RATING AND RANKING WITH PAIRWISE COMPARISONS

ROGER KOENKER

Abstract. A collection of functions and data for estimation of ratings and ranking models based on pairwise comparison data. Included is data and code for ranking journals in econometrics and statistics based on the Bradley-Terry model of journal influence of Stigler.

1. Introduction

It is common practice to construct ratings and rankings of competitors based on pairwise comparisons. The **RRpairwise** package provides some functionality for this purpose. The basic building block is the well-known Bradley and Terry (1952) model and its associated logistic maximum likelihood estimator. In this model, players are assumed to possess a scalar rating, or "ability", $\alpha_i : i = 0, \ldots, p$ and the probability that player i defeats player j in any given match is given by,

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

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

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 $\theta_0 = 0$.

Since $p = \mathcal{O}(\sqrt{n})$ it seems prudent to consider some form of regularization of the unconstrained maximum likelihood estimates. The package offers several options of this type.

The group lasso penalty of Hocking et al. (2011),

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

tries to pull together pairwise differences in parameters in an an attempt to identify groups of players of similar ability. The penalized log likelihood problem,

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

December 29, 2021.

is convex and efficiently solved by modern interior point methods. This is closely related to total variation penalization used for many smoothing problems. This approach is illustrated below in our application to journal citation patterns.

Another approach to regularization is to view the MLE α_i's as approximately independent and Gaussian and generated by a heteroscedastic Gaussian mixture model. A two-step estimation method is used to estimate the mixing distribution, $\hat{G}$, and then posterior mean ratings or ranks can be computed.

The foregoing methods are all incorporated in the function `BTfit` and rely on convex optimization routines from the **REBayes** package and Mosek ApS (2021). In addition, `BTfit` incorporates options to evaluate so-called Borda scores that simply count the number of "wins" in the accumulated matches of each player and use that to rank players.

2. A Simulation Exercise

To evaluate performance of these methods we need to choose two aspects of the experiment: first how the α's are generated; and second, how players are matched. In the experiments reported in Gu and Koenker (2021) we consider two variants for each aspect. The α's are either generated as a noisy mixture of two Dirac's, or as a lognormal, and matching occurs "at random," designated as "RS" or in such a way that players of similar ability are more likely to meet more frequently, designated "LS". The details of the latter option are spelled out in the code for the function `DGP`. To facilitate the computations it is assumed that the players are ordered by ability in the simulation setting.

The following code illustrates one version of the simulation setup for Dirac α's and local matching. Note that the α's are normalized so the smallest element is 1. The other three instances of the simulation simply change how the vector `a` is generated and/or change the call to `DGP` to specify `type = "RS"`

```
> sessionInfo()
> set.seed(1729)
> meths <- c("MLE", "KWPM", "KWPMs", "KWPR","RMLE", "B", "WB")
> ms <- c(1000, 5000, 10000, 50000,100000)
> n <- 100
> R <- 100
> K <- array(0, c(length(meths), length(ms), R))
> kcor <- function(x,a) cor(x, a, method = "kendall")
> for(j in 1:length(ms)){
+   for(i in 1:R){
+     a <- sample(c(4,8), n, prob = c(0.8, 0.2), replace = TRUE) + rnorm(n, sd = 1/3)
+     a <- sort(a)
+     a <- a/a[1]
+     A <- matrix(0,n, length(meths))
+     D <- DGP(a, ms[j], type = "LS")
+     for(k in 1:length(meths)){
+         A[,k] <- BTfit(D, method = meths[k])
+         K[,j,i] <- apply(A, 2, kcor, a = a)
+         }
+   }
+ }
```

3. Journal Ranking and Citation Analysis

As described in Gu and Koenker (2021), one way to measure the influence of academic journals, proposed by Stigler (1994), involves tracing the flow of citations from one journal to another. Journals that are frequently cited by other journals are influential, journals that are less frequently cited by other journals are not. The **RRpairwise** package includes data and code for exploring this approach based on the Clarivate Journal Citation Reports. The data consists of cross journal citation counts for the period 2010-2019 for $J = 86$ selected journals covering ecoonometrics and statistics.

The **RRpairwise** package contains both raw data from this source and a cooked version of the data in the form of binomial counts for each pair of journals that exchanged any citation references. The raw data is stored in the directory `inst/extdata` and can be accessed by the function `Cites`. The cooked form of the data is available in compressed form in the data directory, and can accessed by the invocation `data(citations)` as illustrated in the following example.

We begin by illustrating the group lasso fitting approach. The function `BTfit` fits a group lasso model for a range of λ's from 0 to 10. The resulting lasso plot shows the trajectories of the top 10 ranked journals according to the unconstrained logistic maximum likelihood estimates. Econometrica is the dominant journal exporting more citations than it imports. Two probability journals are also strong performers based on the unconstrained fit, but their ratings decline rapidly as λ increases. Since one of the α parameters is normalized to 1, its corresponding θ is 0, and consequently as $\lambda \to \infty$ all the θ parameters are pulled toward 0 as in the classical lasso, but the nature of the shrinkage is quite different than the classical version of the lasso. The trajectories plotted in Figure 1 reflect this tendency toward zero, however there is also an option to `BTfit` to request that the model is refit with the estimated grouping imposed; a tolerance for the grouping is specified by setting the paraeter `refit` to an integer that determines the degree of rounding used for the grouping. Choice of the grouping parameter λ is inevitably a headache; it is recommended to use something like the BIC criterion that penalizes the log likelihood of the fit by the number of estimated parameters, i.e. the number of estimated groups identified at each λ.

```
> PL10cap <- "Group Lasso Plot for Citation Data"
> data(citations)
> lambdas <- 0:30/3
> ahat0 <- BTfit(citations)
> A <- matrix(0, length(ahat0), length(lambdas))
> A[,1] <- ahat0
> for(i in 2:length(lambdas))
+     A[,i] <- BTfit(citations, method = "lasso", lambdas[i], refit = 0)
> o <- rev(order(A[,1]))
> B <- head(A[o,],10)

> Citing <- Cites(type = "Citing")
> Top10 <- dimnames(Citing)[[1]][o[1:10]]
> matplot(lambdas, t(B), type = 'l', xlab = expression(lambda),
+         ylab = 'Rating', lwd = 2, lty = 1:10, col = 1:10)
> legend("bottomleft", legend = Top10, lwd = 2,
+        lty = 1:10, seg.len = 3, col = 1:10, cex = .75)
```

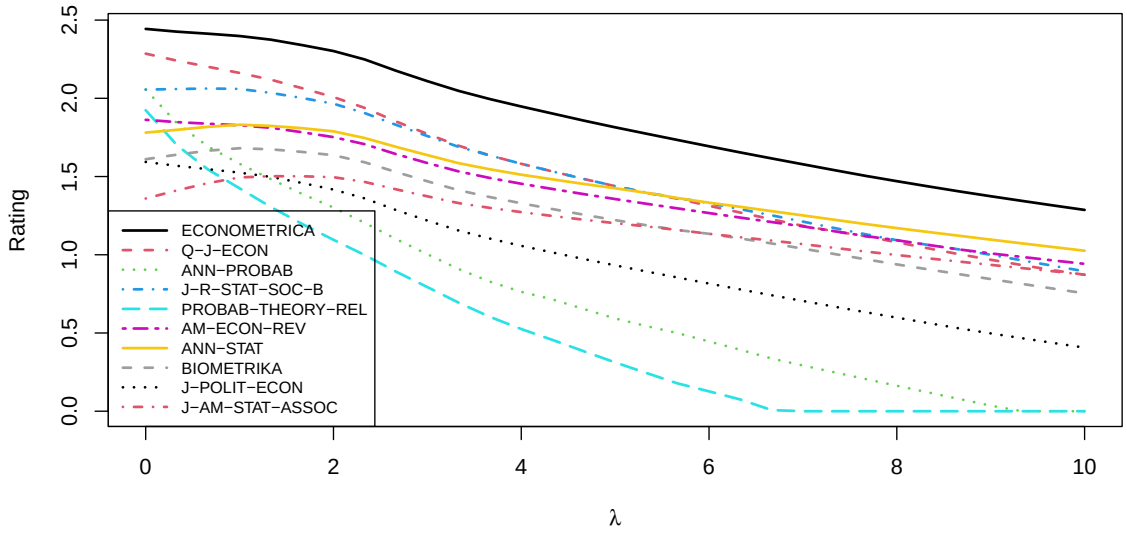

FIGURE 1. Group Lasso Plot for Citation Data

Various other fitting methods can be selected in the `BTfit` function resulting in alternative rating and ranking estimates. As noted in Gu and Koenker (2021) these alternative rankings are all quite similar to the unconstrained MLE rankings except for the Borda procedures which are a bit wonky.

4. Conclusion

This vignette is an attempt to describe some basic features of R package **RRpairwise** designed to explore some regularization schemes for estimating Bradley-Terry type models for ratings based on pairwise comparison data. Commments on any or all aspects would be most welcome.

References

Bradley, R. and Terry, M. (1952), 'Rank analysis of incomplete block designs: I. the method of paired comparisons', *Biometrika* **39**, 324–345.

Gu, J. and Koenker, R. (2021), 'Ranking and selection from pairwise comparisons: Empirical bayes methods for citation analysis'. Available from: `https://arxiv.org/abs/2112.11064`.

Hocking, D., Joulin, A., Bach, F. and Vert, J.-P. (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`.

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