Prediction and Evaluation with Bradley-Terry: 
A College Hockey Case Study

John T. Whelan
jtwsma@rit.edu

School of Mathematical Sciences
and Center for Computational Relativity & Gravitation
Rochester Institute of Technology

Presented at UP-STAT 2018
University of Rochester 2018 April 21
Outline

1. Evaluation of Predictions Using the Bayes Factor

2. Bayesian Bradley-Terry Models
   - Haldane (Maximum Likelihood)
   - Beta (Generalized Logistic)
   - Gaussian

3. Hierarchical Bayesian Bradley-Terry Models
Outline

1. Evaluation of Predictions Using the Bayes Factor

2. Bayesian Bradley-Terry Models
   - Haldane (Maximum Likelihood)
   - Beta (Generalized Logistic)
   - Gaussian

3. Hierarchical Bayesian Bradley-Terry Models
NCAA Hockey Tournament

NCAA Regionals
March 23–25

- Midwest Regional
  (Allentown, PA)
  March 24–25
  1M Ohio State
  4M Princeton
  2M Denver
  3M Penn State

- Northeast Regional
  (Worcester, MA)
  March 24–25
  1N Cornell
  4N Boston Univ.
  2N Michigan
  3N Northeastern

- West Regional
  (Sioux Falls, SD)
  March 23–24
  1W St. Cloud St.
  2W Minn. State
  4W Air Force
  3W Minn.-Duluth

- East Regional
  (Bridgeport, CT)
  March 23–24
  1E Notre Dame
  2E Providence
  4E Mich. Tech
  3E Clarkson

NCAA Men’s Frozen Four™
(St. Paul, MN)
April 5 & 7

1M Ohio State
4M Princeton
2M Denver
3M Penn State
Bayes Factor Evaluation
Bayesian Bradley-Terry Models
Hierarchical Bradley-Terry Models

Example of Probabilistic Predictions

Model predicted probs for outcome of each (possible) game Based on info including regular season results

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cr def BU</td>
<td>0.683</td>
</tr>
<tr>
<td>BU def Cr</td>
<td>0.317</td>
</tr>
<tr>
<td>Mi def NE</td>
<td>0.541</td>
</tr>
<tr>
<td>NE def Mi</td>
<td>0.459</td>
</tr>
<tr>
<td>Cr def Mi</td>
<td>0.624</td>
</tr>
<tr>
<td>Mi def Cr</td>
<td>0.376</td>
</tr>
<tr>
<td>Cr def NE</td>
<td>0.661</td>
</tr>
<tr>
<td>NE def Cr</td>
<td>0.339</td>
</tr>
<tr>
<td>BU def Mi</td>
<td>0.436</td>
</tr>
<tr>
<td>Mi def BU</td>
<td>0.564</td>
</tr>
<tr>
<td>BU def NE</td>
<td>0.476</td>
</tr>
<tr>
<td>NE def BU</td>
<td>0.524</td>
</tr>
</tbody>
</table>

Note: outcomes may be correlated e.g., team strength not well constrained by model

Combine prob for each of $2^{15}$ possible “bracket” outcomes
Assessing Predictions with the Bayes Factor

- Odds ratio for model comparison

\[
\frac{P(\mathcal{M}_1|D, I)}{P(\mathcal{M}_2|D, I)} = \frac{P(D|\mathcal{M}_1, I) \cdot P(\mathcal{M}_1|I)}{P(D|\mathcal{M}_2, I) \cdot P(\mathcal{M}_2|I)}
\]

- Here \(D \equiv \) tourney results

Bayes factor \(\frac{P(D|\mathcal{M}_1, I)}{P(D|\mathcal{M}_2, I)}\) only needs prob of actual set of results

(not all \(2^{15}\) possible)

- Convenient to compare each model to “tossup” model where outcome of each game is 50-50:

\[
\mathcal{B}(\mathcal{M}) = \prod_{\text{game}} 2 \times P(\text{winner}|\mathcal{M}, I)
\]

For each game, pick up a factor between 0 and 2.
Bayes Factor Evaluation
Bayesian Bradley-Terry Models
Hierarchical Bradley-Terry Models

Bayes Factor Example

Bayes factor for KRACH in 2018 NCAAs

Cumulative Bayes factor vs tossup

0.125
0.25
0.5
1
2
4
8

Cumulative Bayes factor vs tossup

0.125
0.25
0.5
1
2
4
8

Bayes factor for KRACH in 2018 NCAAs
Cumulative Bayes Factor, 2003-2018

Bayes factor for KRACH in NCAAs

Cumulative Bayes factor vs tossup

Bayes factor for KRACH in NCAAs
Outline

1. Evaluation of Predictions Using the Bayes Factor

2. Bayesian Bradley-Terry Models
   - Haldane (Maximum Likelihood)
   - Beta (Generalized Logistic)
   - Gaussian

3. Hierarchical Bayesian Bradley-Terry Models
**KRACH = Maximum Likelihood Bradley-Terry**

- **Model:** each team \( i \in \{1, \ldots, t\} \) has log-strength\(^1\) \( \lambda_i \)

\[
P(i \text{ def } j) = \theta_{ij} = \frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j}} = \text{logistic}(\lambda_i - \lambda_j)
\]

Zermelo 1929, Bradley & Terry 1952, (Butler 1993)

- **Given (reg season) results:** \( w_{ij} \) wins in \( n_{ij} \) games for \( i \) against \( j \), KRACH ratings \( \{\hat{\lambda}_i\} \) satisfy

\[
v_i = \sum_{j=1}^{t} w_{ij} = \frac{e^{\hat{\lambda}_i}}{e^{\hat{\lambda}_i} + e^{\hat{\lambda}_j}} = \sum_{j=1}^{t} n_{ij}\hat{\theta}_{ij}
\]

and maximize likelihood

\[
p(\{w_{ij}\} | \{\lambda_i\}, \{n_{ij}\}) = \prod_{i=1}^{t} \prod_{j=1}^{t} \binom{n_{ij}}{w_{ij}} \theta_{ij}^{w_{ij}}
\]

\(^1\)or a strength \( \pi_i = e^{\lambda_i} \)
Bradley-Terry w/Haldane Prior

- ML $\{\hat{\lambda}_i\} \equiv$ maximum a posteriori w/uniform (improper) “Haldane” prior $f(\{\lambda_i\}|l_0) = \text{const}$
- Bayesian approach would marginalize over posterior
  
  $$f(\{\lambda_i\}|w_{ij}, l_0) \propto p(\{w_{ij}\}|\{\lambda_i\}, \{n_{ij}\})$$

  to get $P(D|w_{ij}, l_0) = \int d\lambda P(D|\{\lambda_i\}, l_0) f(\{\lambda_i\}|w_{ij}, l_0)$

- Approximate with Gaussian expansion about MAP point

  $$f_G(\{\lambda_i\}|w_{ij}, l_0) \propto \exp \left( -\frac{1}{2} \sum_{i=1}^{t} \sum_{j=1}^{t} (\lambda_i - \hat{\lambda}_i) H_{ij} (\lambda_j - \hat{\lambda}_j) \right)$$

  with inverse variance-covariance matrix

  $$H_{ij} = -\left. \frac{\partial^2 \ln f(\{\lambda_i\}|w_{ij}, l_0)}{\partial \lambda_i \partial \lambda_j} \right|_{\{\lambda_i=\hat{\lambda}_i\}} = -n_{ij}\hat{\theta}_{ij}\hat{\theta}_{ji} + \delta_{ij} \sum_{k=1}^{t} n_{ik}\hat{\theta}_{ik}\hat{\theta}_{ki}$$

2by analogy to Haldane 1932 in binomial case
Cumulative Bayes Factor, 2003-2018

Bayes factor for KRACH in NCAAs

Haldane MAP
Haldane Gaussian
Bayesian Bradley-Terry Models
Hierarchical Bradley-Terry Models

Bayes Factor Evaluation
Haldane (Maximum Likelihood)
Beta (Generalized Logistic)
Gaussian

Bradley-Terry w/Beta Prior

- Haldane prior improper \( \eta \to 0 \) extreme results w/e.g., undefeated teams
  Prefer proper prior which satisfies desiderata of Whelan 2018, e.g., \( \text{Beta}(\eta, \eta) \) in \( \zeta_i = \text{logistic}(\lambda_i) = \frac{\pi_i}{1+\pi_i} \in (0, 1) \):

\[
f(\lambda_i | I_\eta) \propto \left(1 + e^{-\lambda_i}\right)^{-\eta} \left(1 + e^{\lambda_i}\right)^{-\eta}
\]

\( \eta \to 0 \) is Haldane; \( \eta = 1 \) is uniform in \( \zeta_i \)

- MAP equations: \( v_i + \eta = \sum_{j=1}^{t} n_{ij} \hat{\theta}_{ij} + 2\eta \hat{\zeta}_i \)

MLE w/\( 2\eta \) games “split” vs “fictitious team” w/log-strength 0
\( \eta = \frac{1}{2} \) sometimes used to regularize KRACH (Butler 1993)

- Gaussian approx w/

\[
H_{ij} = -n_{ij} \hat{\theta}_{ij} \hat{\theta}_{ji} + \delta_{ij} \left( \sum_{k=1}^{t} n_{ik} \hat{\theta}_{ik} \hat{\theta}_{ki} + 2\eta \hat{\zeta}_i (1 - \hat{\zeta}_i) \right)
\]
Cumulative Bayes Factor, 2003-2018

Bayes factor for KRACH in NCAAs

- $\eta = 0$ MAP
- $\eta = 0$ Gaussian
- $\eta = 1$ MAP
- $\eta = 1$ Gaussian
Bradley-Terry w/Gaussian Prior

- Convenient to work with Gaussian prior on \{\lambda_i\} (Leonard 1977; Whelan 2018)

\[ f(\lambda_i | I_\sigma) \propto \exp \left( -\frac{\lambda_i^2}{2\sigma^2} \right) \]

\( \sigma \to \infty \) is Haldane

- MAP equations:

\[ v_i = \sum_{j=1}^{t} n_{ij} \hat{\theta}_{ij} + \frac{\hat{\lambda}_i}{\sigma^2} \]

- Gaussian approx w/

\[ H_{ij} = -n_{ij} \hat{\theta}_{ij} \hat{\theta}_{ji} + \delta_{ij} \left( \sum_{k=1}^{t} n_{ik} \hat{\theta}_{ik} \hat{\theta}_{ki} + \frac{1}{\sigma^2} \right) \]
Cumulative Bayes Factor, 2003-2018

Bayes factor for KRACH in NCAAs

η = 0 MAP
η = 0 Gaussian
η = 1 MAP
η = 1 Gaussian
σ = 1 MAP
σ = 1 Gaussian

14/23 John T. Whelan  jtwsma@rit.edu Bradley-Terry for College Hockey  UP-STAT, 2018 Apr 21
Comparison of Prior Distributions

Prior on log-strength

- $\eta = 1$
- $\eta = 1/2$
- $\sigma = 1$
- $\sigma = 2$
- $\sigma = 3$

Prior pdf

$\lambda_i$

Prior on log-strength

$\eta$

$\sigma$

$\eta = 1$

$\eta = 1/2$

$\sigma = 1$

$\sigma = 2$

$\sigma = 3$

Prior pdf

$\lambda_i$

$\eta$

$\sigma$
Outline

1 Evaluation of Predictions Using the Bayes Factor

2 Bayesian Bradley-Terry Models
   - Haldane (Maximum Likelihood)
   - Beta (Generalized Logistic)
   - Gaussian

3 Hierarchical Bayesian Bradley-Terry Models
Motivation for Hierarchical Model

- Values of $\eta$ in $I_\eta$ & $\sigma$ in $I_\sigma$ arbitrary
- Use hierarchical model w/hyperprior on hyperparameter (Phelan & Whelan 2018) e.g.,

$$f(\{\lambda_i\}, \sigma | I_H) \propto \sigma^{-t} \exp \left( -\frac{\sum_{i=1}^{t} \lambda_i^2}{2\sigma^2} \right) f(\sigma | I_H)$$

- Uniform hyperprior on $\sigma$?

MAP eqns

$$v_i = \sum_{j=1}^{t} n_{ij} \hat{\theta}_{ij} + \frac{\hat{\lambda}_i}{\hat{\sigma}^2}$$

and

$$\hat{\sigma}^2 = \frac{1}{t} \sum_{i=1}^{t} \hat{\lambda}_i^2$$

Problem: MAP point is at $\sigma \to 0$

- Phelan & Whelan 2018 used $\Gamma(\alpha, \beta)$ prior motivated by variance of MLE $\{\hat{\lambda}_i\}$ from previous season

Still have $\hat{\sigma} \to 0$ unless $\alpha \geq t + 1$
Standard Deviation of ML Log-Strengths

Compare $\sigma \approx 0.27$ for MLB (Phelan & Whelan 2018)
Standard Deviation of ML Log-Strengths

RMS log-strength values

\[ \hat{\sigma}_\lambda \]

mean(ln(\(\sigma\))) = 0.019, std(ln(\(\sigma\))) = 0.379
Log-Normal Hierarchical Model

- Use log-normal prior on $\sigma$ w/empirical parameters $\ln \sigma_0$ & $\varepsilon$

\[
f(\{\lambda_i\}, \ln \sigma | I_L) \propto \sigma^{-t} \exp \left( -\sum_{i=1}^{t} \frac{\lambda_i^2}{2\sigma^2} - \frac{(\ln \sigma - \ln \sigma_0)^2}{2\varepsilon^2} \right)
\]

- MAP eqns

\[
v_i = \sum_{j=1}^{t} n_{ij} \hat{\theta}_{ij} + \frac{\hat{\lambda}_i}{\hat{\sigma}^2} \quad \text{and} \quad t = \frac{\sum_{i=1}^{t} \hat{\lambda}_i^2}{\hat{\sigma}^2} - \frac{\ln \hat{\sigma} - \ln \sigma_0}{2\varepsilon^2}
\]

- Gaussian approx w/ $H_{ij} = -n_{ij} \hat{\theta}_{ij} \hat{\theta}_{ji} + \delta_{ij} \left( \sum_{k=1}^{t} n_{ik} \hat{\theta}_{ik} \hat{\theta}_{ki} + \frac{1}{\hat{\sigma}^2} \right)$

and $H_{i \ln \sigma} = -\frac{\hat{\lambda}_i}{\hat{\sigma}^2}$ and $H_{\ln \sigma \ln \sigma} = 2 \frac{\sum_{i=1}^{t} \hat{\lambda}_i^2}{\hat{\sigma}^2} + \frac{1}{\varepsilon^2}$
MAP Estimate of Hyperparameter

![Graph showing standard deviation vs hierarchical parameter](image)

- **σ_\hat{λ}**
- **\hat{σ}**

Data points for years 2003 to 2018 are plotted.
Cumulative Bayes Factor, 2003-2018

Bayes factor for KRACH in NCAAs

η = 0 MAP
η = 0 Gaussian
η = 1 MAP
η = 1 Gaussian
σ = 1 MAP
σ = 1 Gaussian
hierarchical MAP

Cumulative Bayes factor vs tossup


1 4 16 64 256 1024 4096 16384 65536 262144

η = 0 MAP
η = 0 Gaussian
η = 1 MAP
η = 1 Gaussian
σ = 1 MAP
σ = 1 Gaussian
hierarchical MAP
Takeaways

- Bayes factors: evaluate probabilistic predictions after the fact
- Bradley-Terry is definitely better than guessing!
- Hard to distinguish details with $15 \times 16$ results
- Hierarchical modelling is harder than it looks (but maybe worth it)
References

- Bradley and Terry 1952 *Biometrika* **39**, 324
- Butler 1993 *Ken’s Ratings for American College Hockey*  
  [link](http://lists.maine.edu/cgi/wa?A2=Hockey-L;68c89935.9307)
- Leonard 1977 *Biometrics* **33**, 121
- Phelan and Whelan 2018 arXiv:1712.05879
- Whelan 2018 arXiv:1712.05311
EXTRA SLIDES
Finiteness of Maximum Likelihood Ratios

Solutions to $v_i = \sum_{j=1}^{t} w_{ij} = \sum_{j=1}^{t} n_{ij} \frac{\hat{\pi}_i}{\hat{\pi}_i + \hat{\pi}_j} = \sum_{j=1}^{t} n_{ij} \hat{\theta}_{ij}$

- Obv, if $k$ undefeated ($v_k = \sum_{j=1}^{t} n_{ik}$), $\hat{\pi}_k \to \infty$ & $\forall j : \hat{\theta}_{kj} = 1$

- More generally (Albert & Anderson 1984, Santner & Duffy 1986), if you can make a “chain of wins” from $i$ to $j$, write $i \triangleright j$
  - If $i \triangleright j$ but $j \not\triangleright i$, then $\hat{\pi}_i / \hat{\pi}_j \to \infty$, and $\hat{\theta}_{ij} = 1$
  - If $i \triangleright j$ and $j \triangleright i$, then $\hat{\pi}_i / \hat{\pi}_j$ finite, and $0 < \hat{\theta}_{ij} < 1$
  - If $i \not\triangleright j$ and $j \not\triangleright i$, then $\hat{\pi}_i / \hat{\pi}_j$ & $\hat{\theta}_{ij}$ undetermined

- Butler & Whelan 2000: teams split into “groups” (equiv classes) within which ML ratios are finite; ML ratios between groups are $0$, $\infty$ or undefined

Can summarize in Directed Acyclic Graph
Problems with Maximum Likelihood Estimates

- Infinite or undetermined MLE ratios problematic
- Problem goes away given enough data, but compromises use of Bradley-Terry to rank teams after a short season (e.g., College Football)
- Counterintuitive:
  - beating an “infinitely worse” team does nothing to MLE
  - impossible to be better than an undefeated team
- Can resort to ad hoc regularization e.g., “fictitious games” to force ratios to be finite (Butler, unpublished)
- Motivates a Bayesian approach with prior information (at least “nobody’s perfect”)
Parametrization for Bradley-Terry

<table>
<thead>
<tr>
<th>$\pi_i$</th>
<th>$\lambda_i = \ln \pi_i$</th>
<th>$\zeta_i = \pi_i \frac{1}{1+\pi_i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_{ij} = \frac{\pi_i}{\pi_i + \pi_j}$</td>
<td>$\gamma_{ij} = \ln \frac{\theta_{ij}}{1-\theta_{ij}} = \lambda_i - \lambda_j$</td>
<td>$\lambda_i = \ln \frac{\zeta_i}{1-\zeta_i}$</td>
</tr>
</tbody>
</table>

- $p(D|\lambda) = p(D|\pi)$ with $\pi_i = e^{\lambda_i}$ but $f(\lambda|X) = e^{\sum_{i=1}^{t} \lambda_i} f(\pi|X)$
- Work with $\lambda$ because $\sum_{i=1}^{t} \frac{\partial \theta_{jk}}{\partial \lambda_i} = 0$, i.e., probabilities $\{\theta_{ij}\}$ depend on combinations “orthogonal” to $\sum_{i=1}^{t} \lambda_i$
- Note if prior $f(\lambda|I)$ is uniform, posterior $f(\lambda|D, I)$ is maximized by maximum likelihood solution $\hat{\lambda}$