Ranking and Selection from Pairwise Comparisons: Empirical Bayes Methods for Citation Analysis

Roger Koenker

University College London

AEA Session on League Tables: 9 January 2021

Joint work with Jiaying Gu (University of Toronto)

League Tables: Invasive Species of Statistical Analysis

- Universities, journals, chess players, football teams, . . .
- Personal favorite: Basel AML League Table of Money Laundering
- May have serious impact on resource allocation:
    - Medical treatments and facilities
    - Teacher and student evaluation
    - A/B testing, . . .
- Room for improvement in methods for ranking and selection.
- I'll focus on rating and ranking from pairwise comparisons.
- Journal rankings based on citation influence are produced.

Pairwise Comparisons and the Ranking Problem

In the simplest setting we have $p + 1$ players of scalar abilities, $\alpha_0, \alpha_1, \ldots, \alpha_p$ who meet in pairs; player i defeats player j with probability,

$$\pi_{ij} = \alpha_i / (\alpha_i + \alpha_j).$$

With a sufficiently rich accumulated history of play, the α's can be estimated by maximum likelihood and thereby ranked.

Pairwise Comparisons and the Ranking Problem

In the simplest setting we have $p + 1$ players of scalar abilities, $\alpha_0, \alpha_1, \ldots, \alpha_p$ who meet in pairs; player i defeats player j with probability,

$$\pi_{ij} = \alpha_i / (\alpha_i + \alpha_j).$$

With a sufficiently rich accumulated history of play, the α's can be estimated by maximum likelihood and thereby ranked.

In accordance with Stigler's law of eponymy, this Bradley-Terry (1952) model for ranking competitors based on paired comparisons was first proposed by Zermelo (1929) for rating chess players.

There has been considerable recent attention to such models in the machine earning community where random pairing assumptions make rankings based on total wins (aka Borda scores) attractive.

The Logistic Model

It is convenient to reparameterize abilities so $\theta_i = \log \alpha_i$ and π_{ij}, becomes,

$$\pi_{ij} = \frac{1}{1 + \exp(-(\theta_i - \theta_j))}$$

and to write the (logistic) log likelihood for n binary outcomes, $y_1, y_2, \ldots, y_n$, with $h_\theta(x_k) = 1/(1 + \exp(-\theta^\top x_k))$, as,

$$\ell(\theta|y) = \sum_{k=1}^{n} y_k \log(h_\theta(x_k)) + (1 - y_k) \log(1 - h_\theta(x_k))$$

where for match k between i and j, x_k is an p vector with ith element 1, and jth element -1, and other elements 0. Wlog, we set $\theta_0 = 0$.

The Logistic Model

It is convenient to reparameterize abilities so $\theta_i = \log \alpha_i$ and π_{ij}, becomes,

$$\pi_{ij} = \frac{1}{1 + \exp(-(\theta_i - \theta_j))}$$

and to write the (logistic) log likelihood for n binary outcomes, $y_1, y_2, \ldots, y_n$, with $h_\theta(x_k) = 1/(1 + \exp(-\theta^\top x_k))$, as,

$$\ell(\theta|y) = \sum_{k=1}^{n} y_k \log(h_\theta(x_k)) + (1 - y_k) \log(1 - h_\theta(x_k))$$

where for match k between i and j, x_k is an p vector with ith element 1, and jth element -1, and other elements 0. Wlog, we set $\theta_0 = 0$.

This model is closely linked to the Elo model used for chess rankings via dynamic updating. Other link functions are possible, among which Cauchit seems particularly attractive, e.g. Aldous (2017).

Selection of the “Best Players”

There is a rich literature on ranking and selection with important early contributions by Bahadur, Robbins, Gupta and Portnoy, who show that ranking by best linear predictors is optimal in certain Gaussian settings, but note that in non-Gaussian settings BLUPS may go badly wrong.

Selection of the “Best Players”

There is a rich literature on ranking and selection with important early contributions by Bahadur, Robbins, Gupta and Portnoy, who show that ranking by best linear predictors is optimal in certain Gaussian settings, but note that in non-Gaussian settings BLUPS may go badly wrong. But such problems have been largely overlooked in econometrics, with present company emphatically excluded.

Regularization of the Logistic Model

If we aggregate to binomial observations for each pair we are in the relatively benign high dimensional world of $p = \mathcal{O}(\sqrt{n})$. Nonetheless, it seems ripe for regularization. An interesting penalty is the group lasso penalty of Hocking, Joulin, Bach and Vert (2011),

$$P(\theta) = \|D\theta\|_1 = \sum_{i<j} |\theta_i - \theta_j|.$$

Pairwise differences in parameters are pulled together in an attempt to identify groups of players of similar ability. The penalized log likelihood problem,

$$-\ell(\theta|X, y) + \lambda\|D\theta\|_1,$$

is convex and efficiently solved by modern interior point methods. This is closely related to total variation penalization for smoothing problems.

A Two-Step Empirical Bayes NPMLE Alternative

Another approach to regularization is to treat the unconstrained logistic MLE estimates, $\hat{\theta}$, as approximately independent draws from a heteroscedastic Gaussian sequence model.

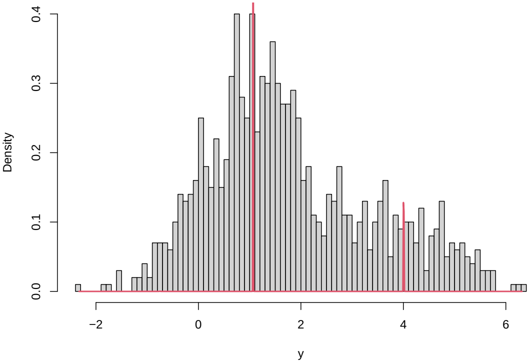
- Coordinates have different variances when the design is unbalanced.
- Each $\hat{\theta}_i$ is a draw from the density, $f_G(\hat{\theta}_i) = \int \varphi_{\hat{\sigma}_i}(\hat{\theta}_i - t)dG(t)$.
- The NPMLE of the mixing distribution G solves.

$$\min_{G \in \mathcal{G}} \{-\sum_{i=1}^{p} \log f_G(\hat{\theta}_i) \mid f_G(\hat{\theta}_i) = \int \varphi_{\hat{\sigma}_i}(\hat{\theta}_i - t)dG(t)\}$$

- The problem is convex and efficiently solved with interior point methods, K and Mizera (2014).
- Given a $\hat{G}$ we can compute posterior means, or medians, for the θ_i's,
- Which might improve upon the raw logistic MLE estimates.

A Simple Discrete Mixture Example

Consider the simple model, $Y_i \sim \mathcal{N}(\theta_i, 1)$, with $\theta \in \{1, 4\}$ with probabilities $(0.75, 0.25)$ respectively. We draw a sample of $n = 1000$, Y's, plot their histogram, and then overplot the Kiefer-Wolfowitz NPMLE in red. It is shockingly accurate!



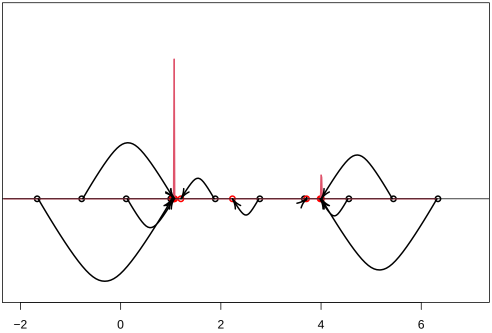
Adaptivity of the NPMLE

The Kiefer-Wolfowitz NPMLE, $\hat{G}$, is self-regularizing, that is no penalization is required to achieve a parsimonious, consistent fit.

- The non-negativity constraint on the probability masses assures only a small number of positive mass points by the Carathéodory Theorem, as already observed by Laird (1978) and Lindsay (1983).
- Even when the true mixing distribution, G, has a density, $\hat{G}$ will have a small number of discrete mass points, of order $\mathcal{O}(\log n)$ as recently shown by Polyanskiy and Wu (2020).
- This is sufficient to approximate the mixture distribution, $f = \varphi * G$ to order $o(1/n)$ in total variation, i.e. there exists an atomic, G_k with $k = \mathcal{O}(\log n)$ mass points such that $\mathsf{TV}(f, f_k) = o(1/n)$ with $f_k = \varphi * G_k$.
- For some practical decision problems, including ranking, it may be advantageous to replace, G_k, by a smoothed approximate.

Tweedie Shrinkage for Posterior Means

Given our $\hat{G}$ we can compute a posterior mean estimate for any value of y. When $\hat{G}$ is Gaussian shrinkage is linear, but otherwise can be quite highly nonlinear,



Posterior mean (Tweedie) shrinkage is quite smart about adapting shrinkage to the form of the G. Black points are shrunken to the red points.

The Machine Learning Perspective

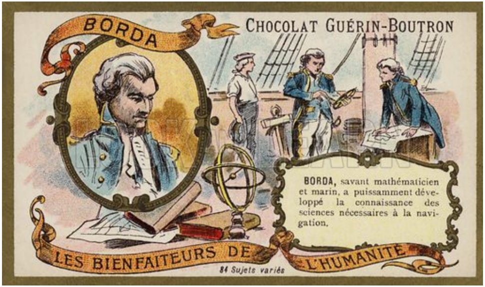
In contrast to the strict parametric, logistic, formulation of the Bradley-Terry model, there is quite an extensive literature on ranking based on pairwise comparisons in machine learning.

- While the ML approach is technically highly virtuosic, it makes what we regard as unrealistic assumptions about how the data is generated.
  - Pairwise matches occur randomly à la Erdős-Rényi resulting in an (asymptotically) balanced design.
  - True “abilities” of the competitors are well separated.
- Under such conditions it is possible to estimate ratings that recover a correct (partial) ranking asymptotically using various methods including support vector machines and Borda (1781) scores.

Notable examples: Wauthier, Jordan, and Jojic (2013), Shah and Wainwright (2018), and Chen, Gao and Zhang (2021).

Borda Scores

When match pairings are random, the vector of total “wins” for each player constitute a sufficient statistic for estimation of the “abilities.” These Borda scores play an important role in the history of social choice.


BORDA
CHOCOLAT GUÉRIN-BOUTRON
BORDA, savant mathématicien et marin, a puissamment développé la connaissance des sciences nécessaires à la navigation.
LES BIENFAITEURS DE
L'HUMANITÉ
84 Sujets variés

However, the number of i versus j matches may convey information about relative abilities of players when players of similar ability tend to be paired together as in tennis, or chess. This is the “strength of schedule” effect.

Posterior Ranks

Given an estimated mixing distribution, $\hat{G}$ we can also compute posterior mean ranks,

$$R_i = \sum_{j \neq i} \mathbb{1}\{\alpha_i \geqslant \alpha_j\}.$$

When our $\hat{\alpha}$'s are Gaussian we can approximate a Bayes rule for the quadratic loss, $\sum_{i=1}^{n} (\hat{R}_i - R_i)^2$,

$$\begin{aligned} \hat{R}_i &= \sum_{j \neq i} \mathbb{P}(\alpha_i \geqslant \alpha_j \mid \hat{\alpha}_1, \dots, \hat{\alpha}_n) \\ &= \sum_{j \neq i} \frac{\int_{\alpha_i \geqslant \alpha_j} \varphi_{ij}((\hat{\alpha}_i, \hat{\alpha}_j)) d\hat{G}(\alpha_i) d\hat{G}(\alpha_j)}{\int \varphi_{ij}((\hat{\alpha}_i, \hat{\alpha}_j)) d\hat{G}(\alpha_i) d\hat{G}(\alpha_j)} \end{aligned}$$

where $\varphi_{ij}(z)$ is a bivariate Gaussian density with mean, $\mu = (\alpha_i, \alpha_j)$ and covariance matrix, $\Sigma(i, j)$, estimated from the Hessian of the MLE.

Some Simulation Evidence

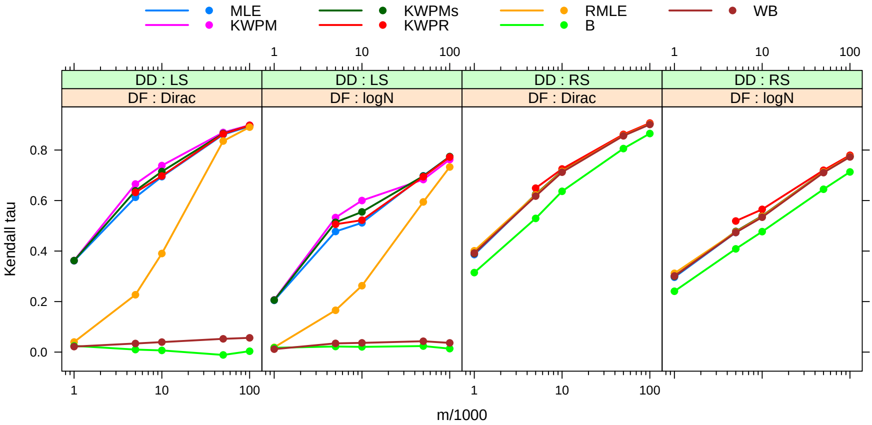
We will compare performance measured by Kendall's τ rank correlation of estimated ratings versus true ratings for 5 estimators:

- MLE: Logistic MLE
- KWPM: Posterior Means Ratings
- KWPMs: Posterior Means Ratings smoothed
- KWPR: Posterior Means Ranks
- RMLE: Group Lasso Logistic MLE
- B: Classical Borda Score
- WB: Weighted Borda Score

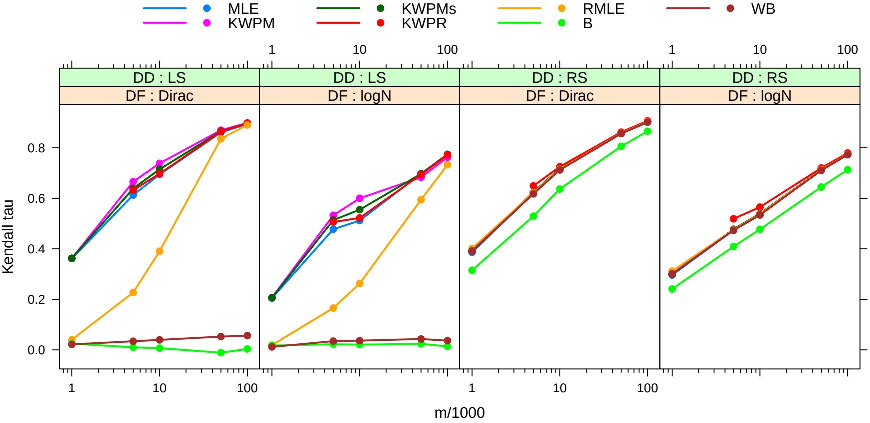
for four distinct data generating schemes:

- $\alpha \sim$ noisy mixture of two Diracs
- $\alpha \sim$ lognormal
- RS: Random matchings
- LS: Local matching

Kendall's τ Performance with 100 players



Kendall's τ Performance with 100 players



- Borda scores are terrible when matching is more probable for similar abilities.
- Regularization is somewhat helpful especially for local matching
- Group lasso is hard to tune.

The Stigler Model of Journal Influence

Stigler and Stigler (1992) and Stigler (1994) consider a Bradley-Terry model of journal influence based on pairwise citation counts.

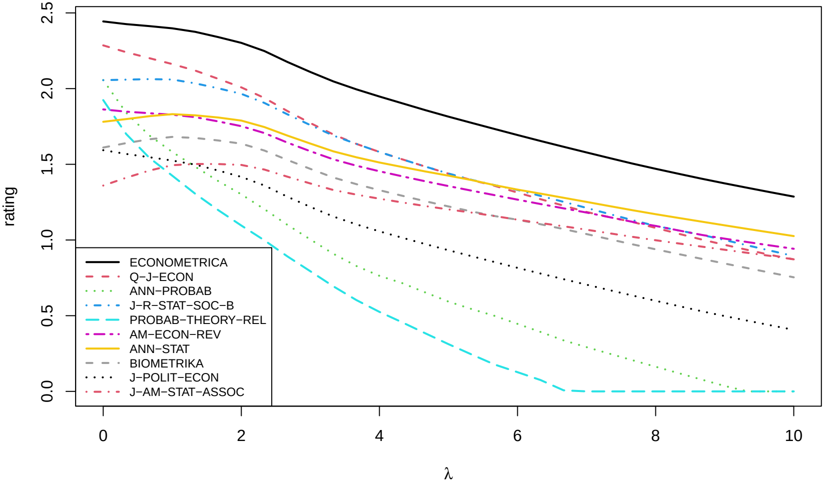
- We compare several rating/ranking methods for 86 journals in statistics and econometrics.
- Based on citation counts from 2010-2019 in the Clarivate Journal Citation Reports.
- Citation counts by journal j of papers in journal i are a measure of influence of i on j.
- These counts are binomial “wins” and “losses” for each journal pair.
- Self-citations, though interesting, are ignored in the analysis.

The Stigler Model of Journal Influence

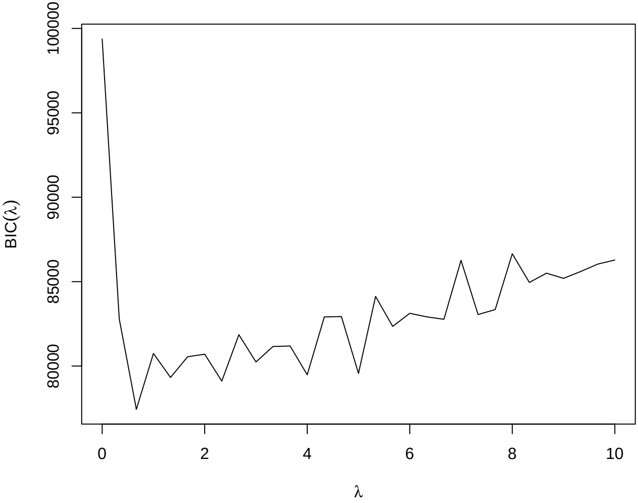
Stigler and Stigler (1992) and Stigler (1994) consider a Bradley-Terry model of journal influence based on pairwise citation counts.

- We compare several rating/ranking methods for 86 journals in statistics and econometrics.
- Based on citation counts from 2010-2019 in the Clarivate Journal Citation Reports.
- Citation counts by journal j of papers in journal i are a measure of influence of i on j.
- These counts are binomial “wins” and “losses” for each journal pair.
- Self-citations, though interesting, are ignored in the analysis.

Group/Ranking Lasso Plot



“How Do You Choose λ?”



Raw Citation Counts for Top 10 Journals

	EMCA	QJE	AnnP	JRSSB	PRF	AER	AnnS	BMKA	JPE	JASA
EMCA	387	82	2	21	0	341	60	25	204	73
QJE	90	203	0	0	0	403	0	0	172	4
AnnP	0	0	313	7	183	0	60	15	0	7
JRSSB	5	3	5	87	4	4	162	105	0	276
PRF	2	0	189	6	149	0	47	3	0	6
AER	211	234	0	2	0	785	2	0	376	10
AnnS	19	2	5	141	44	3	754	203	2	472
BMKA	3	0	0	61	5	3	134	171	3	223
JPE	131	96	0	0	0	268	0	0	201	0
JASA	14	9	3	142	8	27	231	155	8	547

$$C_{ij} = \#\{\text{times journal } i \text{ is cited by journal } j\}$$

Comparison of Journal Rankings for the Top 10

	MLE	RMLE	KWPM	Borda	WBorda
ECONOMETRICA	1	1	1	4	1
Q-J-ECON	2	2	2	5	3
ANN-PROBAB	3	3	4	13	7
J-R-STAT-SOC-B	4	4	3	7	2
PROBAB-THEORY-REL	5	5	5	19	9
AM-ECON-REV	6	6	6	1	6
ANN-STAT	7	7	7	3	4
BIOMETRIKA	8	8	8	6	5
J-POLIT-ECON	9	9	9	9	10
J-AM-STAT-ASSOC	10	10	10	2	8

Comparison of Top Ten Journal Influence Rankings for Five Methods

Comparison of Journal Rankings for the Bottom 10

	MLE	RMLE	KWPM	Borda	WBorda
INT-J-FORECASTING	77	77	79	57	71
J-HUM-CAPITAL	78	78	77	82	81
COMPUTATION-STAT	79	79	80	69	75
J-APPL-STAT	80	80	82	66	79
BRAZ-J-PROBAB-STAT	81	81	81	79	80
COMPUT-ECON	82	82	83	81	82
J-AGR-ECON	83	83	85	83	83
J-CHOICE-MODEL	84	84	84	85	84
TEST	85	85	86	84	85
J-ROY-STAT-SOC-A-STA	86	86	48	86	86

Comparison of Bottom Ten Journal Influence Rankings for Five Methods

The Lanterne Rouge

What about the hapless JRSS(A)?

- It has 51 self-cites, no others.
- MLE rating -16, but should be $-\infty$.
- Its KWPM ranking is only 48, why?
- Because its standard error is 165.
- Tweedie shrinkage moves its rating to ≈ 0.
- Q: What if it had one cite in the AER?
- A: Would move down to 83 for KWPM.
- Because standard error drops to ≈ 1.

Conclusions

- Ranking and selection are difficult with only pairwise comparisons.
- There is a demand for rankings, therefore there should be a supply.
- The grouped/ranked lasso is attractive, but needs delicate tuning.
- Tweedie shrinkage of MLE ratings is more automatic and accurate.
- Machine learning claims about Borda scores are misleading.
- The Stigler model of journal influence via citation flows is copacetic.

R package available on request.

Some References

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

Gu, J., and R. Koenker (2021a): "Invidious Comparisons: Ranking and Selection as Compound Decisions," `https://arxiv.org/abs/2012.12550`.

——— (2021b): "Ranking and Selection from Pairwise Comparisons: Empirical Bayes Methods for Citation Analysis," `https://arxiv.org/abs/2112.11064`.

Hocking, D., A. Joulin, F. Bach, and J.-P. Vert (2011): "Clusterpath: an algorithm for clustering using convex fusion penalties," *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 745–752.

Mosek ApS (2021): "MOSEK Modeling Cookbook," Available from `http://www.mosek.com`.

Shah, N., and M. Wainwright (2018): "Simple, robust and optimal ranking from pairwise comparisons," *Journal of Machine Learning Research*, 18, 1–38.

Stigler, S. (1994): "Citation Patterns in the Journals of Statistics and Probability," *Statistical Science*, 9(1), 94 – 108.

Varin, C., M. Cattelan, and D. Firth (2016): "Statistical modelling of citation exchange between statistics journals," *J. Royal Statist. Soc. A*, 179, 1–63.