Using Bayesian statistics to rank sports teams (or, my replacement for the BCS)

John T. Whelan
jtwsma@rit.edu

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

π-RIT Presentation
2010 October 8

Outline

Outline

The Problem: Who Are The Champions?

- The games have been played; crown the champion (or seed the playoffs)
- If the schedule was balanced, it's easy:
  pick the team with the best record
- If schedule strengths differ, record doesn't tell all
  e.g., college sports (seeding NCAA tourneys)

Evaluating an Unbalanced Schedule

- Most NCAA sports (basketball, hockey, lacrosse, . . .) have a selection committee
- That committee uses or follows selection criteria (Ratings Percentage Index, strength of schedule, common opponents, quality wins, . . .)
- Football (Bowl Subdivision) has no NCAA tournament; Bowl Championship Series “seeded” by BCS rankings
- All involve some subjective judgement (committee or polls)

Requirements for a “Fair” Rating System

- Objective; anyone applying system will get same results
- Only consider this season’s results
  (no historical information, projections, injuries, etc.)
- Don’t consider margin of victory, only game outcome
- Shouldn’t matter which games you win
- Should be open, not secret

These are my personal judgements about what’s “fair”

Requirements for a “Fair” Rating System

- Objective; anyone applying system will get same results
- Only consider this season’s results
  (no historical information, projections, injuries, etc.)
- Don’t consider margin of victory, only game outcome
- Shouldn’t matter which games you win
- Should be open, not secret

These are my personal judgements about what’s “fair”
Note: BCS “computer ratings” satisfy all but the last
but BCS formula designed to make sure polls matter most

RPI

RPI (Ratings Percentage Index)

- Component of most NCAA selection criteria
- 25% winning pct + 50% opponents' winning pct
+ 25% opponents' opponents' winning pct

RPI (Ratings Percentage Index)

- Component of most NCAA selection criteria
- 25% winning pct + 50% opponents' winning pct + 25% opponents' opponents' winning pct

$$R_A = 0.25\,\frac{V_A}{N_A} + 0.50\,O_A + 0.25\sum_B \frac{N_{AB}}{N_A} O_B$$

$$O_A = \sum_B \frac{N_{AB}}{N_A}\,\frac{V_B - V_{BA}}{N_B - N_{BA}}$$

Some Notation

- $A, B, \ldots$ label teams
- N_{AB} number of times A plays B
- V_{AB} number of times A beats B
- $N_A = \sum_B N_{AB}$ total number of games for A
- $V_A = \sum_B V_{AB}$ total number of wins for A

Some Notation

- $A, B, \ldots$ label teams
- N_{AB} number of times A plays B
- V_{AB} number of times A beats B
- $N_A = \sum_B N_{AB}$ total number of games for A
- $V_A = \sum_B V_{AB}$ total number of wins for A

Formula for RPI:

$$R_A = 0.25\,\frac{V_A}{N_A} + 0.50\,O_A + 0.25\sum_B \frac{N_{AB}}{N_A} O_B$$

$$O_A = \sum_B \frac{N_{AB}}{N_A}\,\frac{V_B - V_{BA}}{N_B - N_{BA}}$$

Shortcomings of RPI

…illustrated by NCAA hockey examples

- 25% winning pct + 75% “strength of schedule”
- If your opponent is bad enough, beating them can be worse than not playing them at all (Bowling Green 1995)
- If your opponents play easy schedules, their good records can make your schedule look tougher than it is (Quinnipiac 2000)

Need a more comprehensive way of using all the results

Outline

Basics of the Bradley-Terry Model

- Each team has a rating π_A
- Assign probabilities to outcome of game between A and B:

$$P(A \text{ beats } B) = P_{AB} = \frac{\pi_A}{\pi_A + \pi_B}$$

- Independently invented:
  - 1928 Zermelo (rating chess players)
  - 1952 Bradley & Terry (evaluating taste tests)
  - Equivalent to Bill James's "log5" with $\pi_A = \frac{W\%(A)}{1-W\%(A)}$
- So what are the "right" ratings?

Determining BT Ratings (Sports Fan Method)

$$P_{AB} = \frac{\pi_A}{\pi_A + \pi_B}$$

- Require expected = actual number of wins for each team

$$V_A = \sum_B N_{AB} P_{AB}$$

- System of equations can be solved for the $\{\pi_A\}$

Determining BT Ratings (Classical Statistics Method)

$$P_{AB} = \frac{\pi_A}{\pi_A + \pi_B}$$

- Given ratings $\boldsymbol{\pi} \equiv \{\pi_A\}$, actual numbers of wins $\mathbf{V} \equiv \{V_{AB}\}$ are random variables with pmf from likelihood fcn:

$$p(\mathbf{V}|\boldsymbol{\pi}) = \prod_A \prod_B \binom{V_{AB}}{N_{AB}} P_{AB}^{V_{AB}}$$
$$= \prod_A \prod_B \frac{(V_{AB}+V_{BA})!}{V_{AB}!\,V_{BA}!} \left(\frac{\pi_A}{\pi_A+\pi_B}\right)^{V_{AB}}$$

- Find the ratings $\{\widehat{\pi}_A\}$ which maximize likelihood ML eqns are $V_A = \sum_B N_{AB} \widehat{P}_{AB}$ – same as before!

BT Does Pretty Well

$$P_{AB} = \frac{\pi_A}{\pi_A + \pi_B} \qquad V_A = \sum_B N_{AB} \widehat{P}_{AB}$$

- Popularized for college hockey by Ken Butler:
  Ken's Ratings for American College Hockey (KRACH)
- Winning never hurts, losing never helps
- Harder to "trick" than RPI
  Quinnipiac 2000 was #11 in RPI and #44 in KRACH (of 54)
- Some oddities, though, especially in short seasons . . .

Strange Features of Classical Bradley-Terry

$$P_{AB} = \frac{\pi_A}{\pi_A + \pi_B} \qquad V_A = \sum_B N_{AB} \widehat{P}_{AB}$$

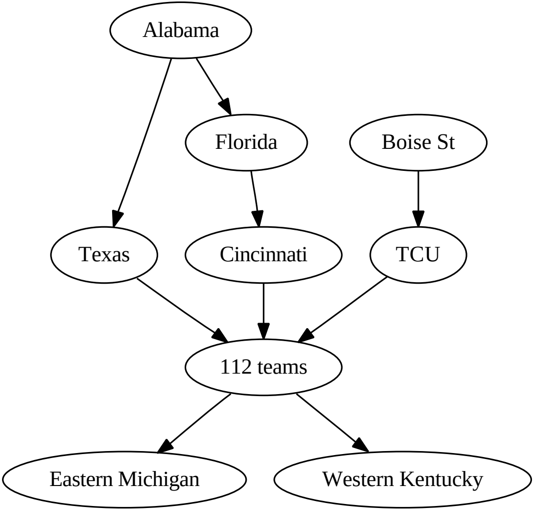
- Ratings only defined up to multiplicative factor
  No big deal; only ratios matter
- Undefeated team has infinite rating;
  in general ratios can be infinite or undefined

Dealing with Infinite or Undefined Ratios

See Butler and Whelan, arXiv: math.ST/0412232

- Remember the old game: Canisius beat SMU beat UAB beat NC State beat Duke
  - If you can make a "chain of wins" from A to B but not from B to A, $\pi_A/\pi_B = \infty$
  - If you can make a "chain of wins" both ways, π_A/π_B finite
  - If you can't make either "chain" π_A/π_B is undefined

Example: 2009 College Football (after the bowls)



Problems for Classical BT w/Short Seasons

$$P_{AB} = \frac{\pi_A}{\pi_A + \pi_B} \qquad V_A = \sum_B N_{AB} \widehat{P}_{AB}$$

- Ratios can be infinite or undefined
- Beating an “infinitely worse” team does nothing to ratings
- It’s impossible to be better than an undefeated team
- No way for more games to give more confidence in ratings

Outline

1. Ranking Systems
2. The Bradley-Terry Model
3. The Bayesian Approach

Bayesian Bradley-Terry

- Recall likelihood fcn

$$p(\mathbf{V}|\boldsymbol{\pi}) = \prod_A \prod_B \binom{V_{AB}}{N_{AB}} P_{AB}^{V_{AB}}$$

  Probability mass fcn for results $\{V_{AB}\}$ given ratings $\{\pi_A\}$

- Bayes's theorem gives us posterior

$$f(\boldsymbol{\pi}|\mathbf{V}) = \frac{p(\mathbf{V}|\boldsymbol{\pi})f(\boldsymbol{\pi})}{p(\mathbf{V})}$$

  Probability density fcn for ratings $\{\pi_A\}$ given results $\{V_{AB}\}$

Focus on the Logarithms

- Since π_A is multiplicative, it's actually more convenient to talk about $\lambda_A = \ln \pi_A$, i.e., $\pi_A = e^{\lambda_A}$.
- Posterior pdf for $\{\lambda_A\}$ given $\{V_{AB}\}$

$$f(\boldsymbol{\lambda}|\mathbf{V}) = \frac{p(\mathbf{V}|\boldsymbol{\lambda})f(\boldsymbol{\lambda})}{p(\mathbf{V})}$$

- Can use peak of posterior pdf in $\boldsymbol{\lambda}$ to choose ratings
- What to use for the prior $f(\boldsymbol{\lambda})$?

Choice of Prior on log-Ratings

- "Fairness" tells us to use same prior pdf for each team
- Assume different ratings are a priori independent

$$f(\boldsymbol{\lambda}) = \prod_A f(\lambda_A)$$

- One possible prior: uniform in λ_A (Jeffreys); then

$$f(\boldsymbol{\lambda}|\mathbf{V}) \propto p(\mathbf{V}|\boldsymbol{\lambda})$$

and we get the same max likelihood eqns

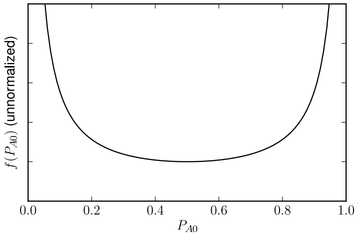
Drawback of Jeffreys Prior

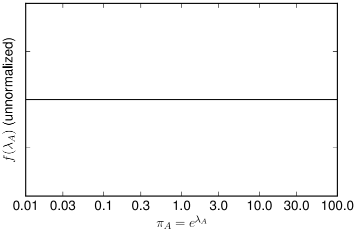
- Still nothing to set the scale
- Consider James's log5 (win prob vs "average" team)

$$P_{A0} = \frac{\pi_A}{1 + \pi_A}$$

Prior on this is

$$f(P_{A0}) = \frac{f(\lambda_A)}{P_{A0}(1 - P_{A0})}$$

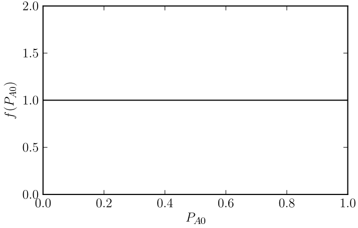


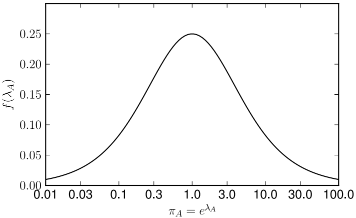


Alternative Prior

- Regularize things by choosing prior uniform in $P_{A0} = \frac{\pi_A}{1+\pi_A}$
- Prior is

$$f(P_{A0}) = 1 \qquad f(\lambda_A) = \frac{\pi_A}{(1+\pi_A)^2}$$





Bayesian BT with Regularizing Prior

- Each team's rating starts spread out around $\pi_A = 1$
- Each game result shapes the posterior
- Undefeated teams can still have lower estimates if results don't overwhelm prior
- Posterior pdf

$$f(\boldsymbol{\lambda}|\mathbf{V}) \propto p(\mathbf{V}|\boldsymbol{\lambda})f(\boldsymbol{\lambda})$$

  is complicated multi-dimensional fcn of $\boldsymbol{\lambda}$
- Near maximum $\widehat{\boldsymbol{\lambda}}$, approximate by Gaussian

$$f(\boldsymbol{\lambda}|\mathbf{V}) \approx f(\widehat{\boldsymbol{\lambda}}|\mathbf{V}) \exp\left(\frac{1}{2}(\boldsymbol{\lambda} - \widehat{\boldsymbol{\lambda}})^{\mathrm{tr}} \boldsymbol{\sigma}^{-2} (\boldsymbol{\lambda} - \widehat{\boldsymbol{\lambda}})\right)$$

Maximum Posterior BT Ratings w/Regularizing Prior

$$f(\boldsymbol{\lambda}|\mathbf{V}) = \frac{p(\mathbf{V}|\boldsymbol{\lambda})f(\boldsymbol{\lambda})}{p(\mathbf{V})} \approx f(\widehat{\boldsymbol{\lambda}}|\mathbf{V}) \exp\left(\frac{1}{2}(\boldsymbol{\lambda}-\widehat{\boldsymbol{\lambda}})^{\mathrm{tr}}\boldsymbol{\sigma}^{-2}(\boldsymbol{\lambda}-\widehat{\boldsymbol{\lambda}})\right)$$

- Can solve for peak $\widehat{\lambda} = \ln \widehat{\pi}$ and find equations

$$1 + V_A = 2\widehat{P}_{A0} + \sum_B N_{AB}\widehat{P}_{AB}$$

  Same as before but w/"fictitious games" vs team w/$\pi_0 = 1$
- Error matrix is

$$\sigma^{-2}{}_{AB} = -N_{AB}\widehat{P}_{AB}\widehat{P}_{BA} + \delta_{AB}\left[2\widehat{P}_{A0}\widehat{P}_{0A} + \sum_C N_{AC}\widehat{P}_{AC}\widehat{P}_{CA}\right]$$

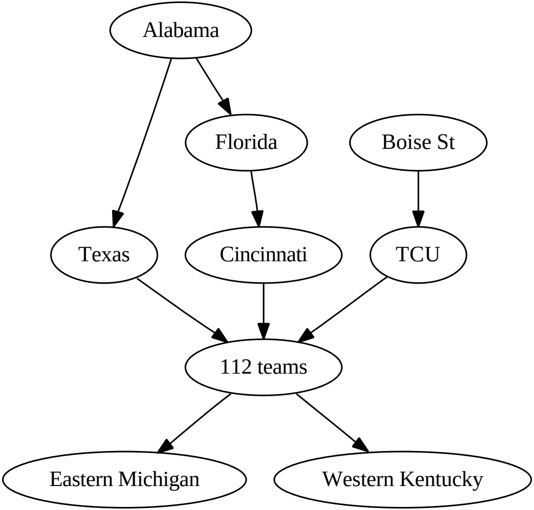
Maximum Posterior BT Ratings w/Error Estimates

- Use Gaussian approx to estimate marginal pdf

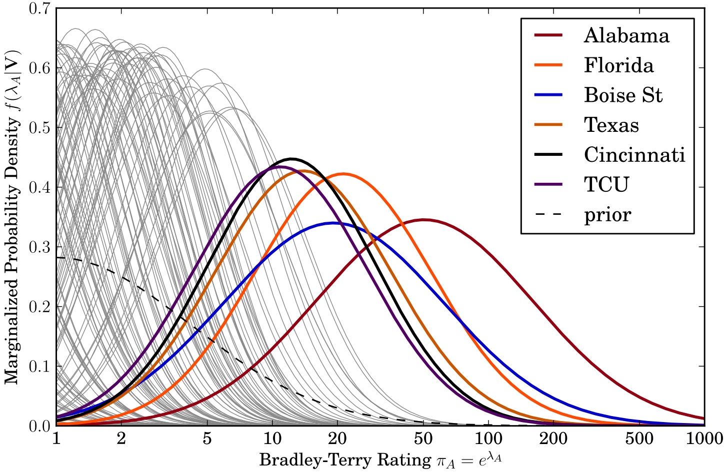
$$
\begin{aligned}
f(\lambda_A|\mathbf{V}) &= \left(\prod_{B\neq A}\int_{-\infty}^{\infty} d\lambda_B\right) f(\boldsymbol{\lambda}|\mathbf{V}) \\
&\approx f(\widehat{\lambda}_A|\mathbf{V}) \exp\left(-\frac{(\lambda_A-\widehat{\lambda}_A)^2}{2\sigma^2{}_{AA}}\right)
\end{aligned}
$$

- For ranking teams, $\widehat{\lambda}_A$ does the job
- Bayesian BT model can do so much more; e.g., σ_{AA} is error estimate

Recall 2009 College Football (post-bowls)



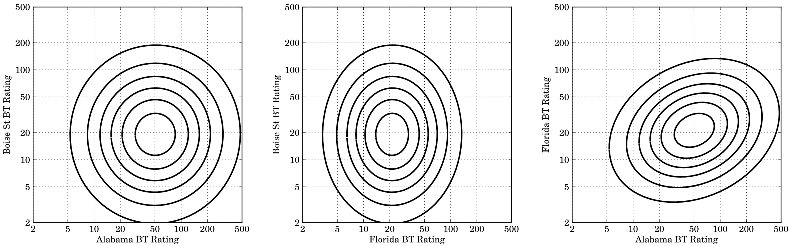
Marginal pdfs for 2009 College Football (post-bowls)



Partially-Marginalized pdfs for Pairs of Teams

- Single-team errors don't tell the whole story
- Can look at correlations w/partially-marginalized posterior

$$f(\lambda_A, \lambda_B|\mathbf{V}) = \left(\prod_{C \neq A,B} \int_{-\infty}^{\infty} d\lambda_B \right) f(\boldsymbol{\lambda}|\mathbf{V})$$

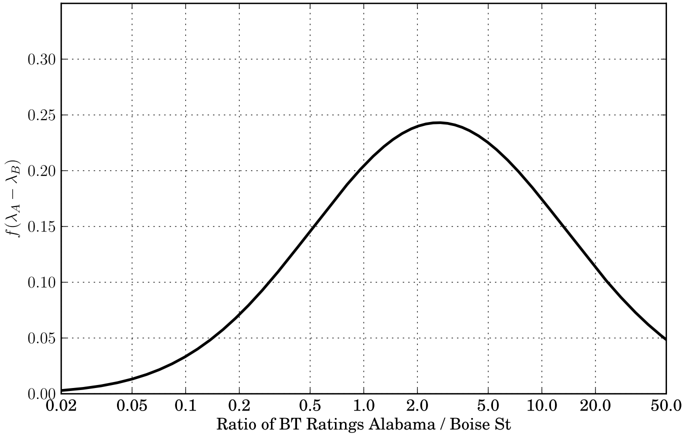


Posterior pdfs for Ratios of Ratings

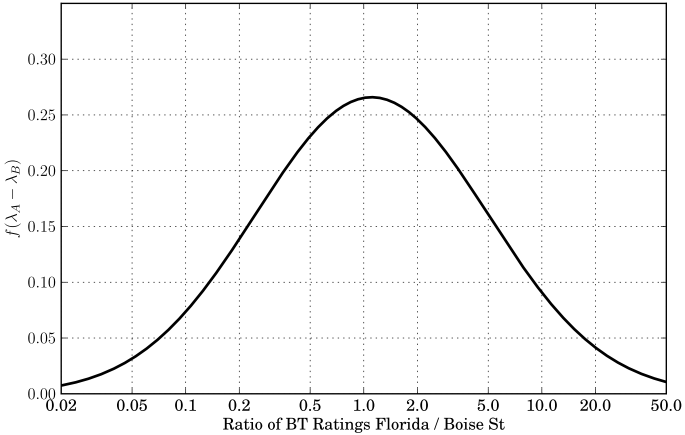
- Meaningful quantity is $\pi_A/\pi_B = e^{\lambda_A-\lambda_B} = e^{\Delta\lambda_{AB}}$
- Posterior PDF

$$f(\Delta\lambda_{AB}|\mathbf{V}) = \int_{-\infty}^{\infty} d\lambda_A \int_{-\infty}^{\infty} d\lambda_B\, f(\lambda_A,\lambda_B|\mathbf{V})$$
$$\approx f(\widehat{\Delta\lambda}_{AB}|\mathbf{V}) \exp\left(-\frac{(\Delta\lambda_{AB}-\widehat{\Delta\lambda}_{AB})^2}{2\sigma^2{}_{\Delta\lambda_{AB}}}\right)$$

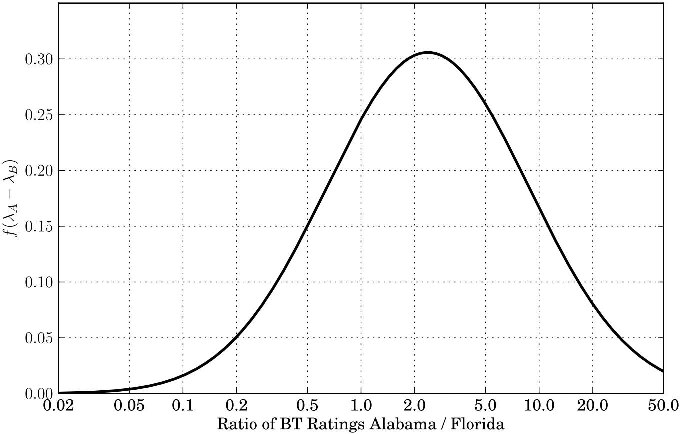
Posterior pdfs for Ratios of Ratings



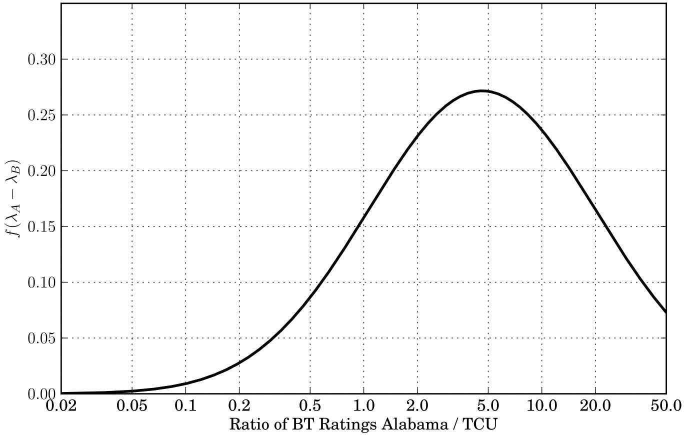
Posterior pdfs for Ratios of Ratings



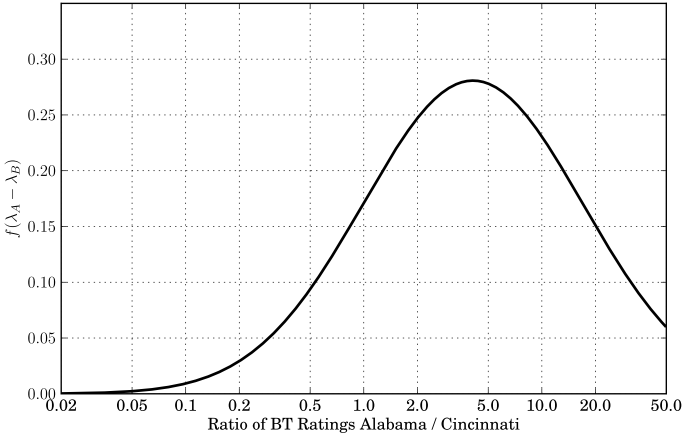
Posterior pdfs for Ratios of Ratings



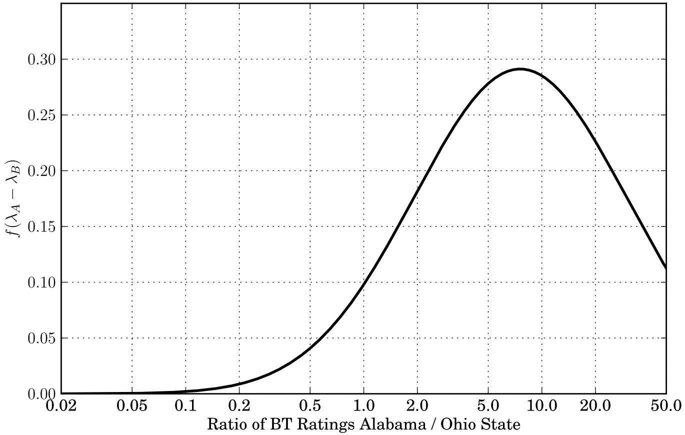
Posterior pdfs for Ratios of Ratings



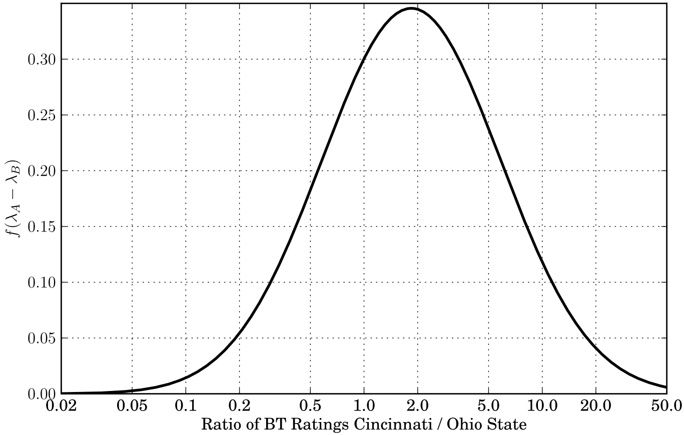
Posterior pdfs for Ratios of Ratings



Posterior pdfs for Ratios of Ratings



Posterior pdfs for Ratios of Ratings



What Have We Done? What Can We Do?

- Objective rating based on unbalanced schedules is tricky
- Bradley-Terry model $P_{AB} = \frac{\pi_A}{\pi_A + \pi_B}$ often works nicely
- Classical application has trouble with short seasons
- Bayesian application w/regularizing prior keeps things finite
- Applications/Investigations beyond just ranking teams
  - Marginalized error estimates on ratings
  - Bayesian model selection: BT vs something else
  - Different priors
  - Utility of extra params (e.g., home field) via odds ratio
  - Checking validity of Gaussian approximation w/monte carlo
  - …?

Addendum: If I Could Replace the BCS

- Play bowls on New Year's; back to traditional matchups
- After the bowls, rank the teams by $\widehat{\lambda}_A$
- Teams 1-6 make the playoffs
- First two rounds at campus sites:
  - Week One: #4 hosts #5; #3 hosts #6
  - Week Two: #1 & #2 host winners from Week One
- Week Three (off-weekend before Super Bowl): National Championship Game @ warm-weather site