Journal of Quantitative Analysis in Sports

Volume 1, *Issue* 1 2005 *Article* 3

Hybrid Paired Comparison Analysis, with Applications to the Ranking of College Football Teams

David H. Annis* Bruce A. Craig†

*Naval Postgraduate School, annis@nps.edu

†Purdue University, bacraig@stat.purdue.edu

Hybrid Paired Comparison Analysis, with Applications to the Ranking of College Football Teams

David H. Annis and Bruce A. Craig

Abstract

Existing paired comparison models used for ranking football teams primarily focus on either wins and losses or points scored (either via each team's total or a margin of victory). While reasonable, each approach fails to produce satisfactory rankings in frequently arising situations due to its ignorance of additional data. We propose a new, hybrid model incorporating both wins and constituent scores and show that it outperforms its competitors and is robust against model mis-specification based on a series of simulation studies. We conclude by illustrating the method using the 2003-04 and 2004-05 college football seasons.

KEYWORDS: Ranking, Football, Paired Comparison

1 Introduction

Football is the only Division I-A varsity sport which does not crown its national champion in a post-season tournament. Rather, this honor is determined by end-of-season rankings according to two well-accepted polls[1] (though it should be noted that the coaches' first-place vote is contractually mandated). While they never agree exactly, in most years, there is a consensus as to which team has demonstrated its superiority over the course of the season. However, between 1950 and 1997, these parties failed to agree on even the best team ten times, culminating in 1997, when the University of Michigan and the University of Nebraska shared the championship.

The perceived failure of the polling system led to the creation of the Bowl Championship Series (BCS) in 1998. By including computer polls and objective measures of performance, the BCS was designed to lessen dependence on popular opinion and determine the two most deserving participants for an end-of-season championship game in a more "objective" manner. A thorough explanation of the BCS system and its shortcomings is given by Stern [22] and discussed extensively by Billingsley [2], Colley [5], Harville [11], Massey [16] and Mease [19]. Perhaps the most common (and scathing) criticism of the current system is that while it considers many factors – among them losses (but not wins), schedule strength[2] and the polls – it does so in an *ad hoc* fashion and fails to optimize any objective measure. Although some feel that this system represents an improvement over the traditional bowl system, it has seen its share of controversy.

For example, in 2001 the Pac-10 champion Oregon Ducks, ranked second in both media polls, were not selected to play in the title game. The prevailing sentiment was that Oregon was judged fairly by the informed media, but hurt by computer polls which considered scoring and margin of victory, as Oregon had many narrow victories. This "problem" was subsequently "rectified," allowing an unbeaten Ohio State team into the Championship game in 2002, despite the Buckeyes' winning six regular season games by seven points or less.

The BCS formula came under increased scrutiny in 2003, when the system produced a split national championship – exactly the outcome it was created to prevent. This occurred because the University of Southern California (USC) finished first in both media polls, yet was excluded from participating in the championship game. Many felt that by *only* considering wins and losses (and thus ignoring scoring margin), the computer polls slighted USC, since all eleven

[1] From 1950–1990, both the Associated Press (AP) and United Press International (UPI) released polls. In 1991 the UPI poll was replaced by the ESPN/USA Today coaches' poll.

[2] Measuring schedule strength is exceedingly difficult. The NCAA often uses measures, such as opponents' winning percentage, which represent the *average* strength of the competition faced. However, for elite teams, facing a good (but not great) team each week may constitute an easier schedule, i.e. result in a higher probability of an unbeaten season, than facing many poor teams and a few other elite teams, even though the average quality of the opposition may be higher for the first schedule.

of their regular season wins were by at least 17 points, while their lone loss was by three points in triple overtime.

The creation of the BCS has sparked increased interest in computerized methods for ranking football teams. However, the experiences of the last few years illustrate the difficulty in creating an equitable ranking system. By considering margin of victory, it is possible that an unbeaten team can be ranked behind one with two or even three losses. On the other hand, by only considering wins and losses and essentially "treating all wins equally," one runs the risk of vastly overstating the ability of an unbeaten or once defeated team, which compiled a gaudy record in less than convincing fashion. To address this problem, we propose a new statistical model – one which incorporates both wins *and* scores, rather than choosing between them.

This paper is organized as follows. Section 2 gives an overview of existing paired-comparison methods used to rank college football teams. Section 3 details the proposed "hybrid" model, which can be viewed as a synthesis of its predecessors. Computational issues and special cases of the proposed method are addressed in Section 4, and the model's performance is examined and compared to that of its peers via simulation studies in Section 5. It is seen to be quite robust to model mis-specification, in sharp contrast to existing methods. Section 6 illustrates the new methodology using college football data. Finally, Section 7 presents some concluding remarks.

2 Current Paired Comparison Ranking Methods

2.1 Win/Loss Models

Existing ranking methods fall into two fundamental classes, one based on wins and losses, and one based on accrued point totals. The first class of models contends that the wins and losses of each team are of primary interest. The motivation for models of this type is obvious since the goal of any team in a game is to win. Due to the data on which they are based, models of this type will be referred to in subsequent discussion as *win/loss models*. Two models in this class are the well-known Bradley-Terry [4] and Thurstone-Mosteller [23], [20] models.

The Bradley-Terry model arises by considering latent (i.e. unobserved) point-scoring processes which follow independent Gumbel[3] distributions with common scale parameters but differing location parameters. When estimated, these location parameters lead to a rank ordering of teams (see Davidson [7] for a discussion of equivalent motivations). Under this formulation, each team is described by its ability, α, that acts as the location parameter governing its cumulative distribution of points scored, $F_\alpha(x) = \exp[-e^{-(x-\alpha)}]$, $-\infty < x, \alpha < \infty$. Team i defeats team j precisely when i's point total ex-

[3]The Gumbel distribution is sometimes referred to as the Extreme Value, Type I distribution.

ceeds j's. It is straightforward to verify that the probability of this event is $\pi_{i,j} = e^{\alpha_i}/(e^{\alpha_i} + e^{\alpha_j})$. Lehmann [13] introduces a class of distributions (of which the Gumbel is a member) which, if chosen to govern point distributions, yield the same probability of victory. All such Lehmann distributions share the feature that their convolution follows a logistic distribution with location parameter $(\alpha_i - \alpha_j)$ and unit scale parameter.

In a similar fashion, the Thurstone-Mosteller model can be derived by considering latent Gaussian random variables with common variance equal to $^1/_2$ and means to be estimated. In this case, the point differential is also Gaussian, and the probability that team i defeats team j is $\pi_{i,j} = \Phi(\alpha_i - \alpha_j)$, where $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function. Both the Bradley-Terry and Thurstone-Mosteller models belong to a more general class of *linear preference scales* proposed by David [6] for which the probability that i defeats j is a function of only the difference in their abilities. That is, $\pi_{i,j} = H(\alpha_i - \alpha_j)$ where $H(\cdot)$ is a symmetric, monotone, non-decreasing function satisfying $H(-\infty) = 0$, $H(+\infty) = 1$ and $H(y) = 1 - H(-y)$. (The function $H(\cdot)$ can be thought of as the cumulative distribution function of a random variable which is symmetric about the origin.)

These and other win/loss models do not provide sufficient grounds for differentiating between teams when the number games played is limited. In particular, when a team, say team k, finishes the season undefeated, the likelihood of either model is maximized by letting α_k tend to infinity, resulting in degenerate fitted probabilities (i.e. $\hat{\pi}_{k,i} = 1$ or, equivalently, $\hat{\pi}_{i,k} = 0$) for all games in which team k competed. While it might not seem unreasonable for an unbeaten team to be top-ranked, a model that mandates this regardless of the given team's schedule is overly restrictive. For example, it is not uncommon for there to be no Division I-A unbeatens and at least one unbeaten team in a lower division. In these cases, such models conclude that an undefeated team in Division II or III is vastly (in fact, infinitely) superior to a one-loss Division I-A team. This last occurred in 2003 when there were no undefeated teams in Division I-A, but three undefeated teams in lower divisions: Pennsylvania (Division I-AA), St. John's (Division III) and Carroll College (NAIA). Despite their better win/loss records, few would argue that any of these teams were as good as Southern California (USC) or Louisiana St. (LSU), both one-loss teams in Division I-A. It might seem that this dilemma could be remedied by inclusion of a "strength-of-schedule" parameter. Alas, this is wishful thinking as these models already use *both* teams to predict the response (in this case wins). Since each team's opposition is tied to its observed performance, the strengths of its opponents are *implicitly* considered even though there is no *explicit* strength-of-schedule parameter. Because strength is implicitly considered, any subsequent *ad hoc* adjustment is unnecessary and tantamount to double-bookkeeping. Further, *ad hoc* parameters are not justifiable in a statistical sense – if they were, they wouldn't be *ad hoc*. It should be noted that since this trouble arises solely because some teams are unbeaten (or winless), there are approaches, such as penalized maximum likelihood (see Mease [18]),

which overcome this difficulty but they are often subjective and thus decisions made by the analyst can affect the rankings.

2.2 Point-Scoring Models

The second model class, by contrast, considers each team's point totals the relevant quantities on which inference should be based. The motivation here is more subtle. As noted earlier, many common binary response models can be derived by considering parallel, latent point totals earned by each team, such that the team scoring the larger number of points is deemed the winner. Since these point-scoring processes are observable in football, use of these models is based on the premise that analysis should use the (uncensored) point totals, rather than the (censored) wins and losses. We will refer to these as *point-scoring models*.

The most common point-scoring model treats observed point totals as realizations drawn from independent, heteroscedastic Gaussian distributions, whose means depend on the teams' offenses and defenses. Under this formulation, the mean number of points scored by team i against team j is given by $\mathbb{E}(Y_{i,j}) = \varphi + \beta_i - \gamma_j$, where φ is an intercept, β_i is the offensive ability of team i and γ_j is the defensive ability of team j. Harville [9], [10] uses variants of this approach to model football scores. Although the Gaussian model is by far the most prevalent in the literature, one might also consider the scores as arising from Poisson distributions, such that $\log[\mathbb{E}(Y_{i,j})] = \varphi + \beta_i - \gamma_j$. The Poisson formulation has two desirable features not found in the Gaussian model. First, the observed scores are non-negative integers, rather than continuous, and possibly negative, random variables. Second, very large scores are assumed to have more variability than very low ones. This is in keeping with intuition, as the winning score in a blowout can take on a wide range of values, while the point total of an overmatched team is likely to fall in a much smaller range. Dixon and Coles [8] investigate models of this type for European soccer data.

Many common point-scoring models, including both the Gaussian and Poisson models, suffer from a sufficiency principle when the canonical link[4] is used. In such cases, the total points scored and allowed over the course of the season by each team are sufficient statistics for the model parameters. Therefore, models of this sort may overemphasize extremely high or low point totals and underemphasize the importance of winning each week. The following example, presented in Table 1, illustrates this idea.

In *Season 1*, team A defeats both teams B and C, and almost any reasonable method of ranking would determine A the champion. In *Season 2*, however, team B defeats both A and C, yet because the marginal point totals are unchanged, parameter estimates (and subsequently rankings) for the two

[4]In the Gaussian case, the canonical link is identity, and in the Poisson case, the canonical link is the natural logarithm.

Table 1: Sufficiency can deemphasize winning.

Season 1	Season 2
A defeats B, 21 – 20.	B defeats A, 24 – 17.
A defeats C, 35 – 7.	A defeats C, 39 – 3.
B defeats C, 35 – 7.	B defeats C, 31 – 11.

scenarios are identical. Thus, by "running up the score," A compensates for its outright loss to B.

Clearly, both wins and points contain information relevant to the ranking problem. The question, therefore, is not whether one should model wins *or* scores. Rather, the question is how to model wins *and* scores simultaneously, thus exploiting all the available information.

3 A Hybrid Ranking System

In this section we develop a model which incorporates both win/loss records and point totals. This hybrid model can be thought of as augmenting the win/loss data with scores that give some measure of the degree of victory. This reflects the reality that a twenty-point victory is much more convincing than winning by only a point or two. From this perspective, each game yields a multivariate response consisting of a win/loss indicator, a winning score and a losing score. These responses are modeled using a two-stage hierarchy, described below.

3.1 Win/Loss Indicators

The win/loss indicator is modeled as a binomial random variable, with success probability related to the teams involved. Specifically, let $\delta_{i,j}$ be the indicator that i wins a game against j. Then $\Pr(\delta_{i,j} = 1) = \pi_{i,j} = H(\alpha_i - \alpha_j)$, as in Section 2.1. Since each team's success depends on its ability to both score points and prevent its opponents from scoring, we define each team's merit as the sum of its offensive and defensive abilities. Mathematically, $\alpha_i = \beta_i + \gamma_i$.

Two common choices of $H(\cdot)$ appropriate for this type of logistic regression are the cumulative logistic and cumulative Gaussian distributions. In Section 2.1, it was shown that if only wins and losses are considered, these correspond to the Bradley-Terry and Thurstone-Mosteller models, respectively. Prentice [21] explores a general class of H-functions which he applied to bioassay. We adopt his general formulation and enforce symmetry about zero (which ensures $\pi_{i,j} = 1 - \pi_{j,i}$).

$$H(\alpha_i - \alpha_j) = \pi_{i,j} = \int_{-\infty}^{\alpha_i - \alpha_j} \frac{e^{zm}\left(1 + e^z\right)^{-2m}}{Be(m,m)}\, dz, \tag{1}$$

where $Be(\cdot,\cdot)$ is the beta function, and m is a non-linear parameter which indexes $H(\cdot)$ over a range of possible functions. Special cases include the

cumulative logistic ($m = 1$) and cumulative Gaussian ($m \to \infty$) distribution functions.

3.2 Conditional Scores

The winning and losing scores are subsequently modeled as conditional responses given the win/loss outcome. The intuition behind conditioning is that knowing team i defeated team j changes our opinion of the likely point totals. Specifically, this knowledge should inflate the estimated point total of the known winner and deflate the estimated point total of the loser. We assume that there is a linear predictor for each point total of the form $g[\mathbb{E}(Y_{i,j})] = \varphi + \beta_i + \gamma_j$ for some monotonic, differentiable function $g(\cdot)$. The conditioning is accomplished by letting $g(\cdot)$ vary (depending on whether the point total being modeled corresponds to a winning or losing score). A convenient, yet flexible, means to accomplish this is to let $g = g_\lambda$ be the family of Box-Cox [3] transformations,

$$\mathbb{E}(Y_{i,j}|\delta_{i,j}) = g_\lambda^{-1}(z) = \begin{cases} (\lambda_{[2-\delta_{i,j}]}z + 1)^{1/\lambda_{[2-\delta_{i,j}]}}, & \lambda_{[2-\delta_{i,j}]} \neq 0; \\ \exp(z), & \lambda_{[2-\delta_{i,j}]} = 0. \end{cases} \tag{2}$$

For any fixed $z \in \{z : z\lambda_j > -1; j = 1, 2\}$ and $\lambda_1 < \lambda_2$, the conditional expectations satisfy $g_{\lambda_1}^{-1}(z) > g_{\lambda_2}^{-1}(z)$. This suggests conditioning the scores through the Box-Cox parameter, λ, as follows. Because the linear predictor, $g_\lambda[\mathbb{E}(Y_{i,j})] = \varphi + \beta_i + \gamma_j$, depends on the offensive and defensive abilities of the teams involved (which do not change depending on i's or j's winning), estimating two different values of the Box-Cox parameter — λ_1 for winning scores and λ_2 for losing scores — allows $\mathbb{E}(Y_{i,j}|i \text{ defeats } j) > \mathbb{E}(Y_{i,j}|j \text{ defeats } i)$, provided $\lambda_1 < \lambda_2$.

In certain instances, one may have reason to believe that the conditional point totals are governed by Gaussian or perhaps Poisson processes. However, in most circumstances, there is no rationale for preferring one of these models to the others *a priori*. Therefore, these conditional point totals are modeled using the more general Tweedie [24] power variance model. In this case, the conditional mean and variance of a score, say $Y_{i,j}$, are given by:

$$\mathbb{E}(Y_{i,j}|\delta_{i,j}) = \mu_{i,j}^{(\lambda)} \qquad \text{Var}(Y_{i,j}|\delta_{i,j}) = \theta_1[\mu_{i,j}^{(\lambda)}]^{\theta_2},$$

where the superscripted λ emphasizes the dependence of the conditional mean point total on the win/loss outcome of the game. Special cases of this model include the Gaussian distribution ($\theta_2 = 0$), the Poisson ($\theta_1 = \theta_2 = 1$) and the Gamma ($\theta_2 = 2$) distributions. Values of θ_2 contained in the open interval $(0, 1)$ do not correspond to proper likelihoods. This is not a problem, as estimation of both mean and variance parameters is achieved using Liang and Zeger's [14] method of Generalized Estimating Equations (*GEE*), which, in lieu of a complete likelihood, requires only specification of the first two moments.

3.3 Complete Formulation

In Section 3.1 and Section 3.2, the mean-variance relationship for the win/loss indicators and conditional point totals were formulated in terms of teams' offensive abilities, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_N)^T$, defensive abilities, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_N)^T$, and in the case of the conditional scores, an intercept, φ. Furthermore, each team's overall ability is the sum of its offensive and defensive strengths: $\boldsymbol{\alpha} = \boldsymbol{\beta} + \boldsymbol{\gamma}$. Therefore if game t is between teams i and j, the response vector is

$$\mathbf{u}_t = \begin{pmatrix} \delta_{i,j} \\ y_{i,j}|\delta_{i,j} \\ y_{j,i}|\delta_{i,j} \end{pmatrix},$$

whose mean and covariance are given by

$$\mathbb{E}(\mathbf{u}_t) = \boldsymbol{\mu}_t = \begin{pmatrix} \mu_{t,1} \\ \mu_{t,2} \\ \mu_{t,3} \end{pmatrix} = \begin{pmatrix} H[(\beta_i + \gamma_i) - (\beta_j + \gamma_j)] \\ g^{-1}_{\lambda_{[2-\delta_{i,j}]}}(\varphi + \beta_i - \gamma_j) \\ g^{-1}_{\lambda_{[1+\delta_{i,j}]}}(\varphi + \beta_j - \gamma_i) \end{pmatrix} \tag{3}$$

$$\mathrm{Var}(\mathbf{u}_t) = \mathbf{V}_t = \begin{bmatrix} \mu_{t,1}(1-\mu_{t,1}) & 0 & 0 \\ 0 & \theta_1[\mu_{t,2}]^{\theta_2} & \theta_3\{\theta_1[\mu_{t,2}\mu_{t,3}]^{\theta_2/2}\} \\ 0 & \theta_3\{\theta_1[\mu_{t,2}\mu_{t,3}]^{\theta_2/2}\} & \theta_1[\mu_{t,3}]^{\theta_2} \end{bmatrix}, \tag{4}$$

where $\theta_3 \in [-1, 1]$ is the correlation between the winning and losing scores given the game outcome. Note that because the scores are modeled conditionally (given the game outcomes) that $\mathrm{Cov}(\delta_{i,j}, Y_{i,j}|\delta_{i,j}) = \mathrm{Cov}(\delta_{i,j}, Y_{j,i}|\delta_{i,j}) = 0$.

An appealing result of this formulation is the transitivity of both winning probability and expected scores. That is, for two teams, i and j, if $(\beta_i + \gamma_i) = \alpha_i > \alpha_j = (\beta_j + \gamma_j)$, then the probability of team i defeating team j is greater than 0.5, and i's expected score against j is greater than j's against i. Therefore, if i is "better" than j (based on either win probability or expected score) and j is "better" than k, then i is also "better" than k. This allows for an unambiguous global ranking[5]. It should be noted that other ranking procedures enjoy this type of transitivity, among them the Bradley-Terry and Thurstone-Mosteller models for win/loss data and the Gaussian and Poisson models for points scored. That these other methods also preserve transitivity should not be surprising, as we will show in the next section that each of these models is a limiting case of the proposed hybrid method.

[5]An example of intransitivity which results in no clear ranking of alternatives is the children's game "Rock/Paper/Scissors." In this game "rock breaks scissors," "scissors cut paper" and "paper covers rock." Because the pairwise preferences are intransitive, rank-ordering is not unique.

3.4 Parameter Estimation

Equations 3 and 4 specify the first two moments of the multivariate response (consisting of a win/loss indicator and the total points scored by each team). However, they do not provide a likelihood to be maximized. Indeed, for certain values of θ_2 there will be no likelihood satisfying the moment constraints. Parameter estimates, therefore, are obtained via generalized estimating equations. The idea underlying *GEE* is that residuals should be set "as close to zero" as possible. In this sense, a naïve estimating function for the mean parameters, $\boldsymbol{\xi} = (\varphi, \boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\lambda}^T)^T$, is

$$\sum_{t=1}^{n} (\mathbf{u}_t - \boldsymbol{\mu}_t) = \mathbf{0}. \tag{5}$$

It is easily shown (see, for example, Heyde [12]) that among estimating functions based on residuals, the optimal estimating function, i.e. the one which achieves the smallest variance, is a weighted version of Equation 5,

$$\sum_{t=1}^{n} \mathbf{D}_t \mathbf{V}_t^{-1} (\mathbf{u}_t - \boldsymbol{\mu}_t) = \mathbf{0}, \tag{6}$$

where $\mathbf{D}_t = [\partial \boldsymbol{\mu} / \partial \boldsymbol{\xi}]$ is the matrix of partial derivatives of the mean functions with respect to the parameters and $\mathbf{V}_t$ is the variance-covariance matrix of the residuals. Aside from their minimal variance optimality, these estimating functions coincide with the maximum likelihood score equations when the elements $\mathbf{u}_t$ are independent realizations from an exponential family distribution. It was this realization that prompted Wedderburn [25] to propose a quasi-likelihood method of parameter estimation. *GEE* takes the same approach; however, in addition to the unknown mean parameters, $\boldsymbol{\xi}$, additional parameters $\boldsymbol{\theta}$, governing the variance-covariance structure, are estimated as well. The estimating equations for $\boldsymbol{\theta}$ take the same form as Equation 6:

$$[\text{Derivative of Mean Function}]^T \times [\text{Covariance Matrix}]^{-1} \times (\text{residual}) = 0.$$

While the estimating equations for $\boldsymbol{\xi}$ are based on the observed responses $\mathbf{u}_t$, the estimating equations for the $\boldsymbol{\theta}$ may be constructed by using the vector of "squared residuals" as a proxy for their unobserved counterparts, specifically, $\mathbf{v}_t = [(u_{t,1} - \mu_{t,1})^2, (u_{t,2} - \mu_{t,2})^2, (u_{t,2} - \mu_{t,2})(u_{t,3} - \mu_{t,3}), (u_{t,3} - \mu_{t,3})^2]^T$.

$$\sum_{t=1}^{n} \left[\frac{\partial \boldsymbol{\nu}_t}{\partial \boldsymbol{\theta}} \right]^T [\text{Cov}(\mathbf{v}_t)]^{-1} (\mathbf{v}_t - \boldsymbol{\nu}_t) = \mathbf{0}, \tag{7}$$

where $\boldsymbol{\nu} = \mathbb{E}(\mathbf{v}_t)$. Although Equation 7 requires specification of third and fourth moments of $\mathbf{u}_t$ through $\text{Cov}(\mathbf{v}_t)$, estimation for $\boldsymbol{\xi}$ is consistent if Equations 6 and 7 are solved in alternating fashion. A commonly used assumption

for $\mathrm{Cov}(\mathbf{v}_t)$ is to treat the $\mathbf{u}_t$ as approximately multivariate normal, in which case

$$\mathrm{Var}[(u_{t,k} - \mu_{t,k})^2] = 2\mathrm{Var}^2(u_{t,k})$$
$$\mathrm{Cov}[(u_{t,k} - \mu_{t,k}), (u_{t,l} - \mu_{t,l})] = 2\mathrm{Cov}^2(u_{t,k}, u_{t,l}).$$

Equations 6 and 7 can be combined into joint estimating equations,

$$\sum_{t=1}^{n} \begin{pmatrix} \mathbf{D}_t^{(1,1)} & \mathbf{0} \\ \mathbf{D}_t^{(2,1)} & \mathbf{D}_t^{(2,2)} \end{pmatrix}^T \begin{pmatrix} \mathbf{V}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_t \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{u}_t - \boldsymbol{\mu}_t \\ \mathbf{v}_t - \boldsymbol{\nu}_t \end{pmatrix} = \mathbf{0}. \tag{8}$$

where

$$\mathbf{D}_t^{(1,1)} = \left[\frac{\partial \boldsymbol{\mu}_t}{\partial \boldsymbol{\xi}}\right],\ \mathbf{D}_t^{(2,2)} = \left[\frac{\partial \mathbf{v}_t}{\partial \boldsymbol{\theta}}\right],\ \mathbf{\Lambda}_t = \mathrm{Cov}(\mathbf{v}_t)$$

and $\mathbf{D}_t^{(2,1)}$ is either a zero-matrix or the matrix of partial derivatives of the variance functions with respect to the mean parameters. When $\mathbf{D}_t^{(2,1)} = \mathbf{0}$, computations are simplified and estimation of the mean parameters (which are of interest) is consistent even if the presumed covariance structure is incorrect. On the other hand, when $\mathbf{D}_t^{(2,1)} = [\partial \mathbf{v}_t / \partial \boldsymbol{\xi}]$ efficiency is increased at the expense of robustness to model mis-specification. Liang et al. [15] refer to these treatments of $\mathbf{D}_t^{(2,1)}$ in Equation 8 as *GEE-1* and *GEE-2*, respectively.

In addition to parameter estimates, standard errors are available at convergence. McCullagh [17] shows that if the covariance structure is correctly specified, the mean parameters are asymptotically Gaussian, and a model-based estimate of parameter covariance is given by

$$\mathrm{Cov}(\hat{\boldsymbol{\xi}}) = \left[\sum_{t=1}^{n} \begin{pmatrix} \mathbf{D}_t^{(1,1)} & \mathbf{0} \\ \mathbf{D}_t^{(2,1)} & \mathbf{D}_t^{(2,2)} \end{pmatrix}^T \begin{pmatrix} \mathbf{V}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_t \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{D}_t^{(1,1)} & \mathbf{0} \\ \mathbf{D}_t^{(2,1)} & \mathbf{D}_t^{(2,2)} \end{pmatrix}\right]^{-1}.$$

In the event that the covariance is mis-specified, White [26] shows that while $\hat{\boldsymbol{\xi}}$ based on *GEE-1* is consistent and its limiting distribution remains Gaussian, the covariance is no longer coincident with McCullagh's expression. Instead, it is a function of the assumed *and* true covariance structure. To hedge against covariance mis-specification, White advocates moment-based estimation of the true covariance and a more complicated "sandwich" estimator of $\mathrm{Cov}(\hat{\boldsymbol{\xi}})$.

4 Computational Considerations

4.1 Limiting Cases

Perhaps the most appealing aspect of the proposed model is its generality. In Section 5, we show that it is robust to model mis-specification in situations

where existing models are not. Indeed, the hybrid model performs well for a wide range of true models. This versatility should not come as a surprise, since many well-known models are embedded in our formulation. In particular, the hybrid model generalizes the Bradley-Terry and Thurstone-Mosteller models for win/loss data as well as the Gaussian and Poisson point-scoring models. Table 2 lists some of the sub-models and the corresponding hybrid model parameters from which they arise.

Table 2: The proposed model generalizes a number of commonly-used paired comparison models.

λ	m	θ	Special Case
—	$m = 1$	$\theta_1 \uparrow \infty$	Bradley-Terry
—	$m \uparrow \infty$	$\theta_1 \uparrow \infty$	Thurstone-Mosteller
$\lambda_1 = \lambda_2 = 1$ (Canonical link)	—	$\theta_1 \downarrow 0$; $\theta_2 = 0$	Gaussian
$\lambda_1 = \lambda_2 = 0$ (Canonical link)	—	$\theta_1 \downarrow 0$; $\theta_2 = 1$	Possion
—	—	$\theta_1 \downarrow 0$; $\theta_2 = 2$	Gamma

The first two cases correspond to win/loss models. These represent one model extreme in which there is no information contained in the scores of the games. At the other extreme, are the point-scoring models. By choosing appropriate values of θ_2 independent Gaussian, Poisson and Gamma point-scoring models may be achieved. Although $\theta_1 \downarrow 0$ may not be realistic, it illustrates a potential situation in which all inference will be based on scores. In reality, however, even if scores were generated from independent Gaussian distributions, there would be *some* information contained in wins and losses. (In fact, this would result in exactly the probit model for logistic regression.) The relationship between the general model and some of these subclasses is investigated further in the next section.

4.2 Numerical Stability

Because there is a limited amount of data in a college football season (roughly eleven games per team) relative to the number of parameters to be estimated (two per team), numerical convergence can be difficult. It has been our experience that inconsistent initial values (resulting in very poor initial fits) may cause singularity in the gradient. For that reason, we recommend starting the procedure at a "known point," for instance, starting $\lambda_1 = 0$, $\lambda_2 = \epsilon$ (for some small, positive number ϵ) and $\boldsymbol{\theta} = (1, 1, 0)^T$. These choices of $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ correspond to a Poisson generalized linear model with log-link. As such, it is appropriate to initialize the mean parameters φ, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ parameters at the maximum likelihood estimators resulting from a Poisson regression on scores. Other reasonable choices are possible. For instance, a linear regression on the

scores suggests initializing $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ at their MLEs, $\boldsymbol{\lambda}$ at $(1, 1+\epsilon)^T$ and $\boldsymbol{\theta}$ at $(\hat{\sigma}^2, 0, 0)^T$, where $\hat{\sigma}^2$ is the estimate of the variance for an ordinary least-squares procedure. Care must be taken, however, when initializing φ. Because the transformation given by Equation 2 is the continuity-corrected version of the Box-Cox transformation, φ should be initialized at $\hat{\varphi} - 1$, rather than $\hat{\varphi}$.

In addition to selecting appropriate initial values, the Fisher scoring method of updating (detailed in Appendix A) tends to suggest overly large steps when starting from a position far from the root of the estimating functions. Therefore, it may be necessary, when programming the algorithm, to limit the maximum absolute step size to prevent divergence. (Bounding step size is a common numerical-methods technique.) Finally, the non-linear parameters governing the mean responses (i.e. the Box-Cox parameters, $\boldsymbol{\lambda}$, and the Prentice win/loss index parameter, m) may influence the absolute scale of the φ, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ parameters. This is to be expected, as varying m is analogous to changing the link function of a logistic regression, and it is well-known that the parameter estimates for two particular choices of links may differ drastically even though the fits can be quite similar. This interdependence of parameters is caused by using the same set of parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, to describe both the win/loss performance and scoring/defending ability of the teams, which is a necessity if these estimates are to lead to an unambiguous rank-ordering of teams. Since the linear parameters are the same, it is necessary that the model be flexible enough in its non-linear transformations so as to provide acceptable fits to both types of data.

In some circumstances, one may wish to simplify the model formulation to expedite computation. Two effective simplifications assume *a priori* knowledge of m and θ_3. When $m = 1$, the win/loss indicators are linked to the linear parameters via the commonly used logit link. When m is a large, fixed value, this link is approximately probit. In a number of studies, assuming one of these values or another did not alter fits substantially (though no formal goodness of fit statistic is available) but resulted in more stable and faster convergence. Similarly, assuming $\theta_3 = 0$ forces the conditional scores to be independent. Again, this results in substantial gains in convergence speed and stability. (In fact, for some data sets, models with non-zero θ_3 will not converge at all.) Although this simplification of the covariance structure may be less efficient than the fully parameterized model *when the covariance is correctly specified*, point estimates on which rankings are based remain consistent estimators of the true parameters.

5 Simulation Studies

We present a series of simulation studies to validate the proposed ranking method. Two groups of studies are given. The first generates data consistent with what might be seen in a college football season, i.e. comparatively few games for the number of teams to be ranked and a dramatically unbalanced

schedule (driven by conference membership). The second study focuses on a situation for which there are comparatively few teams and a large number of games on which to base rankings.

In light of the numerical stability issues discussed in Section 4.2, the binary index parameter, m, was fixed *a priori* to either 1 (corresponding to a logit binary link for the win/loss data) or allowed to tend to infinity (corresponding to a probit binary link for the win/loss data). Furthermore, to expedite computations, a working independence structure (i.e. $\theta_3 = 0$) was assumed for the conditional point totals. Such simplified estimating equations remain consistent for the mean parameters on which the rankings are based. See Liang and Zeger [14] for a discussion.

5.1 Football-Specific Simulations

5.1.1 Scheduling

Determining relative strengths of teams in college football is quite difficult, as there are many teams and a comparatively small number of games by which to judge their performance. This problem is compounded by the disparate schedules faced by top teams. Not only do many highly-ranked teams never play each other, they seldom face common opposition. Our first series of simulation studies mimics this type of scheduling.

We consider a situation in which there are four "conferences" each consisting of six teams. The schedule is constructed such that each team faces each member of its conference once and one out-of-conference opponent. (These out-of-conference games are necessary for a global ranking. Were there no such games, only intra-conference rankings would be possible.) Furthermore, we enforce a pseudo-balance condition that each conference faces opposition from each other conference exactly twice. A typical non-conference schedule is illustrated in Figure 1. Lines connecting two teams represent a game between them. The conferences are labeled A through D, and the teams within each conference are numbered 1 through 6. In the figure, A1 plays B1, B6 plays D4, etc.

5.1.2 Data Generation

The two primary ranking model classes presented in Section 2 are based on wins and losses (ignoring points scored and allowed) or point-scoring data (ignoring wins and losses). Depending on the true data-generating mechanism, one class of models can be expected to consistently outperform the other. To assess our proposed method under varying circumstances, two simulation studies were performed: one in which wins and losses were sufficient statistics for the parameters, and one in which point-totals were.

In the first simulation, data were generated to follow a Thurstone-Mosteller model. For each of 500 simulated seasons, team strengths, $\boldsymbol{\alpha}$ ($\boldsymbol{\alpha} = \boldsymbol{\beta}+\boldsymbol{\gamma}$), were

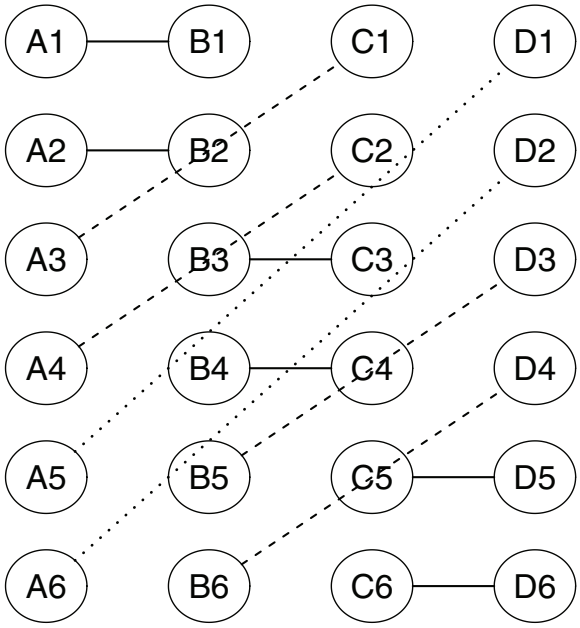

Figure 1: Each team plays one non-conference conference opponent, and each pair of conferences plays twice.

randomly drawn from independent standard Gaussian distributions and the probability that team i defeats team j is $\Phi(\alpha_i - \alpha_j)$, where $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function. Given the wins and losses, the losing scores and the margins of victory were generated from independent χ^2_{15} distributions. Thus, the mean winning and losing scores were 30 and 15, respectively (consistent with college football data) with extreme scores possible. Notice that the individual game scores have nothing to do with the teams involved *once the win/loss has been established*. This type of simulation is a best-case scenario for the win/loss models and a worst-case one for the point-scoring models.

In the second simulation, data were generated from independent Poisson distributions. For each of 500 seasons, the offensive ($\boldsymbol{\beta}$) and defensive ($\boldsymbol{\gamma}$) parameters were randomly drawn from independent Gaussian distributions having mean 0 and standard deviation 0.4, and the intercept, φ, was held constant at 2.5. (As was the case in the win/loss simulation, the parameters were chosen to achieve plausible football scores.) After generating the model parameters, the points scored by team i against team j was a Poisson random variable with mean $\mu_{i,j} = \exp\{\varphi + \beta_i - \gamma_j\}$. Because the Poisson distribution has its support on an integer lattice, there is a non-zero probability of a tied outcome. In these cases (which comprised less than 5% of the outcomes), the scores were kept and the win/loss outcome was determined by "a flip of the coin" i.e. a Bernoulli random variable having probability 0.5. Notice that the wins and losses add nothing to the analysis *once the point totals have been*

established. This type of simulation is a best-case scenario for the point-scoring models and a worst-case one for the win/loss models.

5.1.3 Results

For the first series of simulations, wins and losses were generated based on team strengths and scores, while generated, were meaningless, i.e. *given the win/loss outcomes*, scores were unrelated to team abilities. In such a case, models based on points scored and allowed would be expected to perform poorly compared to those which consider only wins and losses. This was the case in our simulations.

Two metrics were chosen to assess model performance – probability of correctly identifying the true best team and probability of identifying the two best teams (in either order). These are reasonable criteria for college football since the two primary areas of interest center on selecting the single best team after all games are played and selecting the two most deserving participants for a championship game. The results are given in Table 3.

Table 3: Comparison of models based on probability of identifying the best team and best two teams when wins and losses follow a Thurstone-Mosteller model.

	One Team	Two Teams
Bradley-Terry	173/500 (34.6%)	77/500 (15.4%)
Thurstone-Mosteller	167/500 (33.4%)	76/500 (15.2%)
Gaussian	148/500 (29.6%)	66/500 (13.2%)
Poisson	149/500 (29.8%)	62/500 (12.4%)
Hybrid ($m = 1$)	155/500 (31.0%)	59/500 (11.8%)
Hybrid ($m \uparrow \infty$)	159/500 (31.8%)	64/500 (12.8%)

It should be no surprise that the win/loss models performed best in this situation, although since the true model is the Thurstone-Mosteller, it is somewhat interesting that the Bradley-Terry model performed slightly better for these data. When the criterion considered is the probability of correctly identifying the best team, the hybrid model outperforms the "incorrect" (point-scoring) models and lags the "correct" (win/loss) models, as would be expected. Tables 4 and 5 give simultaneous 95% Scheffé-type confidence intervals for the *difference* in these probabilities. They show that our hybrid model is not significantly different from any of the aforementioned models, as all confidence intervals contain zero.

Although there are no significant differences between the hybrid model and the other types for this type of win/loss simulation, we expect that, with a large enough sample, such differences would emerge. We show this to be the case in the point-scoring simulation studies that follow.

In the second series of simulations, scores were generated from independent Poisson distributions. Again, the two criteria by which models were compared

Table 4: When the wins and losses follow a Thurstone-Mosteller formulation, there is no significant difference in the probability of correctly identifying the best team between the the hybrid model and either of the other model classes.

Difference in Probability of Determining Correct #1 Team	Simultaneous Confidence Intervals
Hybrid ($m = 1$) – Bradley-Terry	(-0.115,0.043)
Hybrid ($m \uparrow \infty$) – Bradley-Terry	(-0.102,0.046)
Hybrid ($m = 1$) – Thurstone-Mosteller	(-0.103,0.055)
Hybrid ($m \uparrow \infty$) – Thurstone-Mosteller	(-0.090,0.058)
Hybrid ($m = 1$) – Gaussian	(-0.038,0.066)
Hybrid ($m \uparrow \infty$) – Gaussian	(-0.031,0.075)
Hybrid ($m = 1$) – Poisson	(-0.026,0.050)
Hybrid ($m \uparrow \infty$) – Poisson	(-0.013,0.053)

Table 5: When the wins and losses follow a Thurstone-Mosteller formulation, there is no significant difference in the probability of correctly identifying the best two teams between the the hybrid model and either of the other model classes.

Difference in Probability of Determining Top Two Teams	Simultaneous Confidence Intervals
Hybrid ($m = 1$) – Bradley-Terry	(-0.091,0.019)
Hybrid ($m \uparrow \infty$) – Bradley-Terry	(-0.079,0.027)
Hybrid ($m = 1$) – Thurstone-Mosteller	(-0.090,0.022)
Hybrid ($m \uparrow \infty$) – Thurstone-Mosteller	(-0.078,0.030)
Hybrid ($m = 1$) – Gaussian	(-0.052,0.024)
Hybrid ($m \uparrow \infty$) – Gaussian	(-0.042,0.034)
Hybrid ($m = 1$) – Poisson	(-0.037,0.025)
Hybrid ($m \uparrow \infty$) – Poisson	(-0.025,0.033)

were the probability of correctly identifying the single best team and the probability of identifying the top two teams (irrespective of order). The results are given in Table 6.

Table 6: Comparison of models based on probability of identifying the best team and best two teams when points scored are independent Poisson random variables.

	One Team	Two Teams
Bradley-Terry	274/500 (54.8%)	182/500 (36.4%)
Thurstone-Mosteller	338/500 (67.6%)	246/500 (49.2%)
Gaussian	173/500 (34.6%)	77/500 (15.4%)
Poisson	173/500 (34.6%)	79/500 (15.8%)
Hybrid ($m = 1$)	312/500 (62.4%)	212/500 (42.4%)
Hybrid ($m \uparrow \infty$)	314/500 (62.8%)	217/500 (43.4%)

As expected, the Poisson model performed best in this situation, while the win/loss models performed poorest. This can be attributed to their effective

censoring of each score pair into an indicator that one was larger than the other. Tables 7 and 8 give simultaneous 95% confidence intervals for the *difference* in these probabilities. They show that for both metrics, our hybrid model is significantly better than both "incorrect" (win/loss) models. Furthermore, the difference in probability of correctly identifying the best team between the hybrid ranking method and either of the point-scoring models is not significant. While there is no significant difference between ours and the Gaussian results, the hybrid model is slightly worse than the (true) Poisson model at determining the best two teams. However, were the sample size larger, we expect that the hybrid model would significantly outperform the Gaussian model and lag the true model for both criteria.

Table 7: When the scores are Poisson, the hybrid model has a significantly larger probability of correctly identifying the best team than either of the win/loss models. There is no significant difference between the performance of our model and the true model.

Difference in Probability of Determining Correct #1 Team	Simultaneous Confidence Intervals
Hybrid ($m = 1$) – Bradley-Terry	(0.184,0.372)
Hybrid ($m \uparrow \infty$) – Bradley-Terry	(0.190,0.374)
Hybrid ($m = 1$) – Thurstone-Mosteller	(0.186,0.370)
Hybrid ($m \uparrow \infty$) – Thurstone-Mosteller	(0.192,0.372)
Hybrid ($m = 1$) – Gaussian	(-0.012,0.164)
Hybrid ($m \uparrow \infty$) – Gaussian	(-0.006,0.166)
Hybrid ($m = 1$) – Poisson	(-0.104,0.000)
Hybrid ($m \uparrow \infty$) – Poisson	(-0.095,-0.001)

Table 8: When the scores are Poisson, the hybrid model has a significantly larger probability of correctly identifying the top two teams than either of the win/loss models. There is a slightly significant difference between the performance of our model and the true model.

Difference in Probability of Determining Top Two Teams	Simultaneous Confidence Intervals
Hybrid ($m = 1$) – Bradley-Terry	(0.185,0.355)
Hybrid ($m \uparrow \infty$) – Bradley-Terry	(0.196,0.364)
Hybrid ($m = 1$) – Thurstone-Mosteller	(0.180,0.352)
Hybrid ($m \uparrow \infty$) – Thurstone-Mosteller	(0.191,0.361)
Hybrid ($m = 1$) – Gaussian	(-0.026,0.146)
Hybrid ($m \uparrow \infty$) – Gaussian	(-0.016,0.156)
Hybrid ($m = 1$) – Poisson	(-0.124,-0.012)
Hybrid ($m \uparrow \infty$) – Poisson	(-0.112,-0.004)

5.1.4 Discussion

In the two aforementioned simulation studies, game outcomes (wins, losses and point totals) were generated according to either a win/loss or a point-scoring model. In both cases, the results show that the hybrid model performs better than the "incorrect" models[6] and slightly worse than the true model. Simultaneous confidence intervals were constructed to determine if the results of the hybrid ranking procedure significantly differ from those of the reference models. The win/loss confidence intervals are inconclusive, while the point-scoring ones offer resounding support for our proposed method.

Perhaps the win/loss simulations were inconclusive because the "amount of information" contained in a single 0/1 win/loss outcome is much less than what is contained in two scores which, intuitively, convey both the win/loss outcome as well as the degree (nail-biter, blowout, etc.). In addition, win/loss simulations are particularly difficult to execute as the hybrid model is over-specified since it contains two parameters per team (offensive and defensive ability) when one (overall ability) suffices[7]. Furthermore, when the scores are generated irrespective of team strengths, the intercept parameter (φ) and the Box-Cox parameters ($\boldsymbol{\lambda}$) are highly dependent. Thus, by keeping all the parameters in the model, the quasi-likelihood surface is flat and optimization is slow and has the potential for finding local optima rather than a global solution. In practice, this is less a concern, as our analysis of real data shows (not surprisingly) that there is *some* information conveyed in scores that is not captured in wins and losses.

5.2 Non-Football Simulation

5.2.1 Scheduling

The previous section illustrated the hybrid model's performance in situations designed to mimic college football data. The sparse, unbalanced schedule, the dearth of games per team and the magnitude of scores were all intended to emulate those of a typical college football season albeit on a smaller scale. While our primary interest is ranking college football teams, there is nothing inherent to our model which would limit its applicability to only those situations. Therefore three additional simulation studies were conducted to assess the hybrid model's performance relative to established ranking methods, *in non-football contexts*. In each case, 400 seasons were simulated such that seven teams competed in a round-robin tournament with 13 replications of each pairing. This is the format of Major League Baseball's intra-divisional

[6] By incorrect models, we mean point-scoring models when the data are generated using wins and losses and win/loss models when the data are generated using a point-scoring model.

[7] Point-scoring models are over-specified as well, however randomness in the data prevent singularity.

schedule before the expansion to three divisions within each league and is used by Agresti [1] to illustrate paired comparisons.

5.2.2 Data Generation

In the first study, win totals were simulated using a binomial model with the probability that team i defeats team j given by the Bradley-Terry formulation where $\alpha_i = \log(i);\ i = 1, 2, \ldots, 7$. Thus, the probability that team 1 defeats team 4 is $\left[{}^{1}/_{(1+4)}\right] = 0.2$. Conditional winning scores were assigned a value of 10 with probability 0.9 and 1,000 with probability 0.1, and conditional losing scores were generated according to a discrete uniform distribution on [0,4]. Clearly, scores contain no additional relevant information apart from the binary outcome they imply, and focusing solely on scores can be very misleading due to the presence of extremely large (but meaningless) values.

In the second simulation study, independent scores were generated using a Gaussian model such that the number of points scored by team i against team j is Gaussian with mean $\mu_{i,j} = \varphi + \beta_i - \gamma_j$ and common variance, σ^2. The intercept and variance were fixed at values of $\varphi = 25$ and $\sigma^2 = 30$, respectively, and the offensive and defensive abilities were randomly drawn from independent Gaussian distributions with mean zero and standard deviation 2.

A final simulation generated independent scores using an over-dispersed Poisson model such that the number of points scored by team i against team j is an over-dispersed Poisson variable with $\log(\mu_{i,j}) = \varphi + \beta_i - \gamma_j$. The intercept and over-dispersion parameters were held fixed at 2.5 and 1.5, respectively, while the offensive and defensive parameters were independently drawn from Gaussian distributions with mean zero and standard deviation 0.3.

5.2.3 Results

Because the number of teams is small and the number of games is large, it is reasonable to wonder how often each procedure correctly recovers the true rank order. The results of the simulations are presented in Tables 9 and 10. Table 9 gives the number of times (and percentage parenthetically) that each of six ranking models retrieved the correct rank ordering of teams for each of the three proposed simulation methods. When outcomes are simulated using the win/loss model, the point-scoring models are influenced by the meaningless scores and substantially underperform both binary types. The hybrid model, by contrast, is not fooled by the noisy scores and performs well compared to the binary models (which included the true data-generating mechanism). At the other extreme, when games are simulated using either of the point-scoring models (Gaussian or Poisson), the performance of the win/loss models substantially lags that of both the point-scoring models, because the win/loss models effectively censor the data. Because it considers scores as meaningful data, the hybrid model does not encounter this difficulty. These simulations

illustrate the sensitivity of existing models' performance to the underlying true data generating mechanism and the robustness of the hybrid model to model mis-specification. In each case, the hybrid model performs nearly as well as the known, true model and substantially better than inappropriately specified models.

Table 9: The hybrid model is robust to mis-specification.

Percentage of Perfect Ranks	Win/Loss	Gaussian Scores	Poisson Scores
Bradley-Terry	26/400 (7%)	123/400 (31%)	141/400 (35%)
Thurstone-Mosteller	31/400 (8%)	129/400 (32%)	141/400 (35%)
Gaussian	5/400 (1.2%)	217/400 (54%)	195/400 (49%)
Poisson	2/400 (0.5%)	211/400 (53%)	221/400 (55%)
Hybrid ($m = 1$)	30/400 (8%)	204/400 (51%)	190/400 (48%)
Hybrid ($m \uparrow \infty$)	31/400 (8%)	209/400 (52%)	193/400 (48%)

Table 10 presents simultaneous 95% confidence intervals for the difference in probability of a correct ranking between the hybrid model and models of an inappropriate type (i.e. point-scoring models when the data are, in fact, binary and win/loss models when the data are generated via independent scoring processes). In all cases, the hybrid model performs significantly better than the mis-specified models. Similar confidence intervals for the difference in performance of the hybrid models and models of the appropriate type contain zero, thus are insignificant.

Table 10: The Hybrid Model is Significantly better than Misspecified Types.

Difference in Probability of Perfect Rank Ordering	Simulation Method		
	Win/Loss	Gaussian	Poisson
Hybrid ($m = 1$) – Bradley-Terry	–	(0.141,0.234)	(0.068,0.177)
Hybrid ($m \uparrow \infty$) – Bradley-Terry	–	(0.153,0.247)	(0.074,0.186)
Hybrid ($m = 1$) – Thurstone-Mosteller	–	(0.130,0.275)	(0.067,0.178)
Hybrid ($m \uparrow \infty$) – Thurstone-Mosteller	–	(0.142,0.288)	(0.074,0.186)
Hybrid ($m = 1$) – Gaussian	(0.028,0.097)	–	–
Hybrid ($m \uparrow \infty$) – Gaussian	(0.030,0.100)	–	–
Hybrid ($m = 1$) – Poisson	(0.038,0.102)	–	–
Hybrid ($m \uparrow \infty$) – Poisson	(0.039,0.106)	–	–

5.2.4 Discussion

Unlike the football-specific simulations, these simulations show considerable advantage in employing our hybrid model to rank the teams. One possible explanation is the type of schedule – large amounts of data allow a superior model to distinguish itself. Our model is robust to model mis-specification and performs comparably to the true model. This is to be expected since it

generalizes many win/loss and point-scoring models (see Section 4.1) and the appropriate limiting case or sub-class can be chosen by the data rather than the analyst.

6 An Application: NCAA Football

In this section, we illustrate the hybrid ranking procedure using actual college football data from the 2003-04 and 2004-05 seasons. In addition to being the two most recent seasons, each presents interesting difficulties from a ranking perspective. Both seasons resulted in a "three teams and two slots" controversy in determining which of three apparently even teams would play for the national championship.

Based on the computational considerations discussed in earlier sections, a sub-model of the general formulation was used to rank the teams. Specifically, the correlation between conditional winning and losing scores (θ_3) was taken to be zero and the Prentice binary index (m) was allowed to tend to infinity, corresponding to $H(\cdot) = \Phi(\cdot)$. This choice was made by considering the binomial deviance when logistic regressions were performed on the win/loss data. For the 2003-04 season, the reduction in deviance achieved by using the probit link ($m \uparrow \infty$) instead of the logit link ($m = 1$) was 3. This difference in deviances corresponds to a likelihood ratio slightly larger than 20 ($e^3 \approx 20.09$) and thus suggests a significantly better fit for the probit model. (It should be noted that although the scale of the other $\boldsymbol{\xi}$ parameters varies with differing choices of m, the relative orders of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and consequently team rankings, are relatively unaffected.) The difference between binary links was less pronounced in 2004-05. In fact, the deviance increases slightly (0.61) when using the probit link. However, since this increase corresponds to a likelihood ratio of less than 2 ($e^{0.61} \approx 1.84$), the probit link was deemed acceptable and used to preserve consistency from one season to another.

In addition to these computational simplifications, a linear home-field advantage parameter (ψ) was included to account for the decided edge a team enjoys when hosting an opponent rather than traveling to play the same team on the road. When home-field advantage is considered, the probability that team i defeats team j at home is a modification of Equation 1,

$$\pi_{i,j}^{(i)} = \int_{-\infty}^{\alpha_i - \alpha_j + \psi} \frac{e^{zm}\left(1+e^z\right)^{-2m}}{Be(m,m)}\, dz,$$

where the parenthetical superscript denotes the home team and the venue effect, ψ, enters only via the upper limit of integration. The effect due to venue is assumed to be equal parts advantage to the home team and disadvantage to the road team. As such, the linear predictors that determine the conditional point totals given in Equation 2 were modified so that $g[\mathbb{E}(Y_{i,j})] = \varphi + \beta_i + \gamma_j + \psi/2$ and $g[\mathbb{E}(Y_{j,i})] = \varphi + \beta_i + \gamma_j - \psi/2$.

6.1 2003 Season

6.1.1 Ranking

In 2003, popular opinion held that three one-loss teams (USC, Oklahoma and LSU) were virtually indistinguishable at the regular season's end. However, since there could only be two championship game participants, many felt that Oklahoma should have been excluded, since the Sooners suffered a lopsided loss in their final game (35-7 to Kansas St.) which dropped them to third in both end-of-season polls. The BCS system, however, chose Oklahoma and consensus #2 LSU to play in the title game (meaning that top-ranked USC was left to play in the Rose Bowl). This resulted in a split national championship, the first since the BCS's inception.

While the controversy focused on USC and Oklahoma (and to a lesser extent LSU), there were no Division I-A unbeaten teams. Thus, models which focus solely on wins and losses (which the BCS mandates) would rank all three of these top teams below such lower division teams as Pennsylvania, St. John's (MN) and Carroll College (MT). By also including scores and thus eliminating the degeneracy from the solution, the hybrid model achieves a sensible ranking, one in which no non-Division I-A teams are awarded dubiously high rankings.

The final BCS rankings as well as a number of model-based rankings are compared in Table 11. Although the home-field advantage is extremely significant, many models currently in use by the BCS fail to account for venue. Despite this, all rankings presnted in Table 11 reflect a home-field advantage term in the model. The win/loss rankings, based on the Bradley-Terry and Thurstone-Mosteller models, were computed by only considering games between Division I-A opponents (thereby precluding lower-division schools from achieving the #1 ranking). Some might be surprised to see that although USC was the consensus #1 team in both media polls, the hybrid model ranked them fifth entering the bowl season. However, USC's best win was against 22nd ranked Auburn and their loss came to 35th ranked California[8]. Texas and Kansas St., by contrast, played in the difficult Big-XII conference and benefitted from "quality wins" (for example, Kansas St. defeated Oklahoma 35-7). Even three-loss Georgia was relatively highly ranked. But upon closer inspection, two of Georgia's three losses were to second-ranked LSU (one on the road during the conference season and one at a neutral site in the conference championship game). Since almost *any* team would lose two of two games to LSU, the model does not penalize Georgia for its difficult schedule as the media polls do. In fact Georgia's two losses to LSU are entirely consistent with their sixth place ranking.

Another interesting feature of the results in Table 11 is the strong showing by Mid-American Conference (MAC) champion Miami OH. The Redhawks – like Oklahoma, LSU and USC – finished their season with one loss: a road

[8]Ranks given are hybrid model rankings.

Table 11: Comparison of 2003 pre-bowl game rankings.

BCS	Hybrid ($m \uparrow \infty$)	Bradley Terry	Thurstone-Mosteller	Gaussian	Poisson
Oklahoma	Oklahoma	Miami OH	Miami OH	Oklahoma	LSU
LSU	LSU	LSU	LSU	LSU	Oklahoma
USC	Texas	Oklahoma	Oklahoma	Kansas St.	Kansas St.
Michigan	Kansas St.	USC	USC	Florida St.	Florida St.
Ohio St.	USC	Georgia	Georgia	USC	Georgia
Texas	Georgia	Michigan	Michigan	Texas	Michigan
Florida St.	Florida St.	Tennessee	Tennessee	Michigan	USC
Tennessee	Michigan	Ohio St.	Florida St.	Miami OH	Texas
Miami FL	Miami OH	Texas	Ohio St.	Georgia	Miami FL
Kansas St.	Miami FL	Florida St.	Texas	Miami FL	Miami OH

loss to a Big Ten opponent (Iowa). Because the rankings in the table incorporate home-field, this loss is somewhat discounted. The point estimate of the home-field parameter was 0.386 with a standard error of 0.058 – indicating a significant effect due to venue. Thus when two evenly matched teams play, the home team should be expected to win $\Phi(0.386) \approx 65\%$ of the time.

Table 12 gives the post-bowl game top ten teams according to the hybrid model rankings. Also given are the computer rankings based on two common win/loss models (Bradley-Terry and Thurstone-Mosteller) as well as two common independent point-scoring models (Gaussian and Poisson). Because the hybrid model considers both wins *and* points scored, it can be thought of as a compromise between the two aforementioned modeling paradigms. Viewed in this light, it is not surprising that team's hybrid model ranks are similar to their average ranks given by the other models.

Table 12: The hybrid model can be thought of as an average of win/loss and point-scoring models.

Hybrid ($m \uparrow \infty$)	Bradley-Terry	Thurstone-Mosteller	Gaussian	Poisson	Average
LSU	Miami OH	Miami OH	Oklahoma	LSU	1.75
Oklahoma	LSU	LSU	LSU	Oklahoma	2.25
USC	Oklahoma	Oklahoma	Kansas St.	Kansas St.	5.00
Georgia	USC	USC	Florida St.	Florida St.	6.00
Michigan	Georgia	Georgia	USC	Georgia	6.25
Miami FL	Michigan	Michigan	Texas	Michigan	10.50
Miami OH	Tennessee	Tennessee	Michigan	USC	5.00
Florida St.	Ohio St.	Florida St.	Miami OH	Texas	6.50
Texas	Texas	Ohio St.	Georgia	Miami FL	8.25
Kansas St.	Florida St.	Texas	Miami FL	Miami OH	9.50

It is fascinating to notice that despite their Rose Bowl win over Michigan,

USC did not jump into the top two positions after the bowl games. In fact, since LSU and Oklahoma entered the Sugar Bowl as the top two teams, there was no compelling reason to think that they should leave as anything less (especially given the close final score). Critics, especially those in the media who voted USC #1 in the post-season polls, will claim that there must be "something wrong" with a model that failed to recognize USC's obvious brilliance. However, no matter what model was used to rank the teams, USC finished third or worse. In fact, the third-place finish according to our model was their highest ranking. All of this suggests that voters were swayed more by USC's impressive margins of victory than their suspect strength of schedule.

6.1.2 Prediction

Although the *GEE* technique provides a flexible means of parameter estimation in the absence of a fully-specified likelihood, it does not have an adequate goodness-of-fit measure for assessing the quality of the variance parameters, $\boldsymbol{\theta}$. In light of this, we present bowl predictions as an informal, but convincing, measure of fit. Considering only regular season and conference championship games, teams were ranked, and the resulting point estimates were used to predict the post-season bowl games.

Table 14 (in Appendix B) gives the results and hybrid model predictions for 2003-04 bowl games. Predicted scores are given parenthetically. Although bowls tend to match teams of equal abilities, the hybrid model successfully predicted the winners in 18 of 28 games (64%). The predictions also correlate highly with "expert opinions" in the form of published point-spreads. Of the 27 games for which point-spreads were available (no spread was available for the San Francisco Bowl), the predicted point differential from the hybrid model was within 3 points of the published spread for 15 of them (56%).

6.2 2004 Season

While the 2003 regular season produced no undefeated Division I-A teams, the 2004 season produced five! Because these undefeated teams cause degeneracy, win/loss models are forced to rank all five unbeatens (USC, Oklahoma, Auburn, Utah and Boise St.) and teams whose only losses came to those unbeatens (California and Texas) ahead of all other teams. Our hybrid model is not so constrained. See Table 13 for a comparison of model-based rankings after the regular season and conference championship games. As before, a venue parameter was included in all models to account for home-field advantage for those games which were not played at neutral sites. In 2004, the home team would be expected to win roughly 69% of the time in games between two evenly-matched teams, up slightly from the 65% estimate in 2003, but not a statistically significant difference.

Owing to their previously described degeneracy, both the Bradley-Terry and Thurstone-Mosteller models award their top seven rankings to some per-

Table 13: Comparison of 2004 pre-bowl game rankings.

BCS	Hybrid ($m \uparrow \infty$)	Bradley Terry	Thurstone-Mosteller	Gaussian	Poisson
USC	USC	Oklahoma	Oklahoma	USC	USC
Oklahoma	California	USC	USC	California	California
Auburn	Oklahoma	Auburn	Auburn	Oklahoma	Oklahoma
Texas	Auburn	Utah	California	Louisville	VA Tech
California	Texas	Texas	Texas	Utah	Auburn
Utah	VA Tech	California	Utah	Texas	Miami FL
Georgia	Louisville	Boise St.	Boise St.	Miami FL	Texas
VA Tech	Miami FL	VA Tech	LSU	VA Tech	Louisville
Boise St.	Utah	Louisville	Louisville	Auburn	Florida St.
Louisville	LSU	LSU	VA Tech	Boise St.	Utah

mutation of {USC, Oklahoma, Auburn, Utah, Boise St., California and Texas}. While most would have no problem with USC, Oklahoma and Auburn occupying the top three positions, many might be skeptical of Boise St., which, despite its unbeaten record, played a relatively weak schedule that included a one-point home win against 5-6 Brigham Young and a double-overtime victory over 2-9 San Jose St., the only team that 1-10 Washington was able to defeat. Clearly the win/loss rankings leave something to be desired.

As shown in Table 13 point-scoring models rank undefeated Southeastern Conference (SEC) champion Auburn much lower than would be expected. In particular, when an ordinary linear regression, i.e. Gaussian ranking method, is employed, Auburn is ranked ninth – behind three-loss Miami FL and twice-beaten Virginia Tech (the team which Auburn subsequently defeated in the Sugar Bowl). So while win/loss models have difficulty differentiating between unbeaten teams, point-scoring models can be wholly unimpressed with an undefeated season.

The hybrid model presents a reasonable compromise between the two extremes. Unbeaten teams that fattened up against weak competition are ranked behind teams who have lost to more difficult opponents. In this example, once-beaten California, Texas and Louisville, twice-beaten Virginia Tech and LSU, and three-loss Miami FL are ranked in the Top 10, while undefeated Boise St. is not. On the other hand, despite some narrow victories (10-9 over LSU and 21-13 over arch-rival Alabama), Auburn is ranked higher according to the hybrid model than either of the two point-scoring models. It is interesting to note that while most of the discussion revolved around whether Oklahoma or Auburn deserved to face USC in the championship game, California was ignored (due to their 6-point loss at USC). Though few considered California worthy of a title game appearance, many speculated that their perceived slight at being excluded from the other BCS bowls contributed to their uninspired Holiday Bowl loss to Texas Tech.

7 Conclusion

Methods for ranking college football teams use either wins and losses or points scored to determine teams' relative ranks. Both approaches have their deficiencies. A hybrid model, amalgamating both sources of data, was developed to address this concern. Because many well-known paired comparison models are embedded in its general framework, this model is appropriate for a wide-range of underlying (perhaps hidden) models. Simulation studies illustrate its robustness to model mis-specification and its respectable performance even when compared with a known model. The results can be thought of as a compromise to using either purely win/loss data, or solely points scored, that benefits from the advantages of each data type while avoiding their associated deficiencies.

Appendix A: Mathematical Appendix

I. Constructing Derivative Matrices

In order to construct the estimating equations 8, the following derivatives must be computed:

$$\mathbf{D}_t^{(1,1)} = \left[\frac{\partial \boldsymbol{\mu}_t}{\partial \boldsymbol{\xi}}\right],\ \mathbf{D}_t^{(2,1)} = \left[\frac{\partial \mathbf{v}_t}{\partial \boldsymbol{\xi}}\right],\ \text{and } \mathbf{D}_t^{(2,2)} = \left[\frac{\partial \mathbf{v}_t}{\partial \boldsymbol{\theta}}\right].$$

Beginning with $\mathbf{D}_t^{(1,1)}$, recall that the mean response for game t between teams i and j is

$$\boldsymbol{\mu}_t = \begin{pmatrix} \mu_{t,1} \\ \mu_{t,2} \\ \mu_{t,3} \end{pmatrix} = \begin{pmatrix} H[(\beta_i + \gamma_i) - (\beta_j + \gamma_j)] \\ g^{-1}_{\lambda_{[2-\delta_{i,j}]}}(\varphi + \beta_i - \gamma_j) \\ g^{-1}_{\lambda_{[1+\delta_{i,j}]}}(\varphi + \beta_j - \gamma_i) \end{pmatrix},$$

where

$$\mu_{t,1} = \pi_{i,j} = \int_{-\infty}^{\alpha_i - \alpha_j} \frac{e^{zm}\left(1+e^z\right)^{-2m}}{Be(m,m)}\, dz = \int_{-\infty}^{\alpha_i - \alpha_j} h(z)\, dz = H(\alpha_i - \alpha_j),$$

$\alpha_i = \beta_i + \gamma_i$ is the overall ability of team i and

$$g_\lambda^{-1}(z) = \begin{cases} \left(\lambda_{[2-\delta_{i,j}]} z + 1\right)^{1/\lambda_{[2-\delta_{i,j}]}}, & \lambda_{[2-\delta_{i,j}]} \neq 0; \\ \exp(z), & \lambda_{[2-\delta_{i,j}]} = 0. \end{cases}$$

Beginning with $\mathbf{D}_t^{(1,1)}$, the only non-zero derivatives of $\mu_{t,1}$ are those with respect to β_i, β_j, γ_i and γ_j:

$$\frac{\partial \mu_{t,1}}{\partial \beta_i} = h\left((\beta_i + \gamma_i) - (\beta_j + \gamma_j)\right) \qquad \frac{\partial \mu_{t,1}}{\partial \gamma_i} = h\left((\beta_i + \gamma_i) - (\beta_j + \gamma_j)\right)$$
$$\frac{\partial \mu_{t,1}}{\partial \beta_j} = -\left[h\left((\beta_i + \gamma_i) - (\beta_j + \gamma_j)\right)\right] \quad \frac{\partial \mu_{t,1}}{\partial \gamma_j} = -\left[h\left((\beta_i + \gamma_i) - (\beta_j + \gamma_j)\right)\right]$$

In addition to those with respect to β's and γ's, the only non-zero derivatives of $\mu_{t,2}$ and $\mu_{t,3}$ are those with respect to φ and $\boldsymbol{\lambda}$.

$$\frac{\partial \mu_{t,2}}{\partial \varphi} = \left(\mu_{t,2}\right)^{1-\lambda_{[2-\delta_{i,j}]}} \quad \frac{\partial \mu_{t,2}}{\partial \beta_i} = \left(\mu_{t,2}\right)^{1-\lambda_{[2-\delta_{i,j}]}} \quad \frac{\partial \mu_{t,2}}{\partial \gamma_j} = -\left(\mu_{t,2}\right)^{1-\lambda_{[2-\delta_{i,j}]}}$$
$$\frac{\partial \mu_{t,3}}{\partial \varphi} = \left(\mu_{t,3}\right)^{1-\lambda_{[1+\delta_{i,j}]}} \quad \frac{\partial \mu_{t,3}}{\partial \beta_j} = \left(\mu_{t,3}\right)^{1-\lambda_{[1+\delta_{i,j}]}} \quad \frac{\partial \mu_{t,3}}{\partial \gamma_i} = -\left(\mu_{t,3}\right)^{1-\lambda_{[1+\delta_{i,j}]}}$$

$$\frac{\partial \mu_{t,2}}{\partial \lambda_{[2-\delta_{i,j}]}} = \frac{\mu_{t,2}}{\lambda_{[2-\delta_{i,j}]}}\left[\frac{\varphi + \beta_i - \gamma_j}{\mu_{t,2}^{\lambda_{[2-\delta_{i,j}]}}} - \log(\mu_{t,2})\right]$$

$$\frac{\partial \mu_{t,3}}{\partial \lambda_{[1+\delta_{i,j}]}} = \frac{\mu_{t,3}}{\lambda_{[1+\delta_{i,j}]}}\left[\frac{\varphi + \beta_j - \gamma_i}{\mu_{t,3}^{\lambda_{[1+\delta_{i,j}]}}} - \log(\mu_{t,3})\right]$$

We recommend replacing $\mathbf{D}_t^{(2,1)}$ with a zero matrix. Not only does this simplify computations, but it ensures consistent estimation of, among others, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, which is our primary aim. However, if desired, derivatives of $\mathbf{D}_t^{(2,1)}$ exist and are easily found by noting the dependence of $\mathbf{V}_t$ on $\boldsymbol{\mu}_t$ and applying the chain rule for differentiation.

$$\mathbf{D}_t^{(2,1)} = \left[\frac{\partial \mathbf{v}_t}{\partial \boldsymbol{\xi}}\right] = \left[\frac{\partial}{\partial \boldsymbol{\xi}} \begin{pmatrix} \mu_{t,1}(1-\mu_{t,1}) \\ \theta_1(\mu_{t,2})^{\theta_2} \\ \theta_1\theta_3(\mu_{t,2}\mu_{t,3})^{\theta_2/2} \\ \theta_1(\mu_{t,3})^{\theta_2} \end{pmatrix}\right],$$

where

$$\begin{aligned}\frac{\partial \mathbf{v}_t}{\partial \xi_k} &= \begin{pmatrix} (1-2\mu_{t,1})(\partial\mu_{t,1}/\partial\xi_k) \\ \theta_1\theta_2(\mu_{t,2})^{\theta_2-1}(\partial\mu_{t,2}/\partial\xi_k) \\ \theta_1\theta_2\theta_3(\mu_{t,2}\mu_{t,3})^{(\theta_2/2)-1}[\mu_{t,2}(\partial\mu_{t,3}/\partial\xi_k)+\mu_{t,3}(\partial\mu_{t,2}/\partial\xi_k)] \\ \theta_1\theta_2(\mu_{t,3})^{\theta_2-1}(\partial\mu_{t,3}/\partial\xi_k) \end{pmatrix} \\ &= \begin{pmatrix} (1-2\mu_{t,1})(\partial\mu_{t,1}/\partial\xi_k) \\ (v_{t,2}/\mu_{t,2})(\partial\mu_{t,2}/\partial\xi_k) \\ \theta_2[/(\mu_{t,2}\mu_{t,3})][\mu_{t,2}(\partial\mu_{t,3}/\partial\xi_k)+\mu_{t,3}(\partial\mu_{t,2}/\partial\xi_k)] \\ (v_{t,4}/\mu_{t,4})(\partial\mu_{t,4}/\partial\xi_k) \end{pmatrix}\end{aligned}$$

Lastly, the $\mathbf{D}_t^{(2,2)}$ matrix is computed as

$$\begin{aligned}\mathbf{D}_t^{(2,2)} &= \left[\frac{\partial \mathbf{v}_t}{\partial \boldsymbol{\theta}}\right] = \left[\frac{\partial}{\partial \boldsymbol{\theta}} \begin{pmatrix} \mu_{t,1}(1-\mu_{t,1}) \\ \theta_1(\mu_{t,2})^{\theta_2} \\ \theta_1\theta_3(\mu_{t,2}\mu_{t,3})^{\theta_2/2} \\ \theta_1(\mu_{t,3})^{\theta_2} \end{pmatrix}\right] \\ &= \begin{bmatrix} 0 & 0 & 0 \\ (\mu_{t,2})^{\theta_2} & \theta_1(\mu_{t,2})^{\theta_2}\log(\mu_{t,2}) & 0 \\ \theta_3(\mu_{t,2}\mu_{t,3})^{\theta_2/2} & \frac{1}{2}\theta_1\theta_3(\mu_{t,2}\mu_{t,3})^{\theta_2/2}[\log(\mu_{t,2})\log(\mu_{t,3})] & \theta_1(\mu_{t,2}\mu_{t,3})^{\theta_2/2} \\ (\mu_{t,3})^{\theta_2} & \theta_1(\mu_{t,3})^{\theta_2}\log(\mu_{t,3}) & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & 0 \\ v_{t,2}/\theta_1 & v_{t,2}\log(\mu_{t,2}) & 0 \\ v_{t,3}/\theta_1 & v_{t,3}[\log(\mu_{t,2})\log(\mu_{t,3})] & v_{t,3}/\theta_3 \\ v_{t,4}/\theta_1 & v_{t,4}\log(\mu_{t,3}) & 0 \end{bmatrix}.\end{aligned}$$

II. Numerical Optimization

Denote by $\boldsymbol{\zeta}$ the vector of mean and variance parameters, and by $\boldsymbol{G}(\boldsymbol{\zeta})$ the joint estimating equations. Then the Fisher scoring algorithm suggests an iterative scheme at finding the roots of $\boldsymbol{G}(\cdot)$. Beginning with an approximate root, $\boldsymbol{\zeta}^{(k+1)}$, at the *kth* iteration, the update proceeds by performing a Newton-type update with the *expected* hessian used in place of the unknown derivative matrix

$$\boldsymbol{\zeta}^{(k+1)} = \boldsymbol{\zeta}^{(k)} + \mathbb{E}[\dot{\boldsymbol{G}}(\boldsymbol{\zeta})]^{-1}\boldsymbol{G}(\boldsymbol{\zeta}),$$

where the "dot" denotes differentiation with respect to $\boldsymbol{\zeta}$.

We begin by simplifying the notation for $\boldsymbol{G}(\boldsymbol{\zeta})$ by combining the gradient matrices $\mathbf{D}_t^{(\cdot,\cdot)}$ into a single derivative matrix $\boldsymbol{\Delta}_t$, combining the covariance matrices $\mathbf{V}_t$ and $\boldsymbol{\Lambda}_t$ into a single weight matrix $\mathbf{W}_t$ and by considering a single residual vector $\mathbf{r}_t$ consisting of both raw and "squared" residuals.

$$\begin{aligned}\boldsymbol{G}(\boldsymbol{\zeta}) &= \sum_{t=1}^{n} \begin{pmatrix} \mathbf{D}_t^{(1,1)} & \mathbf{0} \\ \mathbf{D}_t^{(2,1)} & \mathbf{D}_t^{(2,2)} \end{pmatrix}^T \begin{pmatrix} \mathbf{V}_t & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_t \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{u}_t - \boldsymbol{\mu}_t \\ \mathbf{v}_t - \boldsymbol{\nu}_t \end{pmatrix} \\ &= \sum_{t=1}^{n} \boldsymbol{\Delta}_t^T \mathbf{W}_t \left[\mathbf{r}_t - \mathbb{E}(\mathbf{r}_t)\right]\end{aligned}$$

Differentiation is simplified by using the Fisher scoring algorithm, as only the *expected* derivatives are needed. Although the derivative itself consists of three terms, two of them have expectation zero, as they are constant matrices post-multiplied by $[\mathbf{r}_t - \mathbb{E}(\mathbf{r}_t)]$, which, clearly, has expectation zero.

$$\begin{aligned}\dot{\boldsymbol{G}}(\boldsymbol{\zeta}) &= \frac{d}{d\boldsymbol{\zeta}} \left\{ \sum_{t=1}^{n} \boldsymbol{\Delta}_t^T \mathbf{W}_t \left[\mathbf{r}_t - \mathbb{E}(\mathbf{r}_t)\right] \right\} \\ &= \sum_{t=1}^{n} \left\{ \frac{d\boldsymbol{\Delta}_t}{d\boldsymbol{\zeta}}^T \mathbf{W}_t \left[\mathbf{r}_t - \mathbb{E}(\mathbf{r}_t)\right] + \boldsymbol{\Delta}_t^T \frac{d\mathbf{W}_t}{d\boldsymbol{\zeta}} \left[\mathbf{r}_t - \mathbb{E}(\mathbf{r}_t)\right] + \boldsymbol{\Delta}_t^T \mathbf{W}_t \frac{d}{d\boldsymbol{\zeta}} \left[\mathbf{r}_t - \mathbb{E}(\mathbf{r}_t)\right] \right\} \\ \mathbb{E}[\dot{\boldsymbol{G}}(\boldsymbol{\zeta})] &= -\sum_{t=1}^{n} \boldsymbol{\Delta}_t^T \mathbf{W}_t \boldsymbol{\Delta}_t,\end{aligned}$$

where the last line follows because $\boldsymbol{\Delta}_t = [d\mathbb{E}(\mathbf{r}_t)/d\boldsymbol{\zeta}]$, by definition. Finally,

$$\begin{aligned}\boldsymbol{\zeta}^{(k+1)} &= \boldsymbol{\zeta}^{(k)} + \mathbb{E}[\dot{\boldsymbol{G}}(\boldsymbol{\zeta})]^{-1}\boldsymbol{G}(\boldsymbol{\zeta}) \\ &= \boldsymbol{\zeta}^{(k)} + \left[\sum_{t=1}^{n} \boldsymbol{\Delta}_t^T \mathbf{W}_t \boldsymbol{\Delta}_t\right]^{-1} \sum_{t=1}^{n} \boldsymbol{\Delta}_t^T \mathbf{W}_t \left[\mathbf{r}_t - \mathbb{E}(\mathbf{r}_t)\right],\end{aligned}$$

where the quantities on the right hand side, $\boldsymbol{\Delta}_t$, $\mathbf{W}_t$ and $\mathbb{E}(\mathbf{r}_t)$ are evaluated at $\boldsymbol{\zeta}^{(k)}$.

Appendix B: 2003-04 NCAA Bowl Game Predictions

Table 14: Results and hybrid model predictions (parenthetically) for 2003-04 bowl games. Correct bowl predictions are in bold face.

Bowl Game	Team	Score	Team	Score
New Orleans Bowl	Memphis	27 (24)	North Texas	17 (21)
GMAC Bowl	Miami OH	49 (48)	Louisville	28 (20)
Tangerine Bowl	North Carolina State	56 (38)	Kansas	26 (31)
Fort Worth Bowl	Boise State	34 (31)	TCU	31 (20)
Las Vegas Bowl	Oregon State	55 (24)	New Mexico	14 (28)
Hawaii Bowl	Hawaii	54 (41)	Houston	48 (33)
Motor City Bowl	Bowling Green	28 (26)	Northwestern	24 (18)
Insight Bowl	California	52 (29)	Virginia Tech	49 (28)
Continental Tire Bowl	Virginia	23 (24)	Pittsburgh	16 (25)
Alamo Bowl	Nebraska	17 (24)	Michigan State	3 (19)
Houston Bowl	Texas Tech	38 (47)	Navy	14 (28)
Holiday Bowl	Washington State	28 (21)	Texas	20 (36)
Silicon Valley Bowl	Fresno State	17 (16)	UCLA	9 (18)
Music City Bowl	Auburn	28 (26)	Wisconsin	14 (17)
Sun Bowl	Minnesota	31 (35)	Oregon	30 (30)
Liberty Bowl	Utah	17 (23)	Southern Miss	0 (17)
Independence Bowl	Arkansas	27 (33)	Missouri	14 (23)
San Francisco Bowl	Boston College	35 (27)	Colorado State	21 (33)
Outback Bowl	Iowa	37 (19)	Florida	17 (23)
Gator Bowl	Maryland	41 (25)	West Virginia	7 (20)
Capital One Bowl	Georgia	34 (21)	Purdue	27 (13)
Rose Bowl	Southern Cal	28 (33)	Michigan	14 (26)
Orange Bowl	Miami FL	16 (20)	Florida State	14 (19)
Cotton Bowl	Mississippi	31 (27)	Oklahoma State	28 (30)
Peach Bowl	Clemson	27 (20)	Tennessee	14 (25)
Fiesta Bowl	Ohio State	35 (17)	Kansas State	28 (24)
Humanitarian Bowl	Georgia Tech	52 (25)	Tulsa	10 (18)
Sugar Bowl	LSU	21 (21)	Oklahoma	14 (25)

References

[1] Agresti, A. (1990), "Categorical Data Analysis," John Wiley & Sons, Inc., New York.

[2] Billingsley, R. (2004), "Discussion – Statistics and the College Football Championship," *The American Statistician*, 58, 190.

[3] Box, G. E. P. and D. R. Cox (1964), "An analysis of transformations," *Journal of the Royal Statistical Society. Series B*, 26, 211-252.

[4] Bradley, R. A. and M. E. Terry (1952), "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons," *Biometrika*, 39, 324-345.

[5] Colley, W. (2004), "Discussion – Statistics and the College Football Championship," *The American Statistician*, 58, 191-192.

[6] David, H. A. (1988), "The Method of Paired Comparisons," Charles Griffin & Company Ltd., London, second edition.

[7] Davidson, R. R. (1969), "On a relationship between two representations of a model for paired comparisons," *Biometrics*, 25, 597-599.

[8] Dixon, M. J. and S. G. Coles (1997), "Modelling association football scores and inefficiencies in the football betting market," *Applied Statistics*, 46, 265-280.

[9] Harville, D. A. (1977), "The Use of Linear-Model Methodology to Rate High School or College Football Teams," *Journal of the American Statistical Association*, 72, 287-289.

[10] Harville, D. A. (1980), "Predictions for National Football League Games via Linear-Model Methodology," *Journal of the American Statistical Association*, 75, 516-524.

[11] Harville, D. (2004), "Discussion – Statistics and the College Football Championship," *The American Statistician*, 58, 187-189.

[12] Heyde, C. C. (1980), "Quasi-Likelihood and Its Application : A General Approach to Optimal Parameter Estimation," Springer.

[13] Lehmann, E. L. (1953), "The Power of Rank Tests," *The Annals of Mathematical Statistics*, 24, 23-43.

[14] Liang, K-Y and S. L. Zeger (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13-22.

[15] Liang, K-Y and S. L. Zeger and B. Qaqish (1992), "Multivariate Regression Analyses for Categorical Data," *Journal of the Royal Statistical Society. Series B (Methodological)*, 54, 3-40.

[16] Massey, K. (2004), "Discussion – Statistics and the College Football Championship," *The American Statistician*, 58, 185-187.

[17] McCullagh, P. (1983), "Quasi-likelihood functions," *The Annals of Statistics*, 11, 59-67.

[18] Mease, D. (2003), "A Penalized Maximum Likelihood Approach for the Ranking of College Football Teams Independent of Victory Margins," *The American Statistician*, 57, 241-248.

[19] Mease, D. (2004), "Discussion – Statistics and the College Football Championship," *The American Statistician*, 58, 192-194.

[20] Mosteller, F. (1951), "Remarks on the Method of Paired Comparisons: I. The Least Squares Solution Assuming Equal Standard Deviations and Equal Correlations," *Psychometrika*, 16, 3-9.

[21] Prentice, R. (1976), "A Generalization of the Probit and Logit Methods for Dose Response Curves," *Biometrics*, 32, 761-768.

[22] Stern, H. (2004), "Statistics and the College Football Championship," *The American Statistician*, 58, 179-185.

[23] Thurstone, L. L. (1927), "A Law of Comparative Judgment," *Psychological Review*, 34, 273-286.

[24] Tweedie, M. C. K. (1981), "An Index which Distinguishes between some Important Exponential Families," *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, 579-604.

[25] Wedderburn, R. W. M. (1974), Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method," *Biometrika*, 61, 439-447.

[26] White, H. (1982), "Maximum likelihood estimation of misspecified models.," *Econometrica*, 50, 1-26.