How Stable Are Top Choices Over Time? An Investigation Into Preferences Among Popular Baby Names In The United States

Srinath Sampath* Joseph S. Verducci †

Abstract
For the problem of assessing initial agreement between two rankings of long lists, inference in the Fligner and Verducci (1988) multistage model for rankings is modified to provide a locally smooth estimator of stage-wise agreement. An extension to the case of overlapping but different sets of items in the two lists, and a stopping rule to identify the endpoint of agreement, are also provided. Simulations show that this approach performs very well under several conditions. The methodology is applied to a database of popular names for newborns in the United States and provides insights into trends as well as differences in naming conventions between the two sexes.

Key Words: partial rankings, top-K rank list, multistage model, maximum likelihood estimation, stopping rule, consensus

1. Introduction

The measurement of agreement between ranked lists of objects has witnessed many approaches, and has experienced a resurgence following interest in search algorithms. Hall and Schimek (2012) provide a review of the Top-K problem and propose a 0-1 scoring algorithm for matched ranks that uses an elegantly derived moderate deviation bound to assess descent into randomness. The alternative approach investigated here is based on the multistage ranking model of Fligner and Verducci (1988). This new approach incorporates information about the extent of agreement at each stage, and extends the applicability of Top-K methods to ranking lists that may contain overlapping but different sets of items. The statistical properties of the approach are illustrated through simulations, and the method is used to analyze trends in naming newborns in the United States over the last century.

The two rankings themselves are assumed to be ordered, for example by years, with choices made in the current year being compared with the list from the previous year. Under the new method, agreement is assessed in stages determined by the ordering of items in the current year. This ordering starts with the most favored item of the current year and proceeds to the least favored item. At each stage, all remaining items (unselected in the current year) are ranked according to their relative ranking from the previous year. Agreement of the current choice with the previous year's ranking is assessed by its position in this relative ranking. The actual algorithm described in the next section provides a computationally efficient way of calculating these stage-wise agreements.

The main goal is to find the point, if any, where the agreement ceases and the remaining choices become no more than random. To this end a novel stopping rule is proposed that identifies the point where the signal turns to noise. This point estimates the K in the Top-K problem.

*Department of Statistics, The Ohio State University, 1958 Neil Avenue, 404 Cockins Hall, Columbus, OH 43210-1247

†Department of Statistics, The Ohio State University, 1958 Neil Avenue, 404 Cockins Hall, Columbus, OH 43210-1247

An innovation in this paper is the adjustment applied to the stage-wise penalties to handle overlapping but different sets of ranked items in the two lists. A two-part penalty is applied to the newcomers to properly account for their presence in the new list.

A treasure trove of high-quality data on the most popular baby names in the United States, dating back to 1880 and categorized by birth year, is provided by the Social Security Administration. The data set contains the most popular baby names of each sex on an annual basis in a readily accessible format. A detailed analysis of these data using our methodology uncovers rich relationships in multiple dimensions: a) the naming convention differences between the sexes, b) the naming conventions in a given year, and c) the changing trends across years.

The extension of the multistage model developed in Fligner and Verducci (1988) used to build our analytical platform is provided in Section 2. In Section 3, moving average estimators are devised to reduce the variance of maximum likelihood estimators; two-part penalties are proposed to handle the case of overlapping but different lists; and a stopping rule is defined to estimate the point at which stage-wise agreement ends. Simulations that mimic the baby name database are used to evaluate our model's efficacy in Section 4. Section 5 describes the Social Security Administration baby names data set and presents the results discovered by our methodology. A concluding discussion and some questions raised by both the data and our analysis, which will foster future research, are presented in Sections 6 and 7 respectively.

2. Extending The Multistage Model To Estimate Agreement

In this section, Fligner and Verducci (1988)'s multistage model is adapted to capture the agreement between two long lists.

A *ranking* or *permutation*

$$\pi = [\pi(1), \ldots, \pi(n)]$$

of n distinct objects is a vector of length n, with each component corresponding to an object, and the value of the component being the rank of that object, namely the quantity given by (1 plus the number of other objects that are considered superior). Thus $\pi(i)$ represents the rank of the i-th object. For example, if $\pi = [3, 2, 4, 1]$, the rank of the first object is 3, the rank of the second object is 2, the rank of the third object is 4, and the rank of the fourth object is 1.

An *ordering* or *inverse permutation* of n objects, labeled 1 to n, is a vector of length n, with each component i giving the label of the object that has rank $i,\ i = 1, \ldots, n$. The ordering or inverse permutation associated with π is specified by the mapping

$$\pi^{-1}(j) = i, \text{ if } \pi(i) = j,\ i = 1, \ldots, n,\ j = 1, \ldots, n.$$

The above example can be rewritten equivalently as $\pi^{-1} = [4, 2, 1, 3]$, since the object ranked first, namely "4," occupies the first position of the vector π^{-1}; the object ranked second, namely "2," occupies the second position; the object ranked third, namely "1," occupies the third position; and the object ranked fourth, namely "3," occupies the last position.

The baby names data corresponding to two consecutive years can be assumed to represent n objects ordered sequentially according to two independent processes, each representing one year. The earlier year provides the *reference ranking* or *ground truth*, and the later year provides the *observed* or *generated ranking*. The two ranking processes may be initially governed by common parameters and, as one moves further down the list, the ranks can be expected to diverge from each other. A cursory analysis of the baby names data reveals that the most popular names return year after year, often in the same positions, whereas less popular names are not as strongly ordered, having to contend with several close competitors to successfully maintain their ranks. The stage where the two lists become completely uninformative about each other is denoted K, and its estimate measures the length of agreement between the two lists.

The probability model for the penalties is motivated by the following example from the baby names data. In 2010, the dozen most popular male baby names in the United States were, in order,

Jacob, Ethan, Michael, Jayden, William, Alexander, Noah, Daniel, Aiden, Anthony, Joshua, Mason,

while the dozen most popular male names in 2011 were, in order,

Jacob, Mason, William, Jayden, Noah, Michael, Ethan, Alexander, Aiden, Daniel, Anthony, Matthew.

Penalties are now assigned to each of the names from year 2011 based on their discordance with the names from year 2010. Since the most popular name in 2011, *Jacob*, matches the most popular name in 2010, the penalty for the first *stage* is 0. In the second stage, the 2011 pick of *Mason* was the twelfth most popular name in 2010. Since *Jacob* has already been accounted for in 2011, ten other more popular names from 2010 were overlooked before *Mason*, and therefore the penalty for the second stage is 10. Similarly, the penalty for the third stage is 3 (since *Ethan*, *Michael* and *Jayden* were ignored before picking *William*). Continuing in this manner, the stage-wise penalties for the dozen names are

[0, 10, 3, 2, 3, 1, 0, 0, 1, 0, 0, 4].

This penalty assignment can now be generalized and assigned a truncated geometric probability distribution as given below. For illustrative purposes, we describe the calculations for the first stage, and then generalize to the other stages.

Stage 1: Here all n names from the reference year (2010 in the above example) are available for the observed year (2011 in the above example) to choose from. The observed year selects the $(1+v)$th best name overall, as specified by π^{-1}, and incurs the penalty $V_1 = v$, with truncated geometric probability

$$P(V_1 = v) = \left(\frac{1-r_1}{1-r_1^n}\right) r_1^v,\ v = 0, \ldots, n-1,\ 0 < r_1 < 1. \tag{1}$$

Stage j ($j = 2, \ldots, n-1$): Here $n-j+1$ names are available. The observed year picks the $(1+v)$th best name available, as specified by π^{-1}, and incurs a penalty $V_j = v$, with truncated geometric probability

$$P(V_j = v) = \left(\frac{1-r_j}{1-r_j^{n-j+1}}\right) r_j^v,\ v = 0, \ldots, n-j,\ 0 < r_j < 1. \tag{2}$$

The r_j's are unknown and need to be estimated, while the penalties v will be supplied by the data. It is reasonable to assume independent choices at each stage of the ranking process, which causes the $\{V_j \mid j = 1, \ldots, n-1\}$ to become independent. $\{V_1, \ldots, V_{n-1}\}$ is therefore the *penalty* or *discordance vector* between the reference year and the observed year. Since the probabilities in Eq. (1) and Eq. (2) are decreasing functions of v, the model does indeed recognize that higher penalties at any stage are attained with lower probabilities, and furthermore, at every stage it penalizes the observed year appropriately for larger departures from the reference year.

The limiting distribution of each V_j as $r_j \to 1$ is discrete uniform on the set $\{0, \ldots, n-j\}$, $j = 1, \ldots, n-1$, which removes all ranking 'skill' from the observed year with respect to the reference year. At every such stage, the observed year is assumed to select a name from the remaining names at random. A visual analysis of a graph of the stage-wise estimated r_j's can reveal the point where the selections made in the observed year go from deliberate to random. This stage, where the tide turns, is a useful estimate of K, which measures the length of the agreement between the popular names in the two years under study.

For graphical and mathematical reasons, it is easier to work with

$$\theta_j = -\log r_j,\ j = 1, \ldots, n-1,$$

and study the behavior of the θ_j's in lieu of the r_j's. The condition $r_j \to 1$, which leads to the limiting uniform distribution for the V_j's, is now equivalent to the condition $\theta_j \to 0$. Visually, the earliest stage where the θ_j's descend to 0, and remain predominantly 0's from that stage forward, is a good estimate of K. More specifically, determining the value of K where the agreement between the two years ceases, is equivalent to determining the value of K for which $\theta_K > 0$, and $\theta_j = 0$ for all $j > K$. Once past this K, the observed year and the reference year are no longer in agreement.

3. Parameter Estimation, Penalty Assignment, and the Stopping Rule

In this section we modify the estimation of the parameters $\{\hat{\theta}_j\}$ of the multistage ranking model from the usual maximum likelihood estimators to more stable estimators known as moving average maximum likelihood estimators (MAMLEs), which are compelled by the constraint that only a single random ranking is observed. This is due to the fact that the reference year's ranks are anchored as the ground truth, and only the stage-wise deviations of the observed year's ranks from those of the reference year are considered by the algorithm. The underlying idea is that the initial θ_j's should be positive and vary slowly with j, and the later θ_j's should be close to 0. Thus nearby stages should be helpful in estimating θ_j.

Recall that the $r_j,\ j = 1, \ldots, n-1$, are stage-wise measures of the observed year's name agreement with the reference year. Since the probability mass function of the penalties at stage j is inversely proportional to r_j, a lower r_j leads to higher agreement between the two years. It can be reasonably assumed that the r_j's vary gradually from one stage to the next, but not over the changepoint to noise. Specifically, to determine the MAMLE $\hat{r}_j$ for a given stage j, the assumption of a common value r for all the unknown r_i in the backward-looking window of the form $j - w + 1 \leq i \leq j$ of width w, simplifies the computation. The set of MAMLEs, $\hat{r}_j$, determined with this assumption, uses overlapping rank data as the window advances down the stages. An examination of this MAMLE curve across the stages provides a clear understanding of the strength and length of the agreement between the two lists of names.

3.1 Locally Smooth Estimation of Agreement

The MAMLE $\hat{r}_j$ of the parameter r_j is determined from the window $j - w + 1 \leq i \leq j$ as follows. The local likelihood function of the fixed r in the window is given by

$$\begin{aligned} L(r) &= P(V_{j-w+1} = v_{j-w+1}) \times \cdots \times P(V_j = v_j) \\ &= \left(\frac{1-r}{1 - r^{n-(j-w+1)+1}} \right) r^{v_{j-w+1}} \times \cdots \times \left(\frac{1-r}{1-r^{n-j+1}} \right) r^{v_j}. \end{aligned}$$

The local log likelihood is therefore

$$\log L(r) = w \log(1-r) + (\log r) \sum_{i=j-w+1}^{j} v_j - \sum_{k=n-j+1}^{n-j+w} \log(1 - r^k). \tag{3}$$

Differentiating Eq. (3) with respect to r gives the MAMLE $\hat{r}_j$ as the solution to the equation $\overline{V_j} = g_j(r_j)$, where

$$\overline{V_j} = \frac{1}{w} \sum_{i=j-w+1}^{j} V_i, \tag{4}$$

is the mean penalty for the window, and

$$g_j(r) = \frac{r}{1-r} - \frac{1}{w} \sum_{k=n-j+1}^{n-j+w} k \frac{r^k}{1-r^k}$$

is an increasing function of r. Since both the first term on the right hand side and the ratio in the summand take the form of odds ratios, the computation of r becomes straightforward.

3.2 Assigning Penalties to Overlapping but Different Lists

The paper so far has only considered the possibility that the names in the reference and observed years are identical, albeit ordered differently. But an analysis of the full Social Security Administration baby names data reveals that the median number of new female and male name entrants every year are 56 and 53 respectively. It is therefore prudent to assess a two-part penalty on the newcomers; the *position penalty* that considers the stage at which a new name enters the list during the observed year, and the *median penalty* that reflects the size of the group of new names, and which accounts for the fact that the relative ranks among the newcomers are unknown. Specifically, if n_{new} newcomers enter the database for a given sex in a given year, a newcomer that occupies position m on the list will earn a position penalty of $(n - m + 1)$ for 'wrongly' superseding that many names from the reference year, plus a median penalty of $(n_{new} + 1)/2$ to account for the size of the incoming block of names.

3.3 The Stopping Rule

The calculations in Subsection 3.1 provide MAMLEs for the stage-wise r_j's, which in turn are used to compute the stage-wise $\hat{\theta}_j$'s. To determine the rejection bound of the MAMLEs $\{\hat{\theta}_j\}$, a large number of simulations are generated from the multistage model assuming that the θ_j's are all 0 (which is equivalent to all permutations having the same probability $1/n!$), compute the stage-wise $\hat{\theta}_j$'s for each simulation, and, for each stage j, plot $q(j)$, the

$(1-\alpha)$th quantile of $\hat{\theta}_j$. Then the *stopping stage* or *endpoint* K is estimated by

$$\hat{K} = \text{the earliest stage at which } \hat{\theta}_{\hat{K}+w} > q(\hat{K}), \text{ and } \hat{\theta}_j > q(j) \text{ for at most } \alpha \text{ percent of the remaining } j > \hat{K}+w. \quad (5)$$

The reason for moving the region of θ_j's to $\{\hat{K}+w,\ldots,n\}$ is that when $\hat{K}=K$, all $\hat{\theta}_j,\ j>\hat{K}$ are generated by pure noise, but earlier $\hat{\theta}_j,\ j\leq\hat{K}$ are not. This last feature happens because $\hat{\theta}_j$ is based on the previous $w-1$ observations through $\{V_{j-w+1},\ldots,V_j\}$.

The graph of estimated parameters $\{\hat{\theta}_j\}$ should be used as a diagnostic for both window width selection and modification of the estimator $\hat{K}$. In particular we recommend:

1. If the graph of $\{\hat{\theta}_j\}$ is erratic, increase window width.
2. Check if there is a noticeable drop in the $\{\hat{\theta}_j\}$ for $j\in\{\hat{K}-w,\ldots,\hat{K}\}$. If not, increase the window width; $\hat{K}$ is artificially large due to a random tail event.
3. If there is a very steep drop in the $\{\hat{\theta}_j\}$ curve at a single point K^*, choose K^* rather than $\hat{K}$, which is likely to underestimate K due to bias.

We offer these as guidelines until an automatic implementation of the stopping rule can be found. We have checked that the current $\hat{K}$ is already fairly robust to the choice of $\alpha\in(0,1/2)$, so the user need focus only on window width w in estimating the point K where agreement ends.

In the context of baby names, an analysis of the trend in endpoints for each sex over the last century, coupled with a comparison of the endpoints between the sexes across the years, can provide insights into the naming conventions observed by new parents in the United States.

4. Analysis of the Multistage Model using Simulations

A theoretical curve based broadly on actual male name data from the 1950s was created to test the efficacy of our methodology. The solid black line in Figure 1 shows the generated theoretical curve and has the following characteristics: a linear descent over the first half of the stages to represent the gradually declining agreement between the two lists, a cliff event that causes an instantaneous discordance between the two lists, and finally a flat region that represents the random assignment of ranks in the remaining stages. A single instance of the observed list was simulated following Equations (1) and (2) and the probabilities computed from this theoretical curve.

To further match this single simulation with actual data, the number and positions of new names entering every observed year of the Social Security Administration data were exhaustively recorded. Since the median number of new male names entering the database was 53, the histogram of new name positions across the years (10,157 names across 539 unique positions) was sampled to simulate the positions where 53 new names were to enter the simulated observed list. These new names then received both position and median penalties in the manner described in Section 3.2. Lastly, the average penalty for a given window was calculated using a Winsorized mean with trim 1/(window width) to smooth the effect of the larger penalties, but to preserve the impact of a cluster of new names, should more than one enter a window. Implementing the Winsorized mean also gives the user the flexibility to modify the trim assumption for a smoother or rougher fit, as desired.

Figure 1 depicts two MAMLE curves calculated using the above method and different window widths, 40 (dashed red line) and 80 (solid blue line). It is evident that both curves

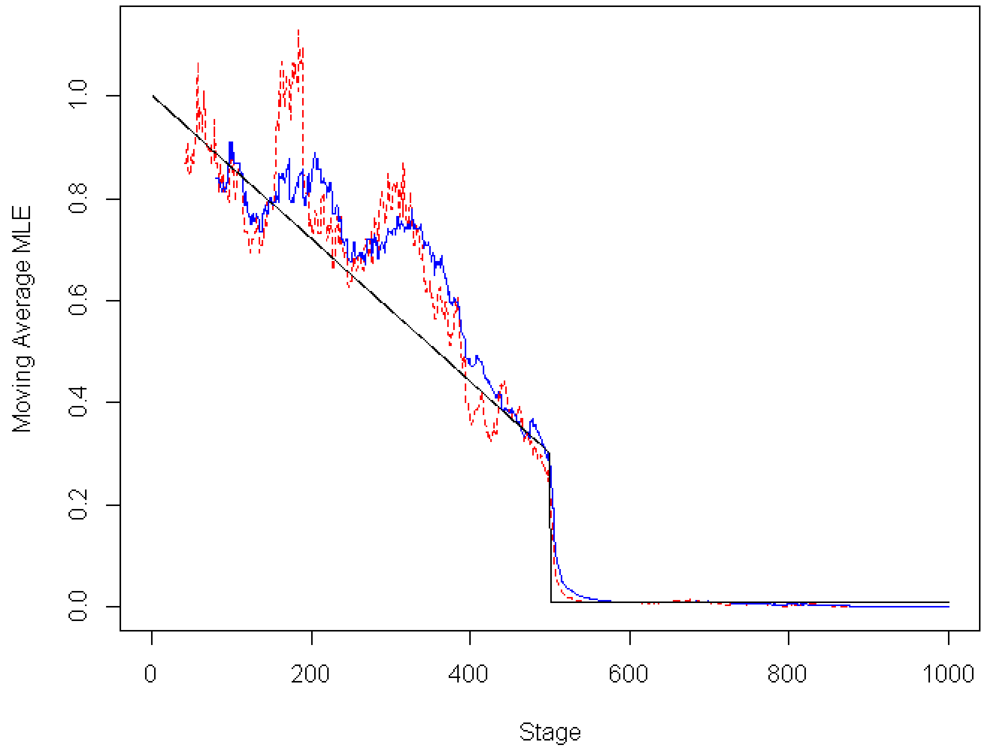

Figure 1: MAMLEs for a 1000-stage baby name simulation. Solid black line: true θ_j, dashed red line: $\hat{\theta}$ with 40-stage window width, solid blue line: $\hat{\theta}$ with 80-stage window width.

capture the cliff event very well. Both curves exhibit a predominantly positive bias due to the downward sloping nature of the theoretical curve, with the curve corresponding to the wider window showing lower bias. Once past the cliff event, both curves succeed in capturing the flatness of the theoretical curve in the later stages. It is also clear that the wider window MAMLE curve exhibits lower variability in the earlier stages.

5. Application: Trends in Popular Baby Names in the United States

This section describes the characteristics of the baby names database compiled by the Social Security Administration, and summarizes trends uncovered therein by the application of the stopping rule proposed in Subsection 3.3.

5.1 The Database of Popular Baby Names

Arguably the largest and best-maintained data source of its kind in the world, the Popular Baby Names website `http://www.ssa.gov/OACT/babynames/` of the Social Security Administration hosts 133 years of the most popular first names given by parents to male and female infants born in the United States, compiled exhaustively from social security card applications since 1880. Known exclusions from the database are those individuals who were born prior to 1937 and never applied for a card, as well as those ap-

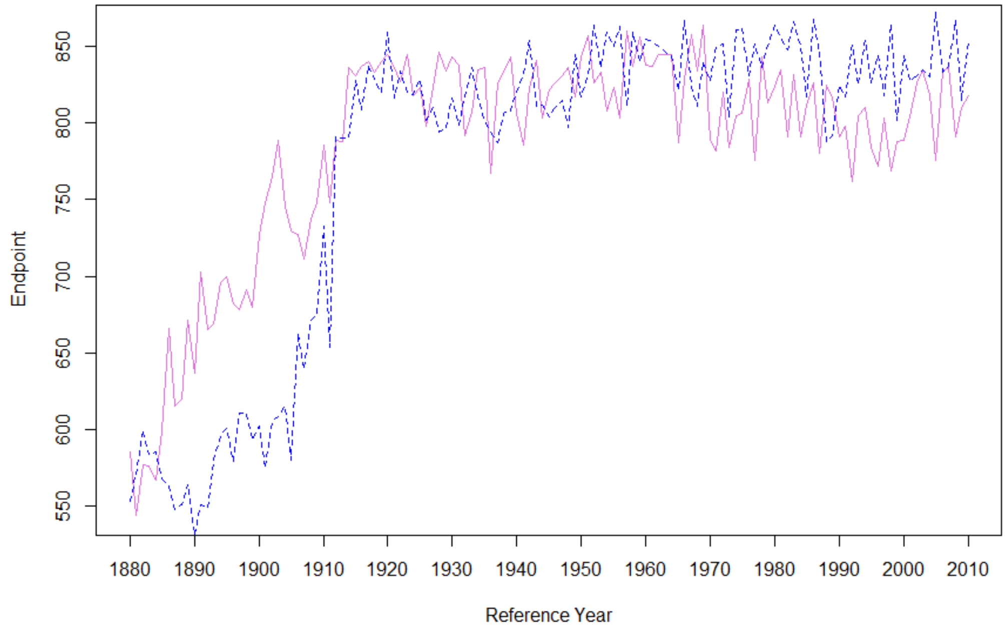

Figure 2: Endpoints for agreement in baby names by reference year. Solid pink line: endpoints for female baby names, dashed blue line: endpoints for male baby names.

plicants who omitted to provide their place of birth during the application process. The data set therefore summarizes all two-character and longer names and counts from records providing complete year of birth, sex, and state information. Name data are not modified and multiple spellings of the same name are listed as separate entries. Where counts are tied in a given year, the competing names are ranked alphabetically. In the event that the same name appears in both male and female categories for a given year (Armani, Charlie, Jaden, Jaiden, Jaidyn, Jayden, Skylar, Zion, etc., in 2011), it is ranked separately for the two sexes since the annual data is maintained by name-sex combination. The genesis of this massive compilation effort is briefly described in Shackleford (1998).

5.2 Descent into Randomness: Trends in Endpoints for the Baby Names Data

The null distribution of the name selection process, representing complete randomness in the observed year and outlined in Subsection 3.3, was simulated 2,500 times. Stage-wise MAMLEs were then computed for each simulation. The curve created by the stage-wise 95th quantiles computed from the simulated MAMLEs provided the rejection bound to which all the MAMLE curves derived from the baby names data were compared. The earliest stage K where each data MAMLE curve exceeded the 95th quantile of the simulated MAMLEs for that stage, and past which there were at most five percent of the remaining stages where the MAMLE curve exceeded the corresponding rejection bound, was denoted the endpoint at which agreement between the reference year and the observed year ceased.

Figure 2 depicts the results of this endpoint methodology. The solid pink and dashed blue lines show the trend in endpoints by reference year for female and male baby names respectively. The chart shows three distinct regimes in endpoint behavior. The dramatic increase in agreement from 1880 to 1915, as demonstrated by the steady climb in endpoints for both sexes, is noteworthy and will be investigated further. From 1916 to 1969, both sexes showed similar trends in their behaviors. Starting with 1970, male names have predominantly exhibited a higher endpoint than female names, and this evidence is supported by other analyses of these data, and will be discussed elsewhere. It is quite intriguing that new parents would choose markedly different strategies to name their newborns based simply on the sex of the child. One explanation for the higher concordance in male names may be the propagation of first names down successive generations. This promises to lead to interesting collaborations with sociologists.

6. Discussion

We have refined estimation in the multistage ranking model by using MAMLEs to reduce the variance of the maximum likelihood estimators of agreement. This is feasible only because both ranking lists are long, and has been achieved while introducing very little bias over those stages where agreement is maintained at a relatively constant level. The MAMLEs enable accurate detection of the endpoint of agreement, and we have introduced a well-defined estimator for this purpose. Additionally, we have introduced position and median penalties, which, in conjunction with Winsorized means, extend the applicability of the method to handle lists of overlapping but different items.

Simulations show that the new approach recognizes the overall shape of the parameter curve; in particular the flat regions and cliff events, and the proposed stopping rule pinpoints the stage at which the association between a pair of long ranked lists degenerates into randomness. The simplicity of the approach enables efficient computation, even when the data under study extends to a thousand stages and several thousand simulations. The application of this method to the database of baby names that are popular in the United States reveals unanticipated trends as well as differences in naming approaches for boys and girls.

As a general comment, while the focus of this paper has been on ranks generated on groups of discrete objects, the work applies equally to instances where the rankings are data reductions from continuous data, as is often the case with data from -omics platforms, and where the problem of reaching random ordering still holds interest. Another noteworthy aspect of our approach is that the number of parameters in the multistage model is the same as the number of items being ranked, so that the dimension of the parameter space is high enough not to be handled well by standard parametric methods. However, the object of interest—the stage at which the relative ordering becomes random—is a feature which is well defined for arbitrary distributions on rankings.

7. Planned Extensions

An interesting next step is to study the performance of the methodology on data generated from copulas, especially those with asymmetric tails. An evaluation of our technique on customer retention data from a financial services company is also under consideration.

The window widths in the methodology are currently user inputs and typically multiples of ten. A valuable extension would be to develop a self-tuning algorithm that determines, at every stage, the optimal window width appropriate for the data, and represented as either a fixed number or a percentage of the total number of stages available for analysis.

To the best of our knowledge, ours is the first comprehensive study of the trends in the baby names database. There are several questions to be investigated: trends in the MAMLE curves from the data analyzed for patterns in naming conventions between male and female newborns, distribution fitting of the MAMLE curves to extract deeper insights into the trends, and even a comparison of popular names at the state level with the national lists to examine how first names are dispersed throughout the United States.

The remarkable crossover between female and male endpoints over the past 133 years, with female names first showing a stronger cohesiveness in the late 19th and early 20th centuries, and subsequently relinquishing this claim to male names in recent years, bears further scrutiny. Whether the cause is data or in fact a real trend will be investigated in future work.

Acknowledgments

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.

REFERENCES

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

Hall, P., and M. G. Schimek (2012), "Moderate-Deviation-Based Inference for Random Degeneration in Paired Rank Lists," *Journal of the American Statistical Association*, 107(498), 661–672.

Shackleford, M.W. (1998), "Name Distributions in the Social Security Area, August 1997," `http://www.ssa.gov/oact/NOTES/note139/original_note139.html`.