, (\phi\omega_s)^{-1}\mathbf{I}_p),$$

where $\boldsymbol{G}$ is the matrix which maps the vector $\boldsymbol{\theta}_{(k,g_k)}$ to $\boldsymbol{\theta}_{(k,g_k)} - \mathrm{ave}(\boldsymbol{\theta}_{(k,g_k)})$, β_s is the between-season regression parameter that measures the degree of shrinkage ($\beta_s < 1$) or expansion ($\beta_s > 1$) in team abilities between seasons, and the product $\phi\omega_s$ is the between-season evolution precision. This particular parametrization for the evolution precision simplifies the distributional calculus involved in model fitting. We let $\sigma_s^2 = (\phi\omega_s)^{-1}$ denote the between-season evolution variance. Then ω_s is the ratio of variances, τ^2/σ_s^2.

The matrix $\beta_s \boldsymbol{G}$ maps the vector $\boldsymbol{\theta}_{(k,g_k)}$ to another vector centered at 0, and then shrunk or expanded around zero. We use this mapping because the distribution of the game outcomes $\boldsymbol{y}_{(k,j)}$ is a function only of differences in the team ability parameters; the distribution is unchanged if a constant is added to or subtracted from each team's ability parameter. The mapping $\boldsymbol{G}$ translates the distribution of team strengths to be centered at 0, though it is understood that shrinkage or expansion is actually occurring around the mean team strength (which may be drifting over time). The season to season variation is due mainly to personnel changes (new players or coaches). One would expect $\beta_s < 1$ since the player assignment process is designed to assign the best young players to the teams with the weakest performance in the previous season.

We model short-term changes in team performance by incorporating evolution of ability parameters between weeks,

$$\boldsymbol{\theta}_{(k,j+1)} \,|\, \beta_w, \boldsymbol{\theta}_{(k,j)}, \phi, \omega_w \;\sim\; \mathrm{N}(\beta_w \boldsymbol{G} \boldsymbol{\theta}_{(k,j)}, (\phi\omega_w)^{-1}\mathbf{I}_p),$$

where the matrix $\boldsymbol{G}$ is as before, β_w is the between-week regression parameter, and $\phi\omega_w$ is the between-week evolution precision. Analogous to the between-season component of the model, we let $\sigma_w^2 = (\phi\omega_w)^{-1}$ denote the variance of the between-week evolution, so that $\omega_w = \tau^2/\sigma_w^2$. Week-to-week changes represent short-term sources of variation, e.g., injuries and team confidence level. It is likely that $\beta_w \approx 1$ since there is no reason to expect such short-term changes would tend to equalize the team parameters ($\beta_w < 1$) or accentuate differences ($\beta_w > 1$).

There are several simplifying assumptions built into this model that are worthy of comment. We model differences in football scores, which can only take on integer values, as approximately normally distributed conditional on team strengths. The rules of football suggest that some outcomes (e.g., 3 or 7) are much more likely than others. Rosner (1976) models game outcomes as a discrete distribution that incorporates the rules for football scoring. However, previous work (Stern 1991, Harville 1980, Sallas and Harville 1988, as well as others) has shown that the normality assumption is not an unreasonable approximation, especially when one is not interested in computing probabilities for exact outcomes, but rather for ranges of outcomes (e.g., whether the score difference is greater than 0). Several parameters, notably the regression variance τ^2, and the evolution variances σ_w^2, σ_s^2, are assumed to be the same for all teams and for all seasons. This rules out the possibility of teams with especially erratic performance. We explore the adequacy of these modeling assumptions using posterior predictive model checks (Rubin 1984; Gelman et al. 1996) in Section 5.

Prior distributions of model parameters are centered at values that seem reasonable based on our knowledge of football. In each case, the chosen distribution is widely dispersed so that before long the data will play a dominant role. We assume the following prior distributions,

$$\begin{aligned} \phi &\sim \text{Gamma}(0.5, 0.5(100)) \\ \omega_w &\sim \text{Gamma}(0.5, 0.5/60) \\ \omega_s &\sim \text{Gamma}(0.5, 0.5/16) \\ \beta_s &\sim \text{N}(0.98, 1) \\ \beta_w &\sim \text{N}(0.995, 1). \end{aligned}$$

Our prior distribution on ϕ corresponds to a harmonic mean of 100, which is roughly equivalent to a ten point standard deviation, τ, for game outcomes conditional on knowing the teams abilities. This is close to, but a bit lower than, Stern's (1991) estimate of $\hat{\tau} = 13.86$ derived from a simpler model. In combination with this prior belief about ϕ, the prior distributions on ω_w and ω_s assume harmonic means of σ_w^2 and σ_s^2 equal to 100/60 and 100/16, respectively, indicating our belief that the changes in team strength between seasons are likely to be larger than short-term changes in team strength. There is little information currently available about σ_w^2 and σ_s^2 which is represented by the 0.5 degrees of freedom. The prior distributions on the regression parameters assume shrinkage towards the mean team strength, with a greater degree of shrinkage for the evolution of team strengths between seasons. It is not necessary in the context of our state-space model to restrict the modulus of the regression parameters, which are assumed to be equal for every week and season, to be less than 1 as long as our primary concern is for parameter summaries and local prediction rather than long-range forecasts.

The only remaining prior distributions are those for the initial team strengths in 1988, $\boldsymbol{\theta}_{(1,1)}$, and the home-field advantage parameters, $\boldsymbol{\alpha}$. For team strengths at the onset of the 1988 season, we could try to quantify our knowledge perhaps by examining 1987 final records and statistics. We have chosen instead to use an exchangeable prior distribution as a starting point, ignoring any pre-1988 information,

$$\boldsymbol{\theta}_{(1,1)} \,|\, \phi, \omega_o \;\sim\; \mathrm{N}(\mathbf{0}, (\omega_o \phi)^{-1}\mathbf{I}_p) \tag{1}$$

where we assume

$$\omega_o \sim \mathrm{Gamma}(0.5, 0.5/6). \tag{2}$$

Let $\sigma_o^2 = (\omega_o \phi)^{-1}$ denote the prior variance of initial team strengths. Our prior distribution for ω_o in combination with the prior distribution on ϕ implies that σ_o^2 has prior harmonic mean of 100/6 based on 0.5 degrees of freedom. The *a priori* difference between the best and worst teams would therefore be about $4\sigma_o = 16$ points.

We assume the α_i have independent prior distributions

$$\boldsymbol{\alpha} \,|\, \phi, \omega_h \;\sim\; \mathrm{N}(3, (\omega_h \phi)^{-1}\mathbf{I}_p) \tag{3}$$

with

$$\omega_h \sim \mathrm{Gamma}(0.5, 0.5/6). \tag{4}$$

We assume a prior mean of 3 for the α_i, believing that competing on one's home field conveys a small but persistent advantage. If we let $\sigma_h^2 = (\omega_h \phi)^{-1}$ denote the prior variance of the home-field advantage parameters, then our prior distributions for ω_h and ϕ implies that σ_h^2 has prior harmonic mean of 100/6 based on 0.5 degrees of freedom.

3 Model Fitting and Prediction

We fit and summarize our model using Markov chain Monte Carlo techniques, namely the Gibbs sampler (Geman and Geman 1984; Gelfand and Smith 1990). Let $\boldsymbol{Y}_{(K,g_K)}$ represent all observed data through week (K, g_K). The Gibbs sampler is implemented by drawing alternately in sequence from the following three conditional posterior distributions:

- $f(\boldsymbol{\theta}_{(1,1)}, \ldots, \boldsymbol{\theta}_{(K,g_K)}, \boldsymbol{\alpha}, \phi \,|\, \omega_o, \omega_h, \omega_w, \omega_s, \beta_s, \beta_w, \boldsymbol{Y}_{(K,g_K)})$

- $f(\omega_o, \omega_h, \omega_w, \omega_s \mid \boldsymbol{\theta}_{(1,1)}, \ldots, \boldsymbol{\theta}_{(K,g_K)}, \boldsymbol{\alpha}, \phi, \beta_s, \beta_w, \boldsymbol{Y}_{(K,g_K)})$
- $f(\beta_s, \beta_w \mid \boldsymbol{\theta}_{(1,1)}, \ldots, \boldsymbol{\theta}_{(K,g_K)}, \boldsymbol{\alpha}, \phi, \omega_o, \omega_h, \omega_w, \omega_s, \boldsymbol{Y}_{(K,g_K)}))$.

A detailed description of the conditional distributions appears in Appendix A. Once the Gibbs sampler has converged, inferential summaries are obtained by using the empirical distribution of the simulations as an estimate of the posterior distribution.

An important use of the fitted model is the prediction of game outcomes. Assume that the model has been fit via the Gibbs sampler to data through week g_K of season K, thereby obtaining m posterior draws of the final team ability parameters $\boldsymbol{\theta}_{(K,g_K)}$, the HFA parameters $\boldsymbol{\alpha}$, and the precision and regression parameters. Denote the entire collection of the parameters as $\boldsymbol{\eta}_{(K,g_K)} = (\omega_w,\ \omega_s,\ \omega_o,\ \omega_h,\ \beta_w,\ \beta_s,\ \phi,\ \boldsymbol{\theta}_{(K,g_K)},\ \boldsymbol{\alpha})$. Given the design matrix for the next week's games, $\tilde{\boldsymbol{X}}_{(K,g_K+1)}$, the posterior predictive distribution of next week's game outcomes, $\boldsymbol{y}_{(K,g_K+1)}$, is given by

$$\boldsymbol{y}_{(K,g_K+1)} \mid \boldsymbol{Y}_{(K,g_K)}, \boldsymbol{\eta}_{(K,g_K)} \sim \mathrm{N}(\tilde{\boldsymbol{X}}_{(K,g_K+1)} \begin{pmatrix} \beta_w \boldsymbol{G} \boldsymbol{\theta}_{(K,g_K)} \\ \boldsymbol{\alpha} \end{pmatrix}, \tau^2 \mathbf{I}_{n_{(K,g_K+1)}} + \sigma_w^2 \tilde{\boldsymbol{X}}_{(K,g_K+1)} \tilde{\boldsymbol{X}}'_{(K,g_K+1)}). \tag{5}$$

A sample from this distribution may be simulated by randomly selecting values of $\boldsymbol{\eta}_{(K,g_K)}$ from among the Gibbs sampler draws, and then drawing $\boldsymbol{y}_{(K,g_K+1)}$ from the distribution in (5) for each draw of $\boldsymbol{\eta}_{(K,g_K)}$. This process may be repeated to construct a sample of desired size. To obtain point predictions, we could calculate the sample average of these posterior predictive draws. It is more efficient, however, to calculate the sample average of the means in (5) across draws of $\boldsymbol{\eta}_{(K,g_K)}$.

4 Posterior Inferences

We use the model described in the preceding section to analyze regular season results of NFL football games for the years 1988–1992, and the first ten weeks of 1993 games. The NFL was comprised of a total of 28 teams during these seasons. During the regular season each team plays a total of 16 games. The 1988–1989 seasons lasted a total of 16 weeks, the 1990–1992 seasons lasted 17 weeks (each team had one off week), and the 1993 season lasted 18 weeks. We use the last 8 weeks of 1993 games to assess the accuracy of predictions from our model. For each game we recorded the final score for each team and the site of the game. Use of covariate information, such as game statistics like rushing yards gained and allowed, might improve the precision of the model fit, but no additional information was recorded.

	Gibbs sampler series						
Parameter	1	2	3	4	5	6	7
ω_w	10.0	100.0	200.0	500.0	1000.0	1000.0	10.0
ω_s	1.0	20.0	80.0	200.0	800.0	1.0	800.0
ω_o	0.5	5.0	15.0	100.0	150.0	150.0	0.5
ω_h	100.0	20.0	6.0	1.0	0.6	0.3	100.0
β_w	0.6	0.8	0.99	1.2	1.8	0.6	1.8
β_s	0.5	0.8	0.98	1.2	1.8	1.8	0.6

Table 1: Starting values for parallel Gibbs samplers

Gibbs sampler implementation:

A single "pilot" Gibbs sampler with starting values at the prior means was run to determine regions of the parameter space with high posterior mass. Seven parallel Gibbs samplers were then run with overdispersed starting values relative to the draws from the pilot sampler. Table 1 displays the starting values chosen for the parameters in the seven parallel runs. Each Gibbs sampler was run for 18000 iterations, and convergence was diagnosed from plots and by examining the potential scale reduction (PSR), as described in Gelman and Rubin (1992), of the parameters ω_w, ω_s, ω_o, ω_h, β_w, and β_s, the home-field advantage parameters, and the most recent team strength parameters. The PSR is an estimate of the factor by which the variance of the current distribution of draws in the Gibbs sampler will decrease with continued iterations. Values near 1 are indicative of convergence. In diagnosing convergence, parameters that were restricted to be positive in the model were transformed by taking logs. Except for the parameter σ_w, all the PSR's were less than 1.2. The slightly larger PSR for σ_w could be explained from the plot of successive draws versus iteration number; the strong autocorrelation in simulations of ω_w slowed the mixing of the different series. We concluded that by iteration 17000 the separate series had essentially converged to the stationary distribution. For each parameter, a sample was obtained by selecting the last 1000 values of the 18000 in each series. This produced the final sample of 7000 draws from the posterior distribution for our analyses.

Parameter summaries:

Tables 2 and 3 show posterior summaries of some model parameters. The means and 95% central posterior intervals for team parameters describe team strengths after the 10th week of the 1993 regular season. The teams are ranked according to their estimated posterior means. The posterior

Parameter	Strength Mean	Strength 95% Posterior Interval	HFA Mean	HFA 95% Posterior Interval
Dallas Cowboys	9.06	(2.26, 16.42)	1.62	(–1.94, 4.86)
San Francisco 49ers	7.43	(0.29, 14.40)	2.77	(–0.76, 6.19)
Buffalo Bills	4.22	(–2.73, 10.90)	4.25	(0.91, 7.73)
New Orleans Saints	3.89	(–3.04, 10.86)	3.44	(–0.01, 6.87)
Pittsburgh Steelers	3.17	(–3.66, 9.96)	3.30	(0.00, 6.68)
Miami Dolphins	2.03	(–4.79, 8.83)	2.69	(–0.81, 6.14)
Green Bay Packers	1.83	(–4.87, 8.66)	2.19	(–1.17, 5.45)
San Diego Chargers	1.75	(–5.02, 8.62)	1.81	(–1.70, 5.12)
New York Giants	1.43	(–5.38, 8.21)	4.03	(0.75, 7.53)
Denver Broncos	1.18	(–5.75, 8.02)	5.27	(1.90, 8.95)
Philadelphia Eagles	1.06	(–5.98, 7.80)	2.70	(–0.75, 6.06)
New York Jets	0.98	(–5.95, 8.00)	1.86	(–1.51, 5.15)
Kansas City Chiefs	0.89	(–5.82, 7.77)	4.13	(0.75, 7.55)
Detroit Lions	0.80	(–5.67, 7.49)	3.12	(–0.31, 6.48)
Houston Oilers	0.72	(–6.18, 7.51)	7.28	(3.79, 11.30)
Minnesota Vikings	0.25	(–6.57, 6.99)	3.34	(–0.01, 6.80)
Los Angeles Raiders	0.25	(–6.43, 7.10)	3.21	(–0.05, 6.55)
Phoenix Cardinals	–0.15	(–6.64, 6.56)	2.67	(–0.69, 5.98)
Cleveland Browns	–0.55	(–7.47, 6.25)	1.53	(–2.04, 4.81)
Chicago Bears	–1.37	(–8.18, 5.37)	3.82	(0.38, 7.27)
Washington Redskins	–1.46	(–8.36, 5.19)	3.73	(0.24, 7.22)
Atlanta Falcons	–2.94	(–9.89, 3.85)	2.85	(–0.55, 6.23)
Seattle Seahawks	–3.17	(–9.61, 3.43)	2.21	(–1.25, 5.52)
Los Angeles Rams	–3.33	(–10.18, 3.37)	1.85	(–1.61, 5.23)
Indianapolis Colts	–5.29	(–12.11, 1.63)	2.45	(–0.97, 5.81)
Tampa Bay Buccaneers	–7.43	(–14.38, –0.68)	1.77	(–1.69, 5.13)
Cincinnati Bengals	–7.51	(–14.74, –0.68)	4.82	(1.53, 8.33)
New England Patriots	–7.73	(–14.54, –0.87)	3.94	(0.55, 7.34)

Table 2: Summaries of the posterior distributions of team strength and home-field advantage parameters after the first 10 weeks of the 1993 regular season.

Parameter	Mean	95% Posterior Interval
τ	12.78	(12.23, 13.35)
σ_o	3.26	(1.87, 5.22)
σ_w	0.88	(0.52, 1.36)
σ_s	2.35	(1.14, 3.87)
σ_h	2.28	(1.48, 3.35)
β_w	0.99	(0.96, 1.02)
β_s	0.82	(0.52, 1.28)

Table 3: Summaries of the posterior distributions of standard deviations and regression parameters after the first 10 weeks of the 1993 regular season.

means range from 9.06 (Dallas Cowboys) to –7.73 (New England Patriots) which suggests that on a neutral field, the best team has close to a 17 point advantage over the worst team. The 95% intervals clearly indicate that a considerable amount of variability is associated with the team strength parameters which may be due to the stochastic nature of team strengths. The distribution of home field advantages varies from roughly 1.6 points (Dallas Cowboys, Cleveland Browns) to over 7 points (Houston Oilers). The 7-point home-field advantage conveyed to the Oilers is substantiated by the numerous "blowouts" they have had on their home field. The home-field advantage parameters are centered around 3.2. This value is consistent with the results of previous modeling (Harville 1980; Sallas and Harville 1988; Glickman 1993).

The distribution of the standard deviation parameters τ, σ_o, σ_h, σ_w and σ_s are shown in Figures 1 and 2. The plots show that each of the standard deviations is approximately symmetrically distributed around its mean. The posterior distribution of τ is centered just under 13 points, indicating that the score difference for a single game conditional on team strengths can be expected to vary by about $4\tau \approx 50$ points. The posterior distribution of σ_o shown in Figure 1 suggests that the normal distribution of teams abilities prior to 1988 have a standard deviation somewhere between 2 and 5, so that the a priori difference between the best and worst teams is near $4\sigma_o \approx 15$. This range of team strength appears to persist in 1993, as can be calculated from Table 2. The distribution of σ_h is centered near 2.3, suggesting that teams' home-field advantages varied moderately around a mean of 3 points.

As shown in the empirical contour plot in Figure 2, the posterior distribution of the between-week standard deviation, σ_w, is concentrated on smaller values and less dispersed than that of the between-season evolution standard deviation, σ_s. This difference in magnitude indicates that the types of changes that occur between weeks are likely to have less impact on a team's ability than the changes that occur between seasons. The distribution for the between-week standard deviation is less dispersed than that for the between-season standard deviation because the data provide much more information about weekly innovations than about changes between seasons. Furthermore, the contour plot shows a slight negative posterior correlation between the standard deviations. This is not terribly surprising if we consider that the total variability due to the passage of time over an entire season is the composition of between-week variability and between-season variability. If between-week variability is small, then between-season variability must be large to compensate. An interesting feature revealed by the contour plot is the apparent bimodality of the joint distribution. This feature was not apparent from examining the marginal distributions. Two modes of (σ_w, σ_s) appear at (0.6, 3) and (0.9, 2).

Figure 3 shows contours of the bivariate posterior distribution of the parameters β_w and β_s. The contours of the plot display a concentration of mass near $\beta_s \approx 0.8$ and $\beta_w \approx 1.0$, as is also indicated in Table 3. The plot shows a more marked negative posterior correlation between these two parameters than between the standard deviations. The negative correlation can be explained in an analogous manner to the negative correlation between standard deviations, viewing the total shrinkage over the season as being the composition of the between-week shrinkages and the between-season shrinkage. As with the standard deviations, the data provide more precision about the between-week regression parameter than the between-season regression parameter.

Prediction for week 11:

Predictive summaries for the week 11 games of the 1993 NFL season are shown in Table 4. The point predictions were computed as the average of the mean outcomes across all 7000 posterior draws. Intervals were constructed empirically by simulating single game outcomes from the predictive distribution for each of the 7000 Gibbs samples. Of the 13 games, six of the actual score difference were contained in the 50% prediction intervals. All of the widths of the intervals were close to 18–19 points. Our point predictions were generally close to the Las Vegas line. Games where predictions differ substantially (e.g., Oilers at Bengals) may reflect information from the previous week that our model does not incorporate, such as injuries of important players.

Predictions for weeks 12 through 18:

Once game results for a new week were available, a single-series Gibbs sampler was run using the entire data set to obtain a new set of parameter draws. The starting values for the series were the posterior mean estimates of ω_w, ω_s, ω_o, ω_h, β_w, and β_s, from the end of week 10. Because the

Week 11 Games			Predicted Score Difference	Las Vegas Line	Actual Score Difference	Central 50% Prediction Interval
Packers	at	Saints	–5.49	–6.0	2	(–14.75, 3.92)
Oilers	at	Bengals	3.35	8.5	35	(–6.05, 12.59)
Cardinals	at	Cowboys	–10.77	–12.5	–5	(–20.25, –1.50)
49ers	at	Buccaneers	13.01	16.0	24	(3.95, 22.58)
Dolphins	at	Eagles	–1.74	4.0	5	(–10.93, 7.55)
Redskins	at	Giants	–6.90	–7.5	–14	(–16.05, 2.35)
Chiefs	at	Raiders	–2.57	–3.5	11	(–11.63, 6.86)
Falcons	at	Rams	–1.46	–3.5	13	(–10.83, 7.93)
Browns	at	Seahawks	0.40	–3.5	–17	(–8.98, 9.72)
Vikings	at	Broncos	–6.18	–7.0	3	(–15.42, 3.25)
Jets	at	Colts	3.78	3.5	14	(–5.42, 13.19)
Bears	at	Chargers	–4.92	–8.5	3	(–14.31, 4.28)
Bills	at	Steelers	–2.26	–3.0	–23	(–11.53, 7.13)

Table 4: Forecasts for NFL games during week 11 of the 1993 regular season.

posterior variability of these parameters is small, the addition of a new week's collection of game outcomes is not likely to have substantial impact on posterior inferences. Thus, our procedure takes advantage of knowing regions of the parameter space *a priori* that will have high posterior mass. Having obtained data from the results of week 11, we ran a single-series Gibbs sampler for 5000 iterations, saving the last 1000 for predictive inferences. This procedure was repeated for weeks 12 through 17 in an analogous manner. Point predictions were computed as described earlier. In practice, the model could be refit periodically using a multiple chain procedure as an alternative to using this one-chain updating algorithm. This might be advantageous in reassessing the propriety of the model or determining if significant shifts in parameter values have occurred.

Comparison with Las Vegas betting line:

We compared the accuracy of our predictions with those of the Las Vegas point spread on the 110 games beyond the 10th week of the 1993 season. The mean squared error for predictions from our model for these 110 games was 165.0. This is slightly better than the mean squared error of 170.5 for the point spread. Similarly, the mean absolute error from our model is 10.50, while the analogous result for the point spread is 10.84. Our model correctly predicted the winners of 64 of the 110 games (58.2%) while the Las Vegas line predicted 63. Out of the 110 predictions from our model, 65 produced mean score differences that "beat the point spread," that is, resulted in predictions that were greater than the point spread when the actual score difference was larger than the point spread, or resulted in predictions that were lower than the point spread when the actual score difference was lower. For this small sample, the model fit outperforms the point spread, though the difference is not large enough to generalize. However, the results here suggest that the state-space model yields predictions that are comparable to those implied by the betting line.

5 Diagnostics

Model validation and diagnosis is an important part of the model fitting process. In complex models, however, diagnosing invalid assumptions or lack of fit often cannot be carried out using conventional methods. In this section, we examine several model assumptions through the use of posterior predictive diagnostics. We include a brief description of the idea behind posterior predictive diagnostics. We also describe how model diagnostics were able to suggest an improvement to an earlier version of the model.

The approach to model checking using posterior predictive diagnostics is discussed in detail in Gelman et al. (1996), and the foundations of this approach are described in Rubin (1984). The strategy is to construct discrepancy measures that address particular aspects of the data that one suspects may not be captured by the model. Discrepancies may be ordinary test statistics or they may depend on both data values and parameters. The discrepancies are computed using the actual data and the resulting values are compared to the reference distribution obtained using simulated data from the posterior predictive distribution. If the actual data is "typical" of the draws from the posterior predictive distribution under the model, then the posterior distribution of the discrepancy measure evaluated at the actual data will be similar to the posterior distribution of the discrepancy evaluated at the simulated data sets. Otherwise, the discrepancy measure provides some indication that the model may be misspecified.

To be concrete, we may construct a "generalized" test statistic, or discrepancy, $D(\boldsymbol{y};\boldsymbol{\theta})$, which may be a function of not only the observed data, generically denoted as $\boldsymbol{y}$, but also model parameters, generically denoted as $\boldsymbol{\theta}$. We compare the posterior distribution of $D(\boldsymbol{y};\boldsymbol{\theta})$ to the posterior predictive distribution of $D(\boldsymbol{y}^*;\boldsymbol{\theta})$, where we use $\boldsymbol{y}^*$ to denote hypothetical replicate data generated under the model with the same (unknown) parameter values. One possible summary of the evaluation is the tail probability, or p-value, computed as

$$\Pr(D(\boldsymbol{y}^*;\boldsymbol{\theta}) \geq D(\boldsymbol{y};\boldsymbol{\theta}) \mid \boldsymbol{y}),$$

or

$$\Pr(D(\boldsymbol{y}^*;\boldsymbol{\theta}) \leq D(\boldsymbol{y};\boldsymbol{\theta}) \mid \boldsymbol{y}),$$

depending on the definition of the discrepancy. In practice, the relevant distributions or the tail probability can be approximated through Monte Carlo integration by drawing samples from the posterior distribution of $\boldsymbol{\theta}$ and then the posterior predictive distribution of $\boldsymbol{y}^*$ given $\boldsymbol{\theta}$.

The choice of suitable discrepancy measures, D, depends on the problem. We try to define measures that evaluate the fit of the model to features of the data that are not explicitly accounted for in the model specification. Here we consider diagnostics that assess the homogeneity of variance assumptions in the model and diagnostics that assess assumptions concerning the home-field advantage. The home-field advantage diagnostics were useful in detecting a failure of an earlier version of the model. Our summary measures D are functions of Bayesian residuals as defined by Zellner (1975) and Chaloner and Brant (1988). As an alternative to focusing on summary measures D, the individual Bayesian residuals can be used to search for outliers or to construct a "distribution" of residual plots; we do not pursue this approach here.

Regression Variance:

For a particular game played between teams i and i' at team i's home field, let

$$e_{ii'}^2 = (y_{ii'} - (\theta_i - \theta_{i'} + \alpha_i))^2$$

be the squared difference between the observed outcome and the expected outcome under the model, which might be called a squared residual. Averages of $e_{ii'}^2$ across games can be interpreted as estimates of τ^2 (the variance of $y_{ii'}$ given its mean). The model assumes τ^2 is constant across seasons and for all competing teams. We consider two discrepancy measures that are sensitive to failures of these assumptions. Let $D_1(\boldsymbol{y};\boldsymbol{\theta})$ be the difference between the largest of the six annual average squared residuals and the smallest of the six annual average squared residuals. Then $D_1(\boldsymbol{y}^*;\boldsymbol{\theta}^*)$ is the value of this diagnostic evaluated at simulated parameters $\boldsymbol{\theta}^*$ and simulated data $\boldsymbol{y}^*$, and $D_1(\boldsymbol{y};\boldsymbol{\theta}^*)$ is the value evaluated at the same simulated parameters but using the actual data. Based on 300 samples of parameters from the posterior distribution and simulated data from the posterior predictive distribution, the approximate bivariate posterior distribution of $(D_1(\boldsymbol{y};\boldsymbol{\theta}^*),\ D_1(\boldsymbol{y}^*;\boldsymbol{\theta}^*))$ is shown in the top panel of Figure 4. The plot shows that large portions of the distribution of the discrepancies lie both above and below the line $D_1(\boldsymbol{y};\boldsymbol{\theta}^*) = D_1(\boldsymbol{y}^*;\boldsymbol{\theta}^*)$ with the relevant tail probability equal to 0.35. This suggests that the year-to-year variation in the regression variance of the actual data is quite consistent with that expected under the model (as evidenced by the simulated data sets).

As a second discrepancy measure we can compute the average $e_{ii'}^2$ for each team and then calculate the difference between the maximum of the 28 team-specific estimates and the minimum of the 28 team-specific estimates. Let $D_2(\boldsymbol{y}^*; \boldsymbol{\theta}^*)$ be the value of this diagnostic measure for the simulated data $\boldsymbol{y}^*$ and simulated parameters $\boldsymbol{\theta}^*$, and let $D_2(\boldsymbol{y}; \boldsymbol{\theta}^*)$ be the value for the actual data and simulated parameters. The approximate posterior distribution of $(D_2(\boldsymbol{y}; \boldsymbol{\theta}^*), D_2(\boldsymbol{y}^*; \boldsymbol{\theta}^*))$ based on the same 300 samples of parameters and simulated data are shown in the bottom panel of Figure 4. The value of D_2 based on the actual data tends to be larger than the value based on the posterior predictive simulations. The relevant tail probability $P(D_2(\boldsymbol{y}^*; \boldsymbol{\theta}) > D_2(\boldsymbol{y}; \boldsymbol{\theta}) | \boldsymbol{y})$ is not terribly small (0.14) so we conclude that there is no evidence of heterogeneous regression variances for different teams. We would not likely be interested, therefore, in extending our model in the direction of a non-constant regression variance.

Site-effect: a model diagnostics success story

We can use a slightly modified version of the game residuals

$$r_{ii'} = y_{ii'} - (\theta_i - \theta_{i'})$$

to search for a failure of the model in accounting for home field advantage. The $r_{ii'}$ are termed site-effect residuals since they take the observed outcome and subtract out estimated team strengths but do not subtract out the home-field advantage. As we did with the regression variance, we can examine differences in the magnitude of the home field advantage over time by calculating the average value of the $r_{ii'}$ for each season, and then examining the range of these averages. Specifically, for a posterior predictive data set $\boldsymbol{y}^*$ and for a draw $\boldsymbol{\theta}^*$ from the posterior distribution of all parameters, let $D_3(\boldsymbol{y}^*; \boldsymbol{\theta}^*)$ be the difference between the maximum and the minimum of the average site-effect residuals by season. Using the same 300 values of $\boldsymbol{\theta}^*$ and $\boldsymbol{y}^*$ as before, we obtain the estimated bivariate distribution of $(D_3(\boldsymbol{y}; \boldsymbol{\theta}^*), D_3(\boldsymbol{y}^*; \boldsymbol{\theta}^*))$ shown at the top of Figure 5. The plot reveals no particular pattern although there is a tendency for $D_3(\boldsymbol{y}; \boldsymbol{\theta}^*)$ to be less than the discrepancy evaluated at the simulated data sets. This seems to be a chance occurrence (the tail probability equals 0.21).

We also include one other discrepancy measure although it will be evident that our model fits this particular aspect of the data. We examined the average site-effect residuals across teams to assess whether the site effect depends on team. We calculated the average value of $r_{ii'}$ for each team. Let $D_4(\boldsymbol{y}^*; \boldsymbol{\theta}^*)$ be the difference between the maximum and minimum of these 28 averages for simulated data $\boldsymbol{y}^*$. It should be evident that the model will fit this aspect of the data because we have used a separate parameter for each team's advantage. The approximate bivariate distribution of $(D_4(\boldsymbol{y}; \boldsymbol{\theta}^*),\ D_4(\boldsymbol{y}^*; \boldsymbol{\theta}^*))$ is shown at the bottom of Figure 5. There is no evidence of lack of fit (tail probability equals 0.32).

This last discrepancy measure is included here, despite the fact that it measures a feature of the data that we have explicitly addressed in the model, because the current model was not the first model that we constructed. We earlier fit a model with a single home-field advantage parameter for all teams. Figure 6 shows that for the single home-field advantage parameter model the observed values of $D_4(\boldsymbol{y}; \boldsymbol{\theta}^*)$ were generally greater than the values of $D_4(\boldsymbol{y}^*; \boldsymbol{\theta}^*)$, indicating that the average site-effect residuals varied significantly more from team to team than was expected under the model (tail probability equal to 0.05). This suggested the model presented here in which each team has a separate home-field advantage parameter.

Sensitivity to heavy tails:

Our model assumes that outcomes are normally distributed conditional on the parameters, an assumption that is supported by Stern (1991). Rerunning the model with t-distributions in place of normal distributions is straightforward since t-distributions can be expressed as scale mixtures of normal distributions (see, e.g., Gelman et al. (1995) and Smith (1983)). Rather than redo the entire analysis, we checked the sensitivity of our inferences to the normal assumption by reweighting the posterior draws from the Gibbs sampler by ratios of importance weights (relating the normal model to a variety of t models). The reweighting is easily done and provides information about how inferences would be likely to change under alternative models. Our conclusion is that using a robust alternative can slightly alter estimates of team strength but does not have a significant effect on the predictive performance. It should be emphasized that the ratios of importance weights can be unstable so that a more definitive discussion of inference under a particular t-model (e.g., 4 degrees of freedom) would require a complete re-analysis of the data.

6 Conclusions

Our model for football game outcomes assumes that team strengths can change over time in a manner described by a normal state-space model. In previous state-space modeling of football scores (Harville 1977, 1980; Sallas and Harville 1988), some model parameters were estimated and then treated as fixed in making inferences on the remaining parameters. Such an approach ignores the variability associated with these parameters. The approach taken here, in contrast, is fully Bayesian in that we account for the uncertainty in all model parameters when making posterior or predictive inferences.

Our data analysis suggests that the model can be improved in several different dimensions. One could argue that teams abilities should not shrink or expand around the mean from week to week, and because the posterior distribution of the between-week regression parameter, β_w, is not substantially different from 1, the model may be simplified by setting it to 1. Also, further exploration may be necessary to assess the assumption of a heavy-tailed distribution for game outcomes. Finally, as the game of football continues to change over time it may be necessary to allow the evolution regression and variance parameters or the regression variance parameter to vary over time.

Despite the room for improvement, we feel that our model captures the main components of variability in football game outcomes. Recent advances in Bayesian computational methods allow us to fit a realistic complex model and diagnose model assumptions that would otherwise be difficult to carry out. Predictions from our model seem to perform as well, on average, as the Las Vegas point spread, so our model does appear to track team strengths in a manner similar to that of the best expert opinion.

A Conditional distributions for MCMC sampling

Conditional posterior distribution of $(\boldsymbol{\theta}_{(1,1)}, \ldots, \boldsymbol{\theta}_{(K,g_K)}, \boldsymbol{\alpha}, \phi)$

The conditional posterior distribution of the team strength parameters, HFA parameters, and observation precision ϕ, is normal-gamma – the conditional posterior distribution of ϕ given the evolution precision and regression parameters $(\omega_o, \omega_h, \omega_w, \omega_s, \beta_s, \beta_w)$ is gamma, and the conditional posterior distribution of the team strengths and home-field parameters given all other parameters is an $(M+1)p$-variate normal distribution, where p is the number of teams and $M = \sum_{k=1}^{K} g_k$ is the total number of weeks for which data is available. It is advantageous to sample using results from the Kalman filter (Glickman 1993; Carter and Kohn 1994; Fruhwirth-Schnatter 1994) rather than consider this $(M+1)p$-variate conditional normal distribution as a single distribution. This idea is summarized below.

The Kalman filter (Kalman 1961; Kalman and Bucy 1961) is used to compute the normal-gamma posterior distribution of the final week's parameters

$$f(\boldsymbol{\theta}_{(K,g_K)}, \boldsymbol{\alpha}, \phi | \omega_o, \omega_h, \omega_w, \omega_s, \beta_s, \beta_w, \boldsymbol{Y}_{(K,g_K)})$$

marginalizing over the previous weeks' vectors of team strength parameters. This distribution is obtained by a sequence of recursive computations that alternately update the distribution of parameters when new data is observed, and then update the distribution reflecting the passage of time. A sample from this posterior distribution is drawn. Samples of team strengths for previous weeks are drawn by employing a back-filtering algorithm. This is accomplished by drawing recursively from the normal distributions for the parameters from earlier weeks

$$f(\boldsymbol{\theta}_{(k,j)} | \boldsymbol{\theta}_{(k,j+1)}, \ldots, \boldsymbol{\theta}_{(K,g_K)}, \boldsymbol{\alpha}, \phi, \omega_o, \omega_h, \omega_w, \omega_s, \beta_s, \beta_w, \boldsymbol{Y}_{(K,g_K)}).$$

The result of this procedure is a sample of values from the desired conditional posterior distribution.

Conditional posterior distribution of $(\omega_o, \omega_h, \omega_w, \omega_s)$

Conditional on the remaining parameters and the data, the parameters ω_o, ω_h, ω_w and ω_s are independent gamma random variables with

$$\begin{aligned} &\omega_o \,|\, \boldsymbol{\theta}_{(1,1)}, \ldots, \boldsymbol{\theta}_{(K,g_K)}, \boldsymbol{\alpha}, \phi, \beta_s, \beta_w, \boldsymbol{Y}_{(K,g_K)} \;\sim \\ &\qquad \text{Gamma}(1 + \frac{(p+1)}{2}, \frac{1}{2}(\frac{1}{6} + \phi \boldsymbol{\theta}'_{(1,1)} \boldsymbol{\theta}_{(1,1)})), \end{aligned}$$

$$\begin{aligned} &\omega_h \,|\, \boldsymbol{\theta}_{(1,1)}, \ldots, \boldsymbol{\theta}_{(K,g_K)}, \boldsymbol{\alpha}, \phi, \beta_s, \beta_w, \boldsymbol{Y}_{(K,g_K)} \;\sim \\ &\qquad \text{Gamma}(1 + \frac{(p+1)}{2}, \frac{1}{2}(\frac{1}{6} + \phi(\boldsymbol{\alpha} - 3)'(\boldsymbol{\alpha} - 3))), \end{aligned}$$

$$\begin{aligned} &\omega_w \,|\, \boldsymbol{\theta}_{(1,1)}, \ldots, \boldsymbol{\theta}_{(K,g_K)}, \boldsymbol{\alpha}, \phi, \beta_s, \beta_w, \boldsymbol{Y}_{(K,g_K)} \;\sim \\ &\qquad \text{Gamma}(2 + \frac{p \sum_{k=1}^{K} (g_k - 1)}{2}, \frac{1}{2}(\frac{1}{60} + \phi \sum_{k=1}^{K} \sum_{j=1}^{g_k - 1} (\boldsymbol{\theta}_{(k,j+1)} - \beta_w \boldsymbol{G} \boldsymbol{\theta}_{(k,j)})'(\boldsymbol{\theta}_{(k,j+1)} - \beta_w \boldsymbol{G} \boldsymbol{\theta}_{(k,j)}))), \end{aligned}$$

and

$$\begin{aligned} &\omega_s \,|\, \boldsymbol{\theta}_{(1,1)}, \ldots, \boldsymbol{\theta}_{(K,g_K)}, \boldsymbol{\alpha}, \phi, \beta_s, \beta_w, \boldsymbol{Y}_{(K,g_K)}) \;\sim \\ &\qquad \text{Gamma}(1 + \frac{(1 + p(K-1))}{2}, \frac{1}{2}(\frac{1}{16} + \phi \sum_{k=1}^{K-1} (\boldsymbol{\theta}_{(k+1,1)} - \beta_s \boldsymbol{G} \boldsymbol{\theta}_{(k,g_k)})'(\boldsymbol{\theta}_{(k+1,1)} - \beta_s \boldsymbol{G} \boldsymbol{\theta}_{(k,g_k)}))). \end{aligned}$$

Conditional posterior distribution of (β_s, β_w)

Conditional on the remaining parameters and the data, the distribution of β_w and β_s are independent random variables with normal distributions. The distribution of β_w conditional on all other parameters is normal, with

$$\beta_w \,|\, \omega_o, \omega_h, \omega_w, \omega_s, \boldsymbol{\theta}_{(1,1)}, \ldots, \boldsymbol{\theta}_{(K,g_K)}, \boldsymbol{\alpha}, \phi, \boldsymbol{Y}_{(K,g_K)} \;\sim\; \mathrm{N}(M_w, V_w)$$

where

$$\begin{aligned}
V_w &= (1 + \phi\omega_w A)^{-1} \\
M_w &= V_w(0.995 + \phi\omega_w B) \\
A &= \sum_{k=1}^{K} \sum_{j=1}^{g_k-1} \boldsymbol{\theta}'_{(k,j)} \boldsymbol{G} \boldsymbol{\theta}_{(k,j)} \\
B &= \sum_{k=1}^{K} \sum_{j=1}^{g_k-1} \boldsymbol{\theta}'_{(k,j+1)} \boldsymbol{G} \boldsymbol{\theta}_{(k,j)}.
\end{aligned}$$

The distribution of β_s conditional on all other parameters is also normal, with

$$\beta_s \,|\, \omega_o, \omega_h, \omega_w, \omega_s, \boldsymbol{\theta}_{(1,1)}, \ldots, \boldsymbol{\theta}_{(K,g_K)}, \boldsymbol{\alpha}, \phi, \boldsymbol{Y}_{(K,g_K)} \;\sim\; \mathrm{N}(M_s, V_s)$$

where

$$\begin{aligned}
V_s &= (1 + \phi\omega_w C)^{-1} \\
M_s &= V_s(0.98 + \phi\omega_w D) \\
C &= \sum_{k=1}^{K-1} \boldsymbol{\theta}'_{(k,g_k)} \boldsymbol{G} \boldsymbol{\theta}_{(k,g_k)} \\
D &= \sum_{k=1}^{K-1} \boldsymbol{\theta}'_{(k+1,1)} \boldsymbol{G} \boldsymbol{\theta}_{(k,g_k)}.
\end{aligned}$$

Bibliography

Amoako-Adu, B., Marmer, H., and Yagil, J.(1985), "The efficiency of certain speculative markets and gambler behavior," *Journal of Economics and Business*, 37, 365–378.

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

Carter, C. K., and Kohn, R.(1994), "On Gibbs sampling for state space models," *Biometrika*, 81, 541–53.

Chaloner, K. and Brant, R. (1988), "A Bayesian approach to outlier detection and residual analysis," *Biometrika*, 75, 651–659.

de Jong, P. and Shephard, N.(1995), "The simulation smoother for time series models," *Biometrika*, 82, 339–350.

Fruhwirth-Schnatter, S.(1994), "Data augmentation and dynamic linear models," *Journal of Time Series Analysis*, 15, 183-202.

Gelfand, A. E. and Smith, A. F. M.(1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85, 972–985.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B.(1995), *Bayesian Data Analysis*, London: Chapman and Hall.

Gelman, A., Meng, X., and Stern, H. S.(1996), "Posterior predictive assessment of model fitness via realized discrepancies (with discussion)," *Statistica Sinica*, 6, 733–807.

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

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

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

Harrison, P. J. and Stevens, C. F.(1976), "Bayesian Forecasting," *Journal of the Royal Statistical Society, Series B*, 38, 240–247.

Harville, D.(1977), "The use of linear model methodology to rate high school or college football teams," *Journal of the American Statistical Association*, 72, 278–289.

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

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

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

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

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

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

Shephard, N.(1994), "Partial non-Gaussian state space," *Biometrika*, 81, 115–31.

Smith, A. F. M. (1983), "Bayesian approaches to outliers and robustness," in *Specifying Statistical Models from Parametric to Nonparametric, using Bayesian or Non-Bayesian approaches*, eds. J. P. Florens, M. Mouchart, J. P. Raoult, L. Simer, A. F. M. Smith (Lecture Notes in Statistics

16), New York: Springer-Verlag, 13–35.

Stern, H.(1991), "On the probability of winning a football game," *The American Statistician*, 45, 179–183.

Stern, H.(1992), "Who's number one? – Rating football teams," *Proceedings of the Section on Statistics in Sports of the American Statistical Association*, 1–6.

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

Stefani, R. T.(1980), "Improved least squares football, basketball, and soccer predictions," *IEEE Transactions on Systems, Man, and Cybernetics*, 10, 116–123.

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

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

Zellner, A. (1975), "Bayesian analysis of regression error terms," *Journal of the American Statistical Association*, 70, 138–144.

Zuber, R. A., Gandar, J. M., and Bowers, B. D.(1985), "Beating the spread: Testing the efficiency of the gambling market for National Football League games," *Journal of Political Economy*, 93, 800–806.

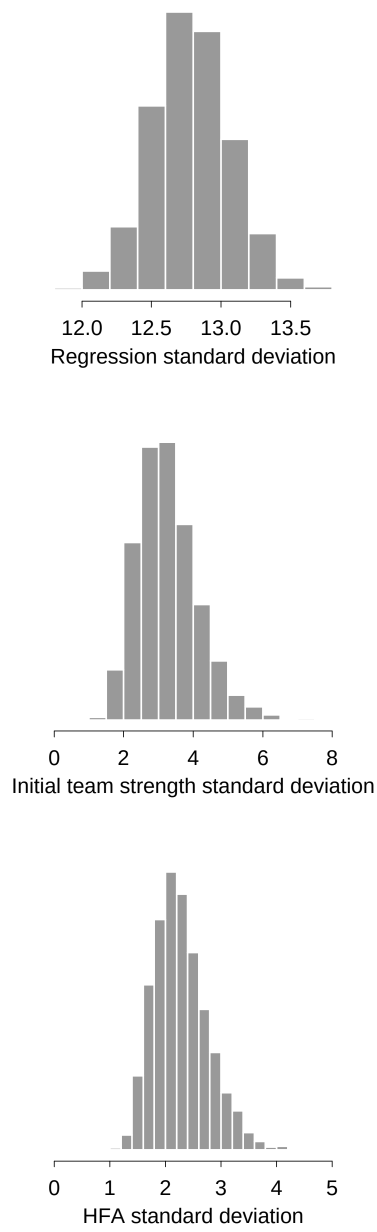


Figure 1: Estimated Posterior Distributions of the Regression Standard Deviation (τ), the Initial Team Strength Standard Deviation (σ_o) and the HFA Standard Deviation (σ_h).

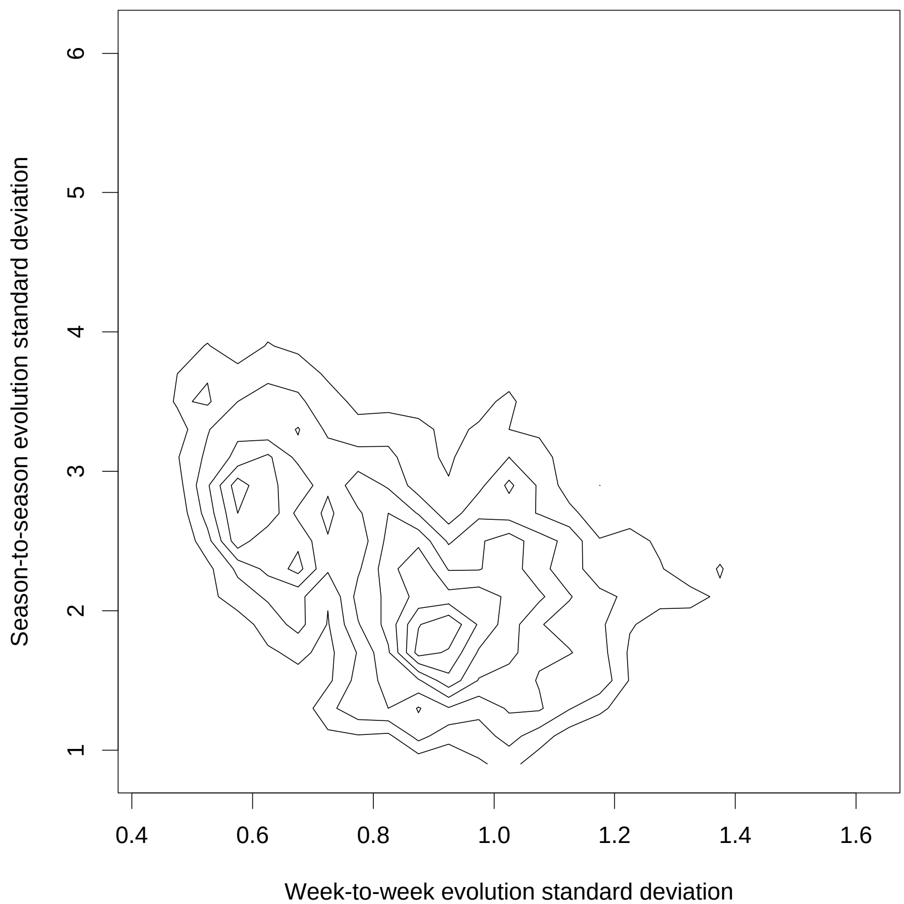


Figure 2: Estimated Joint Posterior Distribution of the Week-to-Week Evolution Standard Deviation (σ_w) and the Season-to-Season Evolution Standard Deviation (σ_s).

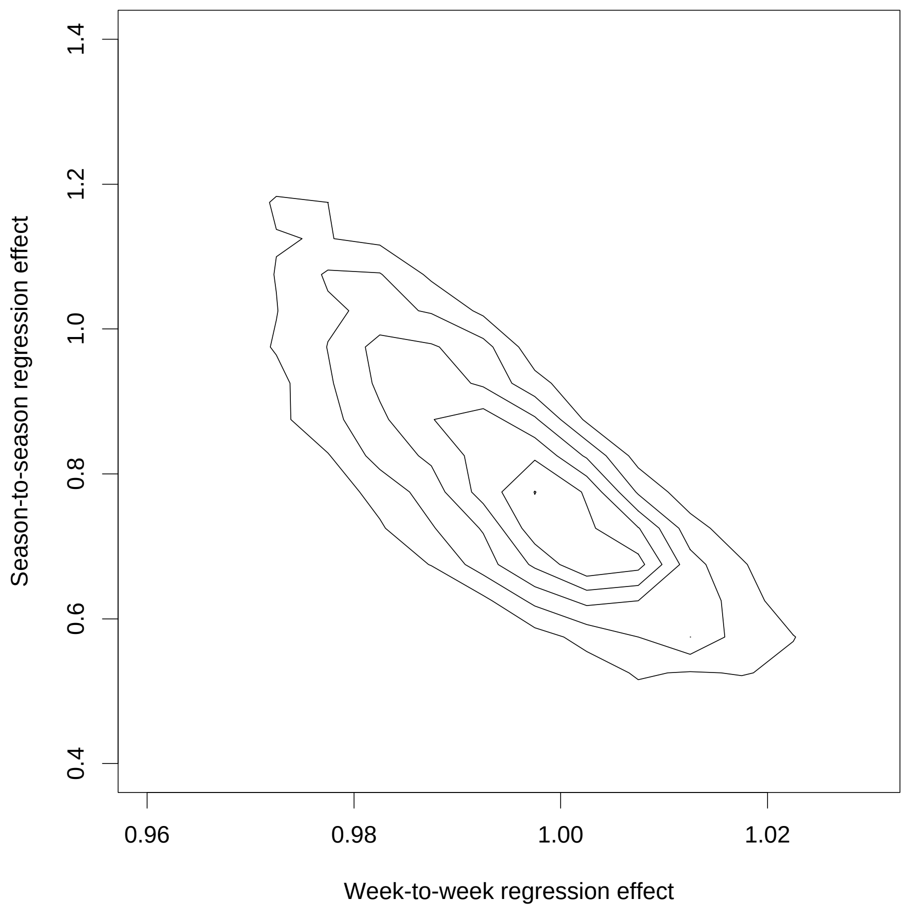
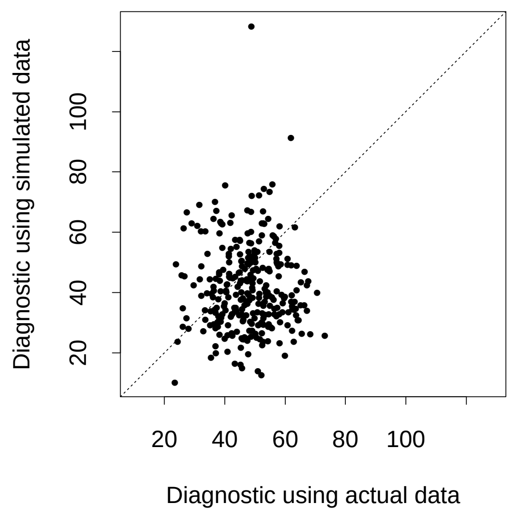


Figure 3: Estimated Joint Posterior Distribution of the Week-to-Week Regression Effect (β_w) and the Season-to-Season Regression Effect (β_s).



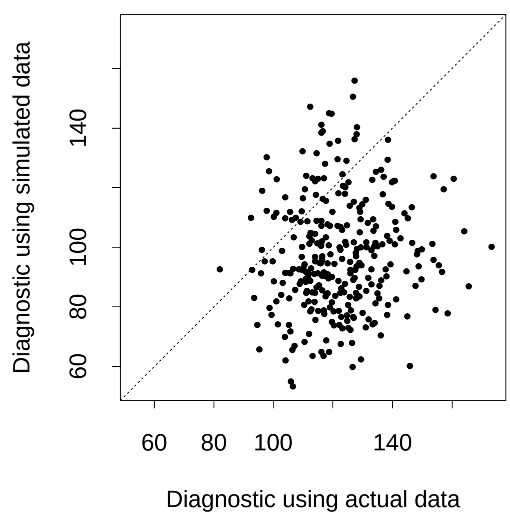


Figure 4: Estimated Bivariate Distributions for Regression Variance Diagnostics. The top panel is a scatterplot of the joint posterior distribution of $D_1(\boldsymbol{y};\boldsymbol{\theta}^*)$ and $D_1(\boldsymbol{y}^*;\boldsymbol{\theta}^*)$ where $D_1(\cdot;\cdot)$ is a discrepancy measuring the range of regression variance estimates among the six seasons and $(\boldsymbol{\theta}^*,\boldsymbol{y}^*)$ are simulations from their posterior and posterior predictive distributions. The bottom panel is a scatterplot of the joint posterior distribution of $D_2(\boldsymbol{y};\boldsymbol{\theta}^*)$ and $D_2(\boldsymbol{y}^*;\boldsymbol{\theta}^*)$ for $D_2(\cdot;\cdot)$ a discrepancy measuring the range of regression variance estimates among the 28 teams.

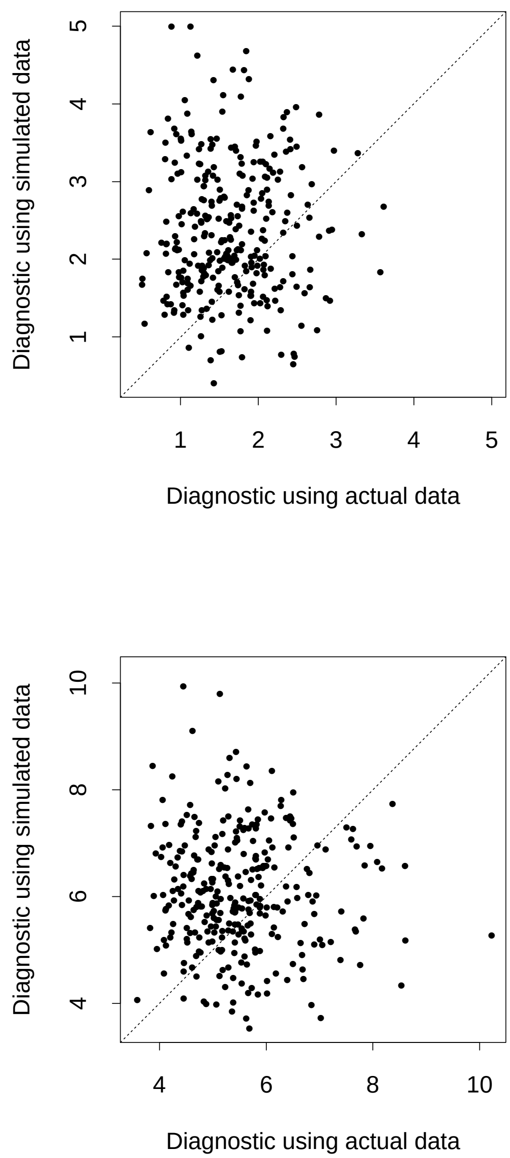


Figure 5: Estimated Bivariate Distributions for Site-Effect Diagnostics. The top panel is a scatterplot of the joint posterior distribution of $D_3(\boldsymbol{y};\boldsymbol{\theta}^*)$ and $D_3(\boldsymbol{y}^*;\boldsymbol{\theta}^*)$ where $D_3(\cdot;\cdot)$ is a discrepancy measuring the range of home-field advantage estimates among the six seasons and $(\boldsymbol{\theta}^*, \boldsymbol{y}^*)$ are simulations from their posterior and posterior predictive distributions. The bottom panel is a scatterplot of the joint posterior distribution of $D_4(\boldsymbol{y};\boldsymbol{\theta}^*)$ and $D_4(\boldsymbol{y}^*;\boldsymbol{\theta}^*)$ for $D_4(\cdot;\cdot)$ a discrepancy measuring the range of home-field advantage estimates among the 28 teams.

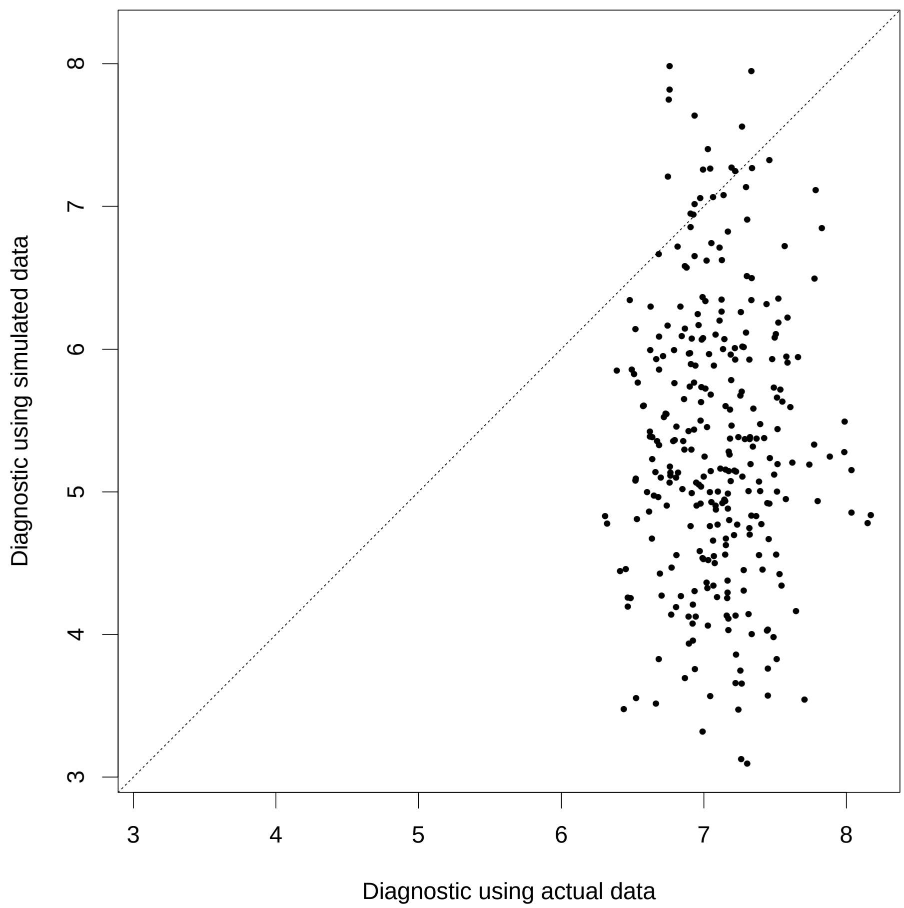


Figure 6: Estimated Bivariate Distribution for Site-Effect Diagnostic from a Poor-Fitting Model. The scatterplot shows the joint posterior distribution of $D_4(\boldsymbol{y};\boldsymbol{\theta}^*)$ and $D_4(\boldsymbol{y}^*;\boldsymbol{\theta}^*)$ for a model that includes only a single parameter for the site-effect rather than 28 separate parameters, one for each team. The values of $D_4(\boldsymbol{y};\boldsymbol{\theta}^*)$ are generally larger than the values of $D_4(\boldsymbol{y}^*;\boldsymbol{\theta}^*)$, suggesting that the fitted model may not be capturing a source of variability in the observed data.