Math. Appl. **7** (2018), 71–84
DOI: 10.13164/ma.2018.07

# HIERARCHICAL BAYESIAN BRADLEY–TERRY FOR APPLICATIONS IN MAJOR LEAGUE BASEBALL

GABRIEL C. PHELAN AND JOHN T. WHELAN

*Abstract.* A common problem faced in statistical inference is drawing conclusions from paired comparisons, in which two objects compete and one is declared the victor. A probabilistic approach to such a problem is the Bradley–Terry model [5, 20], first studied by Zermelo in 1929 and rediscovered by Bradley and Terry in 1952. One obvious area of application for such a model is sporting events, and in particular Major League Baseball. With this in mind, we describe a hierarchical Bayesian version of Bradley–Terry suitable for use in ranking and prediction problems, and compare results from these application domains to standard maximum likelihood approaches. Our Bayesian methods outperform the MLE-based analogues, while being simple to construct, implement, and interpret.

## 1. Background

### 1.1. The Bradley–Terry model

Amongst a set of $N$ objects, which we will call "teams", the Bradley–Terry model associates a "strength" $\pi_i \in \mathbb{R}^+$ to each team and assumes that

$$\mathbb{P}\{\text{team } i \text{ defeats team } j\} = \frac{\pi_i}{\pi_i + \pi_j},$$

where $i, j \in \{1, 2, \dots, N\}$. If we define $V_{ij}$ to be the number of times in a season that team $i$ defeats team $j$, and $n_{ij}$ the number of games between them, an entire season can be described in terms of the probability mass function

$$p(\mathbf{V} \mid \boldsymbol{\pi}) = \prod_{i=1}^{N-1} \prod_{j=i+1}^{N} \binom{n_{ij}}{V_{ij}} \left(\frac{\pi_i}{\pi_i + \pi_j}\right)^{V_{ij}} \left(\frac{\pi_j}{\pi_i + \pi_j}\right)^{V_{ji}},$$

where $\mathbf{V}_{N\times N}$ is a matrix of records and $\boldsymbol{\pi}_{N\times 1}$ a vector of team strengths.

### 1.2. The Bayesian approach

In practice, it is often the case that $\mathbf{V}$ is known and the goal is to perform inference on the strengths $\boldsymbol{\pi}$. In a frequentist interpretation, this would proceed by maximum likelihood estimation, for which there is no closed-form solution but a number of numerical algorithms have been suggested [9,11,12]. However, as noted by Ford [11], pathologies may exist under this approach. Maximum likelihood ratios of some teams' strengths may be zero, infinite, or undetermined, leading to 0, 1, or

*MSC (2010):* primary 62F15; secondary 62F07.
*Keywords:* Bayesian inference, Major League Baseball, MCMC, probabilistic modeling.

undetermined probabilities. The conditions under which these pathologies arise has been studied by various researchers [1,6,17]. Settings like Major League Baseball (unlike, for instance, college football) are practically guaranteed immunity from these issues [6], but taking a Bayesian approach fully guards against them.

In a Bayesian interpretation of the model, we place a prior distribution $p(\boldsymbol{\pi})$ over the strengths, avoiding the aforementioned difficulties. It also affords us the use of full Bayesian inference, in which we compute a posterior distribution $p(\boldsymbol{\pi} \mid \mathbf{V}) \propto p(\mathbf{V} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi})$ over the team strengths in light of the records. $p(\boldsymbol{\pi} \mid \mathbf{V})$ captures and quantifies all of our uncertainty conditional on our knowledge. Various takes on Bayesian Bradley–Terry have been studied [7,8,10,13,19]; a common concern is the choice of prior distribution. In this regard we draw on the work of Whelan, who advocates for two classes of distributions in particular [19].

### 1.3. Desiderata and choice of prior distribution

In specifying a prior for Bayesian Bradley–Terry, one approach is to require that it adheres to a list of *desiderata*. These formalize our intuition about how the model should behave under a suitable prior distribution. We adopt the desiderata of Whelan [19], which, with applications to ranking systems in mind, attempts to construct priors that make no unfair distinctions between individual teams. Roughly speaking, this means that we should choose a prior, possibly over a transformed parameter $\mathbf{T}(\boldsymbol{\pi})$, that:

(1) Ensures invariance under the interchange of teams.
(2) Ensures invariance under the interchange of winning and losing.
(3) Ensures invariance under the elimination of teams.
(4) Is a proper (normalizable) prior.

Most families of prior distribution fail at least one of these requirements, but there are two families that are known to satisfy all four [19]. The first is a separable Gaussian distribution in the log-strengths with 0 mean and common variance:

$$\lambda_i \sim \mathcal{N}(0, \sigma^2), \qquad \lambda_i = T_1(\pi_i) = \log \pi_i.$$

The second is a Beta distribution in what can be interpreted as the probability of a particular team defeating an "imaginary opponent" of unit strength, with common scale and shape parameters:

$$\zeta_i \sim \beta(\eta, \eta), \qquad \zeta_i = T_2(\pi_i) = \frac{\pi_i}{1 + \pi_i}.$$

It is clear based on these definitions that $\boldsymbol{\lambda} \in \mathbb{R}^N$ and $\boldsymbol{\zeta} \in (0,1)^N$. We consider only these two families of prior distribution, denoted as $I_{\mathcal{N}}$ and $I_\beta$, guaranteeing the desiderata are satisfied. As we proceed, we will estimate posterior densities using Markov Chain Monte Carlo (MCMC), which is known to prefer unconstrained parameter spaces [18]. It is therefore helpful to transform the prior on $\boldsymbol{\zeta}$ into $\boldsymbol{\lambda}$-space. This parameterization will also ease any comparisons we may wish to make between the two priors.

**Lemma 1.1.** *$\lambda_i \mid I_\beta$ has the generalized logistic distribution of the third kind, which we denote as $\lambda_i \mid I_\beta \sim \mathrm{GL}_3(\eta)$.*

*Proof.* For each $\zeta_i$ we have

$$p(\zeta_i \mid I_\beta) \propto [\zeta_i(1-\zeta_i)]^{\eta-1}$$

and

$$\zeta_i = \frac{e^{\lambda_i}}{1+e^{\lambda_i}} = \text{logistic}(\lambda_i).$$

Thus,

$$\left|\frac{\mathrm{d}\zeta_i}{\mathrm{d}\lambda_i}\right| = \frac{e^{\lambda_i}}{(1+e^{\lambda_i})^2},$$

and by change of variables,

$$\begin{aligned} p(\lambda_i \mid I_\beta) &\propto \left[\frac{e^{\lambda_i}}{1+e^{\lambda_i}}\left(1-\frac{e^{\lambda_i}}{1+e^{\lambda_i}}\right)\right]^{\eta-1} \frac{e^{\lambda_i}}{(1+e^{\lambda_i})^2} \\ &\propto \left[\frac{e^{\lambda_i}}{(1+e^{\lambda_i})^2}\right]^{\eta-1} \frac{e^{\lambda_i}}{(1+e^{\lambda_i})^2} \\ &\propto \left[\frac{e^{\lambda_i}}{(1+e^{\lambda_i})^2}\right]^{\eta}, \end{aligned}$$

which is the form of a GL$_3$ distribution. □

This family of distributions is well-known, the most general form of which is given by

$$p(\lambda_i \mid \varphi,\eta,\gamma) = \frac{1}{\mathrm{B}(\gamma,\eta)}\left(\frac{\varphi e^{-\varphi\eta\lambda_i}}{(1+e^{-\varphi\lambda_i})^{\eta+\gamma}}\right),$$

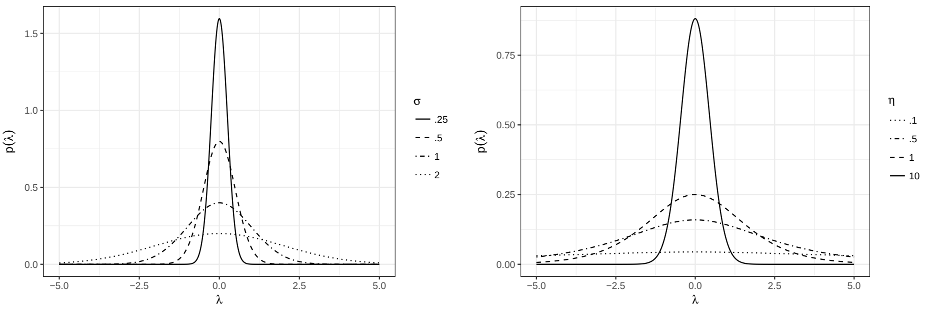
where $\mathrm{B}(\gamma,\eta) = \frac{\Gamma(\gamma)\Gamma(\eta)}{\Gamma(\gamma+\eta)}$ is the Beta function. We say that $\lambda_i \sim \mathrm{GL}(\varphi,\eta,\gamma)$. A complete overview is given in [14], where the following useful properties are shown:

$$\begin{aligned} \mathbb{E}[\lambda_i \mid \varphi,\eta,\gamma] &= \frac{1}{\varphi}[\psi(\gamma)-\psi(\eta)], \\ \mathbb{V}[\lambda_i \mid \varphi,\eta,\gamma] &= \frac{1}{\varphi^2}[\psi'(\gamma)+\psi'(\eta)], \end{aligned}$$

for $\psi(\cdot)$ and $\psi'(\cdot)$ the digamma and trigamma functions respectively. It is evident that the GL$_3(\eta)$ distribution is equivalent to the $\mathrm{GL}(1,\eta,\eta)$ distribution, and so we immediately have

$$\mathbb{E}[\lambda_i \mid I_\beta] = 0 \qquad \text{and} \qquad \mathbb{V}[\lambda_i \mid I_\beta] = 2\psi'(\eta).$$

For reference, $\eta = 1$, which would produce a uniform prior in $\zeta_i$, corresponds to a Gaussian-like prior with variance $2\psi'(1) \approx 3.3$ in $\lambda_i$. With $p(\lambda_i \mid I_\beta)$ established, we restrict our discussion to $\boldsymbol{\lambda}$-space from here onward.



(a) $p(\lambda_i \mid I_{\mathcal{N}})$ for different values of $\sigma$ (we use $\sigma$ due to R's parameterization of the Gaussian distribution).

(b) $p(\lambda_i \mid I_\beta)$ for different values of $\eta$. This family is known as the type III generalized logistic distribution.

**Figure 1.** Prior distributions in $\lambda_i$ that satisfy the desiderata described in Subsection 1.3

# 2. The model

## 2.1. Motivation

Recent advances in Bayesian computation mean that for models of a reasonable size, we no longer have to restrict ourselves to using conjugate priors, Laplace approximations, or hand-tuned hyperparameters. Modern MCMC methods such as Hamiltonian Monte Carlo (HMC) allow for rich hierarchical models that can be implemented easily in probabilistic programming languages such as Stan [18]. These efficient MCMC algorithms mean we can integrate over our uncertainty in the form of expectations, which, from a Bayesian perspective, are preferable to optimization-based procedures [3]. Thus, we take the stance that hierarchical models are the ideal way to approach Bayesian inference, especially now that the computational tools to exploit such models exist. The hallmark of hierarchical modeling is to place additional priors over hyperparameters of interest. In our case, this will mean deriving a suitable prior distribution for $\eta$ or $\sigma$.

The model is informed by our applications, which are discussed in the final section. We wish to exploit the advantages of Bayesian inference, while remaining objective in our treatment of individual teams. This will lead us to a weakly-informative hierarchy that encompasses both the objective and subjective approaches to the Bayesian paradigm. As seen in section 2.3, we incorporate prior information that pertains to the league as a whole at the uppermost layer of the model, but insist that teams be evaluated only based on their performance against one another. Bradley–Terry provides an ideal framework for such a philosophy since it relies solely on the $\binom{30}{2}$ head-to-head records. This affords us a rich framework for handling uncertainty, while keeping the model primarily data-driven.

### 2.2. Likelihood

We restate the likelihood, this time in terms of $\boldsymbol{\lambda}$:

$$\begin{aligned} p(\mathbf{V} \mid \boldsymbol{\lambda}) &= \prod_{i=1}^{N-1} \prod_{j=i+1}^{N} \binom{n_{ij}}{V_{ij}} \left( \frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j}} \right)^{V_{ij}} \left( \frac{e^{\lambda_j}}{e^{\lambda_i} + e^{\lambda_j}} \right)^{V_{ji}} \\ &\propto \prod_{i=1}^{N} \prod_{j=1}^{N} \left( \frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j}} \right)^{V_{ij}}. \end{aligned}$$

### 2.3. Choosing between $I_\beta$ and $I_\mathcal{N}$

One initial complication is that there is no "obvious" way to choose between $I_\mathcal{N}$ and $I_\beta$. Figure 1 illustrates the similarity between the densities of the two families, a fact that is discussed in [14]. $I_\mathcal{N}$ is attractive on the grounds that Gaussians are often easy to work with, but we can also motivate its usage by appealing to the principle of maximum entropy. Note that the variance of the $\lambda_i$ should be prior-independent; that is $\mathbb{V}[\lambda_i \mid I_\mathcal{N}] = \mathbb{V}[\lambda_i \mid I_\beta]$. Thus the first two moments are fixed at 0 and $\sigma^2$ respectively. It is well known that under these circumstances the differential entropy

$$\mathbb{H}(\lambda_i) = -\int_{\mathbb{R}} p(\lambda_i) \log p(\lambda_i) \, \mathrm{d}\lambda_i$$

is maximized for $\lambda_i \sim \mathcal{N}(0, \sigma^2)$. Equivalently, this maximizes the relative entropy[1] under a uniform measure. So, in this sense $I_\mathcal{N}$ carries less information about $\lambda_i$, and we select it for this reason.

### 2.4. Hyperparameters and hyperpriors

In hierarchical modeling, one foregoes the hand-tuning of hyperparameters and instead builds another layer of prior distributions into the model, called hyperpriors. This creates the added difficulty of determining good hyperpriors to use. Unfortunately, there is rarely a principled approach to determining this final layer of the model (we again reject the use improper priors, to ensure the stability of MCMC methods [4]). Often, the construction is made through some use of maximum likelihood estimation.

We will take the approach of using prior seasons' data to produce a hyperprior for the hyperparameter $\sigma$. In principle, one could carry out MCMC on a previous season with a weakly informative hyperprior, and produce a marginalized posterior for $\sigma$ which could serve as a hyperprior for a subsequent season. We opt for a computationally-simpler approach based on an approximate maximum a posteriori expansion. The result will be a point estimate $\widehat{\sigma}$ with an associated variance $\widehat{\varsigma}$. Rather than a Gaussian approximation for the posterior on $\sigma$ (which would extend to negative values of $\sigma$), we instead choose a Gamma distribution (using the shape and rate parameterization) with the same mean $\widehat{\sigma}$ and variance $\widehat{\varsigma}$, i.e., $\sigma \sim \Gamma\left(\frac{\widehat{\sigma}^2}{\widehat{\varsigma}}, \frac{\widehat{\sigma}}{\widehat{\varsigma}}\right)$.

[1] The relative entropy is defined to be $\mathbb{H}_R(x) = -\int_\Omega p(x) \log\left(\frac{p(x)}{m(x)}\right) \mathrm{d}x$, where $\Omega$ is the support space and $m$ is an invariant measure, meaning it transforms like $p$ under a change of variables.

Formally assuming a uniform prior on $\sigma$, the log-posterior can be written

$$\ell = \log p(\boldsymbol{\lambda}, \sigma \mid \mathbf{V}) = \sum_{i=1}^{N} \left\{ \sum_{j=1}^{N} V_{ij} \left[ \lambda_i - \log(e^{\lambda_i} + e^{\lambda_j}) \right] - \frac{\lambda_i^2}{2\sigma^2} \right\} - N \log \sigma + \text{const.}$$

and the MAP point can be found by taking the partial derivatives

$$\frac{\partial \ell}{\partial \sigma} = \sigma^{-3} \sum_{i=1}^{N} \lambda_i^2 - N\sigma^{-1} \qquad \text{and} \qquad \frac{\partial \ell}{\partial \lambda_i} = \sum_j \left( V_{ij} - n_{ij} \frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j}} \right) - \sigma^{-2} \lambda_i.$$

Setting these to zero and rearranging produces the coupled MAP equations

$$\widehat{\sigma} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \widehat{\lambda}_i^2} \qquad \text{and} \qquad \widehat{\lambda}_i = \log \left\{ \frac{V_i - \widehat{\lambda}_i / \widehat{\sigma}^2}{\sum_{j=1}^{N} \left( n_{ij} \Big/ \left[ e^{\widehat{\lambda}_i} + e^{\widehat{\lambda}_j} \right] \right)} \right\}$$

where $V_i = \sum_{j=1}^{N} V_{ij}$ is the total number of games won by team $i$. These MAP equations could be solved iteratively by a method analogous to that of Ford [11], but we make the assumption that, with each team playing 162 games in a full season, $V_i$ is large compared to $\widehat{\lambda}_i / \widehat{\sigma}^2$ and we can use the maximum likelihood estimates $\left\{ \widehat{\lambda}_i^{\text{MLE}} \right\}$, determined by iteratively solving

$$\widehat{\lambda}_i^{\text{MLE}} = \log \left\{ \frac{V_i}{\sum_{j=1}^{N} \left( n_{ij} \Big/ \left[ e^{\widehat{\lambda}_i^{\text{MLE}}} + e^{\widehat{\lambda}_j^{\text{MLE}}} \right] \right)} \right\} \tag{2.1}$$

in place of the $\left\{ \widehat{\lambda}_i \right\}$, and writing

$$\widehat{\sigma} \approx \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left( \widehat{\lambda}_i^{\text{MLE}} \right)^2}. \tag{2.2}$$

Note that the maximum likelihood equations (2.1) only determine the $\left\{ \widehat{\lambda}_i^{\text{MLE}} \right\}$ up to an overall additive constant, which we set by requiring $\sum_{i=1}^{N} \widehat{\lambda}_i^{\text{MLE}} = 0$. The variance $\widehat{\varsigma}$ can be estimated by considering the matrix of second derivatives

$$\mathbf{H} = \begin{pmatrix} -\frac{\partial^2 \ell}{\partial^2 \sigma} & \left\{ -\frac{\partial^2 \ell}{\partial \sigma \, \partial \lambda_j} \right\} \\ \left\{ -\frac{\partial^2 \ell}{\partial \lambda_i \, \partial \sigma} \right\} & \left\{ -\frac{\partial^2 \ell}{\partial \lambda_i \, \partial \lambda_j} \right\} \end{pmatrix}_{\sigma = \widehat{\sigma};\, \boldsymbol{\lambda} = \widehat{\boldsymbol{\lambda}}}$$

and defining $\widehat{\varsigma} = \left[ \mathbf{H}^{-1} \right]_{\sigma\sigma}$. The second derivatives are

$$\begin{aligned} H_{\sigma\sigma} &= 3\widehat{\sigma}^{-4} \sum_{i=1}^{N} \widehat{\lambda}_i^2 - N\widehat{\sigma}^{-2} = 2N\widehat{\sigma}^{-2}, \\ H_{\sigma\lambda_i} &= 2\widehat{\lambda}_i \widehat{\sigma}^{-3}, \end{aligned}$$

$$H_{\lambda_i\lambda_j} = \delta_{ij}\left(\widehat{\sigma}^{-2} + \sum_{k=1}^{N} n_{ik}\frac{e^{\widehat{\lambda}_i+\widehat{\lambda}_k}}{e^{\widehat{\lambda}_i}+e^{\widehat{\lambda}_k}}\right) - n_{ij}\frac{e^{\widehat{\lambda}_i+\widehat{\lambda}_j}}{e^{\widehat{\lambda}_i}+e^{\widehat{\lambda}_j}}.$$

In practice, given a whole season's worth of data, we don't need to invert the full matrix; the terms involving $\{n_{ij}\}$ will dominate to leading order, and we can approximate the matrix as block-diagonal[2] to write

$$\widehat{\varsigma} \approx \frac{1}{H_{\sigma\sigma}} = \frac{\widehat{\sigma}^2}{2N}.$$

Following this procedure, we arrive at a $\Gamma\left(2N, (2N)/\widehat{\sigma}^2\right)$ hyperprior. In keeping with the Bayesian philosophy, we avoid using the data from the season to be modeled in setting that season's hyperprior. Instead, we account for any trends in league parity by constructing the hyperprior using the previous season's data. Since Major League Baseball has consisted of 30 teams throughout the seasons we model, the hyperprior is $\Gamma\left(60, 60/\widehat{\sigma}^2\right)$ where $\widehat{\sigma}^2$ is the estimated variance of the $\{\lambda_i\}$ during the previous season.

**Table 1.** The estimates $\widehat{\sigma}$ and $\sqrt{\widehat{\varsigma}}$ for the 2010 - 2016 seasons, computed according to the above prescription. We construct the hyperprior for a a given season using the estimates from the previous season. $\sqrt{\widehat{\varsigma}}$ can be interpreted as one standard deviation of uncertainty in $\widehat{\sigma}$.

| year | $\widehat{\sigma}$ | $\sqrt{\widehat{\varsigma}}$ |
|---|---|---|
| 2010 | 0.264 | 0.034 |
| 2011 | 0.267 | 0.034 |
| 2012 | 0.316 | 0.041 |
| 2013 | 0.289 | 0.037 |
| 2014 | 0.235 | 0.030 |
| 2015 | 0.274 | 0.035 |
| 2016 | 0.262 | 0.034 |

### 2.5. Full model

The following describes the full Bayesian hierarchical model in matrix notation:

$$\begin{aligned} \sigma &\sim \Gamma\left(2N, \frac{2N}{\widehat{\sigma}^2}\right) \\ \boldsymbol{\lambda} \mid \sigma &\sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \\ \mathbf{V} \mid \boldsymbol{\lambda} &\sim \text{Bradley–Terry}(\exp\{\boldsymbol{\lambda}\}). \end{aligned}$$

[2] Note that the $\widehat{\sigma}^{-2}$ is important for inversion of the other block, which is otherwise degenerate since $\sum_{j=1}^{N} H_{\lambda_i\lambda_j} = \widehat{\sigma}^{-2}$.

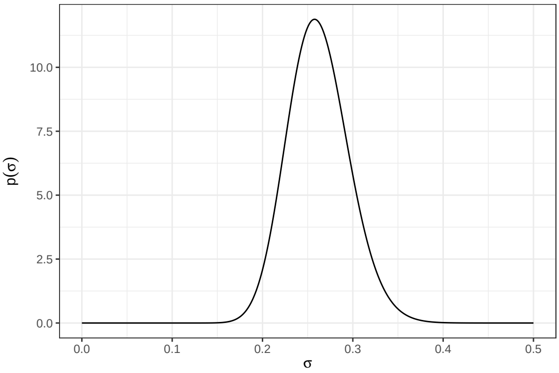


**Figure 2.** The hyperprior $p(\sigma)$ used to model the 2017 season, generated according to estimates from the 2016 season.

**Table 2.** The estimates for $\left\{\widehat{\lambda}_i^{\text{MLE}}\right\}$ and $\left\{\widehat{\lambda}_i^{\text{MAP}}\right\}$ during the 2017 season, computed according to Ford's iterative algorithm. Note that the values are indeed close, justifying the simplification made in equation (2.2). Teams in **boldface** made the postseason.

| | team | $\widehat{\lambda}_i^{\text{MLE}}$ | $\widehat{\lambda}_i^{\text{MAP}}$ | | team | $\widehat{\lambda}_i^{\text{MLE}}$ | $\widehat{\lambda}_i^{\text{MAP}}$ |
|---|---|---|---|---|---|---|---|
| 1 | **CLE** | 0.52 | 0.47 | 16 | SEA | −0.02 | −0.03 |
| 2 | **HOU** | 0.51 | 0.46 | 17 | TEX | −0.03 | −0.04 |
| 3 | **LAN** | 0.50 | 0.45 | 18 | TOR | −0.04 | −0.06 |
| 4 | **BOS** | 0.33 | 0.29 | 19 | BAL | −0.06 | −0.07 |
| 5 | **NYA** | 0.29 | 0.26 | 20 | OAK | −0.09 | −0.10 |
| 6 | **WAS** | 0.26 | 0.25 | 21 | PIT | −0.18 | −0.15 |
| 7 | **ARI** | 0.26 | 0.25 | 22 | MIA | −0.19 | −0.15 |
| 8 | **CHN** | 0.19 | 0.19 | 23 | SDN | −0.25 | −0.22 |
| 9 | **MIN** | 0.13 | 0.12 | 24 | CHA | −0.27 | −0.26 |
| 10 | **COL** | 0.12 | 0.11 | 25 | ATL | −0.30 | −0.26 |
| 11 | MIL | 0.07 | 0.08 | 26 | CIN | −0.33 | −0.29 |
| 12 | TBA | 0.05 | 0.03 | 27 | NYN | −0.34 | −0.29 |
| 13 | LAA | 0.03 | 0.02 | 28 | DET | −0.34 | −0.33 |
| 14 | KCA | 0.02 | 0.02 | 29 | SFN | −0.41 | −0.36 |
| 15 | SLN | 0.00 | 0.01 | 30 | PHI | −0.43 | −0.37 |

## 3. Applications to Major League Baseball

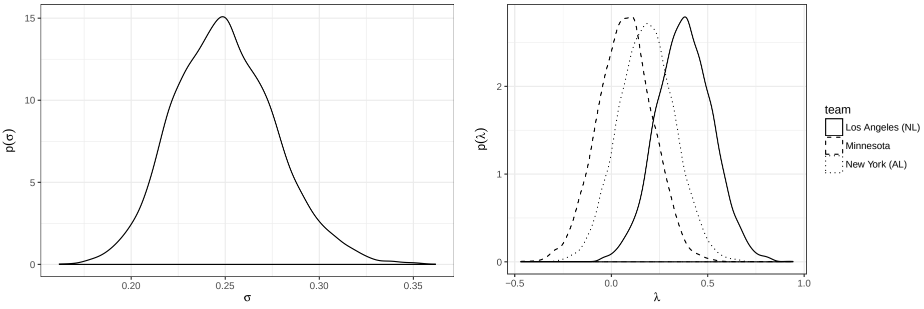
### 3.1. A word on data acquisition and computation

The present authors have obtained all data from `baseballreference.com` [2] and `retrosheet.org` [16]. Modifications of data and numerical computations were performed in the R and Stan programming languages [15,18]. Stan is a probabilistic programming language for performing HMC. HMC allows for efficient MCMC, capable of computing marginal posterior distributions for complex Bayesian models.

Stan also permits the drawing of samples from the posterior predictive distribution. For more information about Stan, see [18]; for more information about HMC, see [4].

## 3.2. Application I: Ranking systems

The first application we present is that of a ranking system based on the log-strengths. Such a system could generate weekly or monthly rankings more nuanced than that provided by simple win-loss comparisons. Much of our motivation for treating the teams objectively is related to this application; a reliable ranking system should be based on team performance alone. The Bayesian approach permits us to assign ranks based on $\mathbb{E}[\lambda_i|\mathbf{V}]$, which integrates over possible outcomes rather than finding an optimum based on the data. Table 3 shows the final rankings from the 2017 season, with teams in **boldface** having made the postseason. Note that teams may be compared via the distance between their respective log-strengths (which correspond to ratios in $\boldsymbol{\pi}$-space). In table 4, we compare these results to those found by maximum likelihood estimation. This aptly illustrates the effect of the Bayesian approach. The prior serves as a regularizer and promotes shrinkage, protecting against over-fitting. Unsurprisingly, $\mathbb{E}[\lambda_i \mid \mathbf{V}]$ is more correlated with a team's true record than is $\widehat{\lambda}_i^{\text{MLE}}$.



(a) $p(\sigma \mid \mathbf{V})$

(b) $p(\lambda_{\text{LAN}} \mid \mathbf{V})$ vs. $p(\lambda_{\text{MIN}} \mid \mathbf{V})$ vs. $p(\lambda_{\text{NYA}} \mid \mathbf{V})$

**Figure 3.** The left image shows the marginal posterior density for $\sigma$ during the 2017 season. As expected, it looks much like our informative hyperprior. The right image compares the marginal posterior densities of three postseason teams from 2017. The superiority of the NL-champion Dodgers is clear.

## 3.3. Application II: Predictive modeling

Bayesian probability offers a particularly elegant way of handling prediction. For our model, the posterior predictive distribution is given by

$$p(\tilde{\mathbf{V}} \mid \mathbf{V}) = \int_{\mathbb{R}^N} p(\tilde{\mathbf{V}} \mid \boldsymbol{\lambda}) p(\boldsymbol{\lambda} \mid \mathbf{V}) \mathrm{d}\boldsymbol{\lambda},$$

which integrates over all uncertainty in the model and gives a distribution over unobserved data $\tilde{\mathbf{V}}$ conditional on the observed data $\mathbf{V}$. The point estimate used

**Table 3.** Final 2017 Major League Baseball rankings based on hierarchical Bayesian Bradley–Terry. Whereas $\widehat{\lambda}_i^{\mathrm{MLE}}$ is found by optimizing the likelihood, $\mathbb{E}[\lambda_i \mid \mathbf{V}]$ is found by integrating over a posterior probability density.

| | team | $\mathbb{E}[\lambda_i \mid \mathbf{V}]$ | wins |
|---|---|---|---|
| 1 | **LAN** | 0.38 | 104 |
| 2 | **CLE** | 0.35 | 102 |
| 3 | **HOU** | 0.35 | 101 |
| 4 | **WAS** | 0.22 | 97 |
| 5 | **BOS** | 0.21 | 93 |
| 6 | **ARI** | 0.20 | 93 |
| 7 | **NYA** | 0.18 | 91 |
| 8 | **CHN** | 0.16 | 92 |
| 9 | **COL** | 0.10 | 87 |
| 10 | **MIN** | 0.07 | 85 |
| 11 | MIL | 0.07 | 86 |
| 12 | SLN | 0.02 | 83 |
| 13 | TBA | 0.00 | 80 |
| 14 | LAA | 0.00 | 80 |
| 15 | KCA | −0.01 | 80 |
| 16 | SEA | −0.04 | 78 |
| 17 | TEX | −0.04 | 78 |
| 18 | TOR | −0.06 | 76 |
| 19 | BAL | −0.08 | 75 |
| 20 | OAK | −0.09 | 75 |
| 21 | MIA | −0.10 | 77 |
| 22 | PIT | −0.12 | 75 |
| 23 | SDN | −0.17 | 71 |
| 24 | ATL | −0.19 | 72 |
| 25 | NYN | −0.21 | 70 |
| 26 | CHA | −0.22 | 67 |
| 27 | CIN | −0.22 | 68 |
| 28 | DET | −0.27 | 64 |
| 29 | PHI | −0.28 | 66 |
| 30 | SFN | −0.28 | 64 |

**Table 4.** Rankings from the 2017 season using maximum likelihood estimates. Maximum likelihood tends to produce results that diverge more from a team's actual record.

| | team | $\widehat{\lambda}_i^{\mathrm{MLE}}$ | wins |
|---|---|---|---|
| 1 | **CLE** | 0.52 | 102 |
| 2 | **HOU** | 0.51 | 101 |
| 3 | **LAN** | 0.50 | 104 |
| 4 | **BOS** | 0.33 | 93 |
| 5 | **NYA** | 0.29 | 91 |
| 6 | **WAS** | 0.26 | 97 |
| 7 | **ARI** | 0.26 | 93 |
| 8 | **CHN** | 0.19 | 92 |
| 9 | **MIN** | 0.13 | 85 |
| 10 | **COL** | 0.12 | 87 |
| 11 | MIL | 0.07 | 86 |
| 12 | TBA | 0.05 | 80 |
| 13 | LAA | 0.03 | 80 |
| 14 | KCA | 0.02 | 80 |
| 15 | SLN | 0.00 | 83 |
| 16 | SEA | −0.02 | 78 |
| 17 | TEX | −0.03 | 78 |
| 18 | TOR | −0.04 | 76 |
| 19 | BAL | −0.06 | 75 |
| 20 | OAK | −0.09 | 75 |
| 21 | PIT | −0.18 | 75 |
| 22 | MIA | −0.19 | 77 |
| 23 | SDN | −0.25 | 71 |
| 24 | CHA | −0.27 | 67 |
| 25 | ATL | −0.30 | 72 |
| 26 | CIN | −0.33 | 68 |
| 27 | NYN | −0.34 | 70 |
| 28 | DET | −0.34 | 64 |
| 29 | SFN | −0.41 | 64 |
| 30 | PHI | −0.43 | 66 |

for prediction is $\mathbb{E}[\tilde{\mathbf{V}} \mid \mathbf{V}]$. The accuracy of the predictive distribution can be readily measured by splitting the sample. We fit the data in a given season up to a certain date, and predict team records for the remainder of the season. This can of course be validated against the known outcomes. In machine learning terminology, the date at which we partition the data represents the separation between the training and the test sets. We can define a loss function, or error metric, to evaluate the overall validity of this approach, and compare it to predictions based on generating samples from maximum likelihood estimates alone. The respective error metrics for a given team are

$$\begin{aligned}\text{error}_i^{\text{Bayes}} &= \left|\mathbb{E}\left[\tilde{V}_i^{\text{test}} \mid V_i^{\text{train}}\right] - V_i^{\text{test}}\right|, \\ \text{error}_i^{\text{MLE}} &= \left|\mathbb{E}\left[\tilde{V}_i^{\text{test}} \; ; \; \widehat{\lambda}_i^{\text{train}}\right] - V_i^{\text{test}}\right|.\end{aligned}$$

In words, these are the absolute distances between the predicted wins and actual wins in the test set. An overall error metric can be given by

$$\begin{aligned}\text{error}^{\text{Bayes}} &= \frac{1}{N}\sum_{i=1}^{N}\left(\text{error}_i^{\text{Bayes}}\right), \\ \text{error}^{\text{MLE}} &= \frac{1}{N}\sum_{i=1}^{N}\left(\text{error}_i^{\text{MLE}}\right),\end{aligned}$$

the means of the individual error metrics. Similarly,

$$\begin{aligned}\text{sd}^{\text{Bayes}} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\text{error}_i^{\text{Bayes}} - \text{error}^{\text{Bayes}}\right)^2}, \\ \text{sd}^{\text{MLE}} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\text{error}_i^{\text{Bayes}} - \text{error}^{\text{MLE}}\right)^2}\end{aligned}$$

measures how variable these estimates are. The results for the 2017 season are shown in table 5. In figure 4, we plot at each partition date the average overall predictive metrics during the seasons 2011–2017. With sufficient data, the two methods perform similarly. However, the Bayesian approach offers much-improved performance when data is scarce. This is true for both error and variability. In this sense, our model is preferable early in the season, and continues to outperform MLE-based predictions into July, after which the two methods begin to converge in accuracy. Without doubt, higher accuracy could be achieved under different approaches if that were the sole goal; we aim to strike a balance between our two described applications.

## 4. Conclusions

Our proposed Bayesian Bradley–Terry model provides a useful and coherent framework for assessing and predicting the performance of Major League Baseball teams. By adhering to Whelan's desiderata [19], we construct a hierarchical model that is weakly informative at the level of individual teams, but includes prior knowledge with respect to the entire league. This permits a model that combines the subjective and objective approaches to Bayesian inference, capable of for use in a range of applications. Specifically, we demonstrate the merit of Bayesian Bradley–Terry in applications to ranking and prediction, finding a balance between inferring latent structure and making respectable forecasts. In both cases, our model outperforms

**Table 5.** Comparison of error rates from predictions based on hierarchical Bayesian Bradley–Terry and maximum likelihood for the 2017 season. “Partition” indicates where the data was split into a training and test. The Bayesian approach performs significantly better during the first half of the season.

| partition | $\text{error}^{\text{Bayes}}$ | $\text{error}^{\text{MLE}}$ | $\text{sd}^{\text{Bayes}}$ | $\text{sd}^{\text{MLE}}$ |
|---|---|---|---|---|
| Apr15 | 8.82 | 24.65 | 6.58 | 17.34 |
| May1 | 7.31 | 12.49 | 6.17 | 10.39 |
| May15 | 6.20 | 9.84 | 5.68 | 5.84 |
| Jun1 | 4.72 | 6.90 | 4.87 | 4.27 |
| Jun15 | 4.32 | 4.81 | 4.65 | 4.79 |
| Jul1 | 4.04 | 4.17 | 3.46 | 3.43 |
| Jul15 | 3.90 | 4.01 | 3.72 | 3.96 |
| Aug1 | 3.58 | 3.89 | 3.17 | 3.15 |
| Aug15 | 3.32 | 3.26 | 2.74 | 2.91 |
| Sep1 | 2.57 | 2.59 | 2.31 | 2.39 |
| Sep15 | 1.75 | 1.83 | 1.14 | 1.11 |

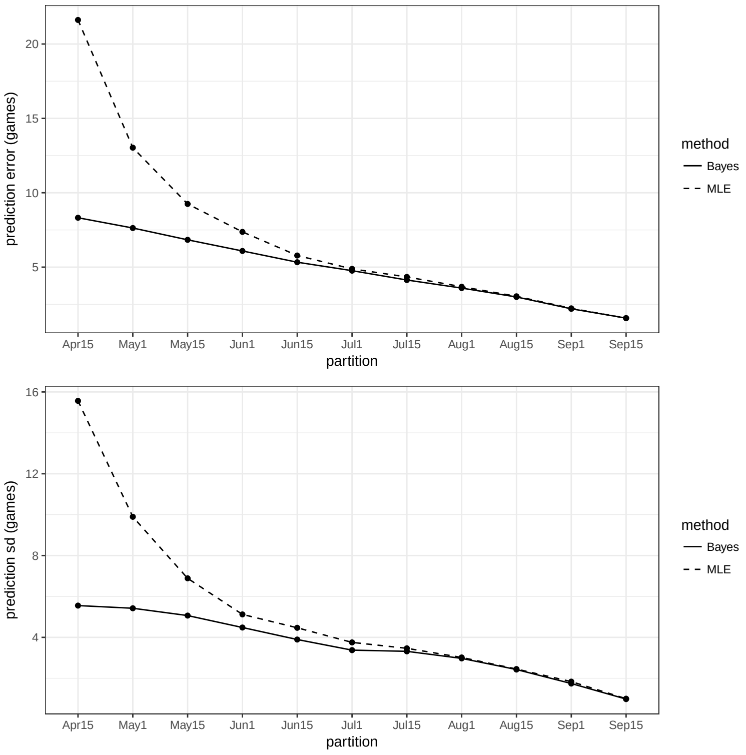


**Figure 4.** A comparison of predictions based on $\mathbb{E}[\tilde{\mathbf{V}}^{\text{test}} \mid \mathbf{V}^{\text{train}}]$ and $\mathbb{E}[\tilde{\mathbf{V}}^{\text{test}} ; \widehat{\lambda}_i^{\text{train}}]$, averaged across the seasons 2011–2017. “Partition” indicates when the data was split into a training and test set. In general, Bayesian Bradley–Terry matches or beats the performance of MLE-based prediction for the entirety of a season, in terms of both error rate and error variability.

maximum likelihood estimation by integrating over uncertainty, preventing overfitting. In summary, hierarchical Bayesian Bradley–Terry offers good performance in application, while being simple, interpretable, and compliant with the desirable properties of the desiderata.

## References

[1] A. Albert and J. A. Anderson, *On the existence of maximum likelihood estimates in logistic regression models*, Biometrika **71** (1984), 1–10.

[2] Baseball Reference, MLB scores, Data, 2017.

[3] M. Betancourt, *Efficient Bayesian inference with Hamiltonian Monte Carlo*, Lecture, 2014.

[4] M. Betancourt, *A conceptual introduction to Hamiltonian Monte Carlo*, arXiv:1701.02434, 2016.

[5] R. A. Bradley and M. E. Terry, *Rank analysis of incomplete block desings: The method of paired comparisons*, Biometrika **39** (1952), 324–345.

[6] K. Butler and J. T. Whelan, *The existence of maximum-likelihood estimates in the Bradley–Terry model and its extensions*, arXiv:math/0412232, 2000.

[7] F. Caron, A. Doucet, *Efficient Bayesian inference for generalized Bradley–Terry models*, arXiv:1011.1761, 2010.

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

[9] V. Csiszár, *EM algorithms for generalized Bradley–Terry models*, Annales Universitatis Scientiarium Budapestinensis de Rolando Eötvös Nominatae, Sectio Computatorica **36** (2012), 143–157.

[10] R. R. Davidson and D. L. Solomon, *A Bayesian approach to paired comparisons*, Biometrika **60** (1973), 477–487.

[11] L. R. Ford Jr., *Solution of a ranking problem from binary comparisons*, The American Mathematical Monthly **64** (1957), 28–33.

[12] D. R. Hunter, *MM algorithms for generalized Bradley–Terry models*, The Annals of Statistics **32** (2004), 384–406.

[13] T. Leonard, *An altervative Bayesian approach to the Bradley–Terry model for paired comparisons*, Biometrics **33** (1977), 121–132.

[14] M. M. Nassar and A. Elmasry, *A study of generalized logistic distributions*, Journal of the Egyption Mathematical Society **20** (2012), 126–133.

[15] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016.

[16] Retrosheet, Regular Season Event Files, Data, 2017.

[17] T. J. Santner and D. E. Duffy, *A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models*, Biometrika **73** (1986), 755–758.

[18] Stan Development Team, RStan: the R interface to Stan, R package version 2.14.1, 2016.

[19] J. T. Whelan, *Prior distributions for the Bradley–Terry model of paired comparisons*, arXiv:1712.05311v1, 2017.

[20] E. Zermelo, *Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung*, Mathematische Zeitschrift **29** (1929), 436–460.

Gabriel C. Phelan, School of Mathematical Sciences, Rochester Institute of Technology, 85 Lomb Memorial Drive, Rochester, New York 14623
*e-mail*: gxp3900@rit.edu

John T. Whelan, John T. Whelan, School of Mathematical Sciences and Center for Computational Relativity and Gravitation, Rochester Institute of Technology, 85 Lomb Memorial Drive, Rochester, New York 14623
*e-mail*: jtwsma@rit.edu