An Application of Endpoint Detection to Bivariate Data in Tau-Path Order

Srinath Sampath
Hamilton Capital Management
Columbus, OH
sampath.5@osu.edu

Joseph S. Verducci
Department of Statistics
The Ohio State University
jsv@stat.osu.edu

September 25, 2014

ABSTRACT

The Fligner and Verducci (1988) multistage model for rankings is modified to create the moving average maximum likelihood estimator (MAMLE), a locally smooth estimator that measures stage-wise agreement between two long ranked lists, and provides a stopping rule for the detection of the endpoint of agreement. An application of this MAMLE stopping rule to bivariate data set in tau-path order (Yu, Verducci and Blower (2011)) is discussed. Data from the National Cancer Institute measuring associations between gene expression and compound potency are studied using this application, providing insights into the length of the relationship between the variables.

KEYWORDS: partial rankings, top-K rank list, multistage model, maximum likelihood estimation, stopping rule, tau-path

1. INTRODUCTION

This paper provides a brief description of the application of the moving average maximum likelihood estimator (MAMLE) as a stopping rule that determines the length of relationship between two long ranked lists. The MAMLE algorithm is described in detail in Sampath and Verducci (2013); it adapts the Fligner and Verducci (1988) multistage ranking model into a locally smooth estimator to measure stage-wise agreement between two ranked lists using a rolling maximum likelihood estimation approach. The efficacy of the stopping rule is demonstrated on data maintained by the National Cancer Institute, where gene expression and treatment potency levels are examined to determine the endpoint of their relationship. This paper is meant to serve merely as a companion to the presentation made at the 2014 Joint Statistical Meetings in Boston; full details of the methodology are provided in Sampath and Verducci (2013).

Briefly, the MAMLE algorithm provides a stage-wise estimate of agreement between two long lists of ranks. Known as the Top-K problem in the literature, the algorithm attempts to answer the question: given two long ranked lists, at which stage does signal degenerate into noise? If two assessors were to independently order a long list of items, the MAMLE algorithm anchors one list of ranks as the reference ranking, and computes stage-wise deviations of the second assessor's ranks from the reference ranking. These deviations are used as penalties in a truncated geometric distribution framework to arrive at a nuanced estimator of the stage-wise level of agreement between the two assessors.

The focus here is on the use of the MAMLE approach as a stopping rule to determine the endpoint of association for data organized in tau-path order. Yu et al. (2011) use Kendall's tau measure to extract strongly associated subsets from bivariate data, and this method is described next.

2. THE TAU-PATH METHOD

Natural phenomena are replete with instances where a modest overall association between two variables masks a strong association between these variables in a subset of the population. If it were possible to isolate such a subset, subsequent analyses could potentially lead to a better understanding of the underlying phenomenon that caused this association.

The tau-path algorithm described by Yu et al. (2011) arranges bivariate data (X, Y) in decreasing order of Kendall's tau measure. Early in the ordered list, the contributions to correlation will be high. If, further down this ordered list, contributions to the correlation decrease to 0 for a stretch of data, it can be inferred that the association between the two variables is confined to the earlier subpopulation. This would cause the sample tau coefficients for the sequence of ordered and growing subsets to form a monotone decreasing path toward Kendall's tau.

Using the notation of Yu et al. (2011), the probabilities of concordance and discordance for independent variables (X_1, Y_1) and (X_2, Y_2) from a bivariate distribution are, respectively,

$$\begin{aligned} p_c &= P[(X_1 - X_2)(Y_1 - Y_2) > 0] \\ p_d &= P[(X_1 - X_2)(Y_1 - Y_2) < 0] \end{aligned}$$

Kendall's tau for the distribution is given by $\tau = p_c - p_d$, and has an unbiased estimator in Kendall's tau coefficient, which, for a random sample of n pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, is given by

$$T = \Big(\sum_{1 \le i < j \le n}\sum c_{[i,j]}\Big) / \binom{n}{2}$$

where

$$c_{[i,j]} = \begin{cases} 1, & \text{if the } (i,j)\text{th pair is concordant} \\ -1, & \text{if the } (i,j)\text{th pair is discordant.} \end{cases}$$

The goal, then, is to find the permutation π such that the sequence of tau coefficients computed from the increasing stream of subsets of the bivariate data in π order is sequentially maximal monotone decreasing. Such a sequence of tau coefficients is denoted the tau-path and is determined by either of the two backward conditional search algorithms given by Yu et al. (2011).

3. ENDPOINT DETECTION IN NCI-60 DATA

If it were possible to transform the tau-path into a stage-wise penalty function of the form used by the MAMLE algorithm, a stopping rule to detect the endpoint of association of the bivariate data in tau-path order could be engineered. Denoting the number of concordant and discordant pairs of observations in tau-path order, respectively, $A_k(\pi)$ and $D_k(\pi),\ k = 1, \ldots, n$, Kendall's tau coefficient for the first k observations in the permuted sample is simply

$$T_k(\pi) = \Big(A_k(\pi) - D_k(\pi)\Big) / \binom{k}{2},\ k = 2, \ldots, n.$$

Since $A_k(\pi) + D_k(\pi) = \binom{k}{2},\ k = 2, \ldots, n,$

$$D_k(\pi) = \binom{k}{2}\left(\frac{1 - T_k(\pi)}{2}\right),\ k = 2, \ldots, n.$$

Finally, the stage-wise penalty function becomes

$$V_k(\pi) = D_k(\pi) - D_{k-1}(\pi),\ k = 3, \ldots, n.$$

The NCI-60 data maintained by the National Cancer Institute consists of five genes *ASNS*, *IGFBP6*, *LDOC1*, *NQO1* and *PIK3R3*, and five cancer treatments 40, 757, 1771, 2039 and 3062, each measured across 60 cancer cell-lines. The tau-path and the order of observations from strongest to weakest association were extracted for all 25 gene-compound pairs, and these were fed into the MAMLE algorithm to determine the endpoint of agreement, using a window width of five stages. The null distribution for the MAMLE curve was generated next, assuming that the second assessor's ranking scheme was completely random relative to the first, and these random permutations were fed first into the tau-path algorithm and the latter outputs into the MAMLE algorithm. Figure 1 shows the rejection region, formed by the stage-wise 95th quantiles of the MAMLEs computed from the simulations.

For one of the gene-compound pairs – gene LDOC1 and compound 1771 – the MAMLE approach with a window width of five stages suggests that the endpoint is best estimated by stage 28 (Figure 2). When using a window width of ten stages, the algorithm estimates the endpoint as stage 27, suggesting that the subset with strongest association ends about halfway into the bivariate data series.

4. DISCUSSION

The MAMLE methodology discussed above was developed in the `R` programming environment, version 3.0.1 (R Core Team 2013). Even long lists are handled efficiently by this algorithm. The application of this methodology as a stopping rule on data in tau-path order provides a reliable method to estimate the length of the strongly associated subset in bivariate populations.

ACKNOWLEDGEMENTS

We thank the Nationwide Center for Advanced Customer Insights for research support for the first author, and the National Institutes of Health for partial support for the second author.

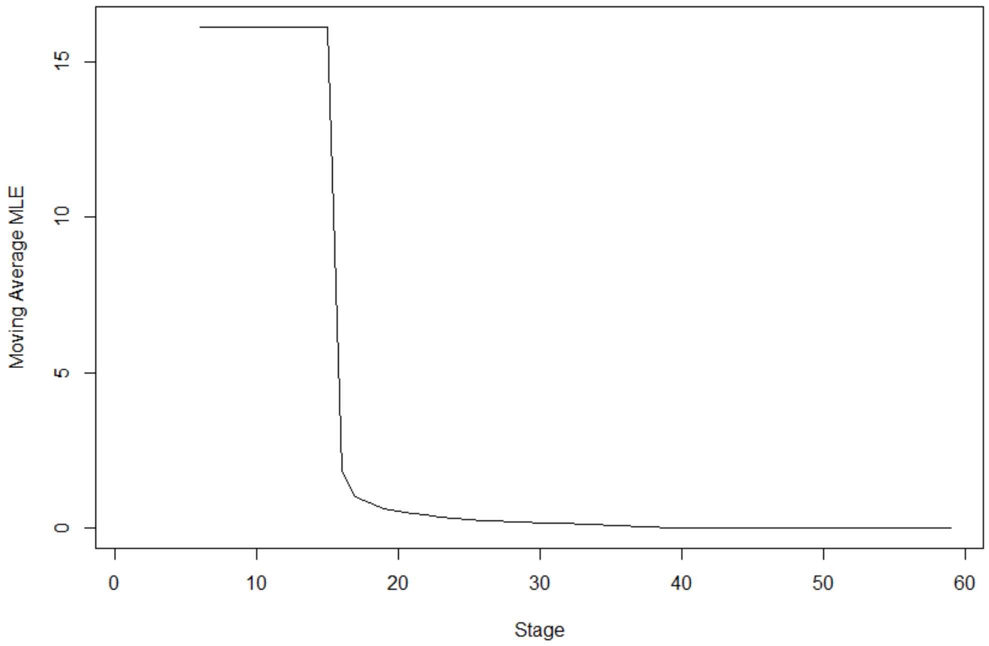

Figure 1: Stage-wise 95th quantile MAMLEs under the null distribution of NCI-60 gene-compound data, using a window width of five stages.

REFERENCES

Fligner, M. A., and Verducci, J. S. (1988), “Multistage Ranking Models,” *Journal of the American Statistical Association*, 83(403), 892–901.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`

Sampath, S., and Verducci, J. S. (2013), “Detecting the End of Agreement Between Two Long Ranked Lists,” *Statistical Analysis and Data Mining*, 6(6).

Yu, L., Verducci, J. S., and Blower, P. E. (2011), “The Tau-Path Test for Monotone Association in an Unspecified Population: Application to Chemogenomic Data Mining,” *Statistical Methodology*, 8, 97–111.

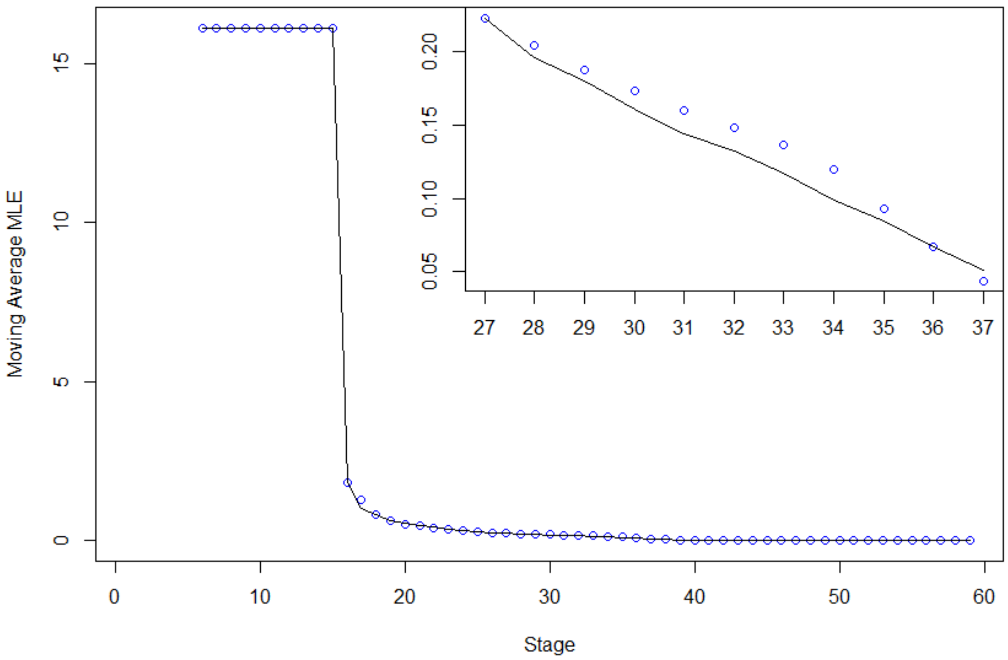

Figure 2: Comparison of gene LDOC1 and compound 1771 association with the null distribution, using MAMLEs with a window width of five stages. Solid black line: stage-wise 95th quantile MAMLEs under the null distribution. Blue circles: MAMLEs for association between gene LDOC1 and compound 1771.