Analysis of Agreement Between Two Long Ranked Lists

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy in the Graduate School of The Ohio State
University

By

Srinath Sampath, MS, ASA, CFA

Graduate Program in Statistics

The Ohio State University

2013

Dissertation Committee:

Joseph S. Verducci, Advisor

Douglas E. Critchlow

Douglas A. Wolfe
Abstract

An alternative approach to the problem recently posed by Hall and Schimek (2012) is proposed: determining at what point the agreement between two rankings of a long list of objects degenerates into noise. To this end the method of estimation in Fligner and Verducci (1988)’s multistage model for rankings is modified from maximum likelihood of conditional agreement over a sample of rankings to a locally smoothed estimator of stage-wise agreement. An extension to the case of overlapping but different sets of objects in the two lists is also provided. Simulations show that this approach performs very well under several conditions. The technique is next applied as a stopping rule to augment the tau-path algorithm, developed by Yu, Verducci and Blower (2011), in an analysis of associations between gene expression and compound potency in cancer data, and the detection of the endpoint of agreement. The methodology is also applied to a database of popular names for newborns in the United States and insights into trends as well as differences in naming conventions between the two sexes are uncovered.

KeyWords: partial rankings, top-$K$ rank list, multistage model, maximum likelihood estimation, stopping rule, consensus
To my family
Acknowledgments

I would like to thank my adviser Professor Joe Verducci for making the Ph.D. a stimulating and insightful journey. Researching and publishing a dissertation should never be this much fun. I am going to miss our regular conversations on a sweeping range of topics, including, most recently, the naming of estimators after ‘hybrid’ creatures: the mule, the hinny, and time permitting, the liger and the tigon. I look forward to many more years of collaboration.

Thank you, Professors Doug Critchlow and Doug Wolfe, for guidance through and beyond the candidacy examination. I enjoyed taking your classes, came out of every assignment, midterm and final a wiser statistician than before, and have adopted several of your best teaching ideas in my own communication style. Dr. Shili Lin, my utmost thanks for serving on the candidacy examination and providing context and a great many ideas.

My gratitude to the good folks, present and past, at the Nationwide Center for Advanced Customer Insights for the funding and the wonderful atmosphere that fosters both academic and corporate progress: Tom Bishop, Mike McCaslin, Chris Nicholas, Yiem Sunbhanich, Tim Welch and all the graduate research associates. NCACI is a phenomenal place to learn and grow.
Chris Holloman, I thoroughly enjoyed our weekly interactions and learned a great deal about applied statistics, statistical consulting and R from you. Thank you for your perspectives.

A great debt of gratitude goes to all the faculty and staff at the Ohio State Department of Statistics. I have interacted with just about every one of you on multiple occasions. You have shown me kindness and made me feel at home every single time. Professor Elizabeth Stasny and Lisa Van Dyke have patiently helped me navigate paperwork and protocol, and provided sage advice whenever I sought them out. I appreciate your availability and good counsel.

And a special thank you to my family for making all this worthwhile.
Vita

June 1992 ........................... B.Sc. Statistics
Loyola College, India

June 1992 ........................... Advanced Diploma in Systems
Management, NIIT, India

June 1994 ........................... M.S. Statistics
The Ohio State University
Columbus, OH

1996–2011 ......................... Nationwide Insurance
Columbus, OH

2003 ............................... Chartered Financial Analyst (CFA)
CFA Institute, Charlottesville, VA

2005 ............................... Associate, Society of Actuaries (ASA)
Society of Actuaries, Schaumburg, IL

2011–present ..................... Graduate Research Associate
Nationwide Center for Advanced Customer Insights
The Ohio State University
Columbus, OH

Publications

Research Publications


**Fields of Study**

Major Field: Statistics
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xi</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2. Background and Framework for Comparison</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Moderate Deviation-Based Inference for Random Degeneration in Paired Rank Lists</td>
<td>7</td>
</tr>
<tr>
<td>2.1.1 Problem Statement and Methodology</td>
<td>7</td>
</tr>
<tr>
<td>2.1.2 Estimating the Stopping Stage</td>
<td>8</td>
</tr>
<tr>
<td>2.1.3 Theorems</td>
<td>9</td>
</tr>
<tr>
<td>2.2 The Consolidation of Ranked Lists</td>
<td>9</td>
</tr>
<tr>
<td>2.2.1 Description of the Algorithm</td>
<td>10</td>
</tr>
<tr>
<td>2.2.2 Experiments and Conclusions</td>
<td>11</td>
</tr>
<tr>
<td>3. The Moving Average Maximum Likelihood Estimator (MAMLE)</td>
<td>12</td>
</tr>
<tr>
<td>3.1 Forward Multistage Model for Rankings</td>
<td>13</td>
</tr>
<tr>
<td>3.2 Parameter Estimation and the Stopping Rule</td>
<td>15</td>
</tr>
<tr>
<td>3.2.1 Locally Smooth Estimation of Agreement</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>viii</td>
</tr>
</tbody>
</table>
3.2.2 The Stopping Rule ........................................... 17
3.3 Descent Models .................................................... 18
3.3.1 MAMLE Behavior under Various Multistage Ranking Models 19
3.3.2 Comparing the MAMLE and Hall–Schimek Stopping Rules  under Thurstone Ranking Models ................................. 27
3.4 Discussion .......................................................... 34

4. Application to the Tau-Path Methodology .......................... 38
  4.1 Detecting the End of Agreement in Associated Subpopulations ... 39
  4.2 The Tau-Path Method ............................................. 40
  4.3 Detecting the End of Association Between Gene Expression and  Chemical Activity Over Subpopulation of Cancer Cell-Lines .... 45

5. Application to the Naming of Newborns in the United States ...... 50
  5.1 Extending The Multistage Model To Estimate Agreement ......... 51
  5.2 Assigning Penalties to Overlapping but Different Lists ............ 53
  5.3 Analysis of the Multistage Model using Simulations ............... 53
  5.4 Application: Trends in Popular Baby Names in the U.S. .......... 55
    5.4.1 The Database of Popular Baby Names ....................... 56
    5.4.2 Descent into Randomness: Trends in Endpoints for the Baby  Names Data ................................................. 56
  5.5 Discussion .......................................................... 58

6. Conclusions and Future Work ....................................... 60

Appendices 64

A. Relevant MAMLE Code Extracts .................................... 64

B. The Hall–Schimek Algorithm ....................................... 68

Bibliography ............................................................. 72
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Comparison of MAMLE and Hall–Schimek stopping rules under Thurstone descent models.</td>
<td>33</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 MAMLEs for a 200-stage simulation with a steep cliff. Top: window width 20, bottom: window width 40. Solid black line: true ( { \theta_j } ), dashed blue line: ( { \hat{\theta}_j } ), dotted red line: reverse ( { \hat{\theta}_j } ).</td>
<td>21</td>
</tr>
<tr>
<td>3.2 MAMLEs for a 400-stage simulation with a shallow cliff. Top: window width 40, bottom: window width 80. Solid black line: true ( { \theta_j } ), dashed blue line: ( { \hat{\theta}_j } ), dotted red line: reverse ( { \hat{\theta}_j } ).</td>
<td>22</td>
</tr>
<tr>
<td>3.3 MAMLEs for a 200-stage simulation with window widths 10, 20, 30 and 40.</td>
<td>24</td>
</tr>
<tr>
<td>3.4 MAMLEs for a 400-stage simulation and window widths 20, 40, 60 and 80.</td>
<td>25</td>
</tr>
<tr>
<td>3.5 MAMLEs with window width 20, and Winsorized means applied to stage-wise penalties.</td>
<td>26</td>
</tr>
<tr>
<td>3.6 MAMLE endpoint using window width 40 for Thurstone simulation with identity covariance matrix. Solid black line: MAMLE curve from Thurstone simulation, dashed red line: stage-wise 95th quantile MAMLEs under the null distribution.</td>
<td>29</td>
</tr>
<tr>
<td>3.7 MAMLE endpoint using window width 40 for Thurstone simulation with block diagonal covariance matrix. Solid black line: MAMLE curve from Thurstone simulation, dashed red line: stage-wise 95th quantile MAMLEs under the null distribution.</td>
<td>30</td>
</tr>
<tr>
<td>3.8 ( \Delta )-plot between reference ranks and observed ranks in a Thurstone simulation with identity covariance matrix, used by the Hall–Schimek stopping rule.</td>
<td>32</td>
</tr>
</tbody>
</table>
4.1 Fast Backward Conditional Search (Fast BCS) algorithm.

4.2 Full Backward Conditional Search (Full BCS) algorithm.

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

4.4 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.

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

5.2 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.
Chapter 1: Introduction

Ranking a group of objects is a fundamental activity in practically every field of inquiry. Judges—both human and machine—show their preferences among a group of objects by assigning them ranks. For example, voters of candidates running for political office may express their preferences with a rank of one for their favorite candidate, a rank of two for their second favorite candidate, and so on. Two judges who rank the same group of objects may exhibit similar or dissimilar preference patterns. The earliest papers on ranking objects to reflect one’s preferences date back to at least the late 18th century, when de Borda (1781) formally provided a contradiction of the ‘ordinary method’ used in the electoral process of the time of having each voter pick only one best candidate.

Thurstone (1927) posed that people arrive at rankings through a ‘discriminal process’ in which the internal stimulus for an object $i$ is represented as a random point $X_i$ on a continuous scale. For several stimuli these $X_i$ are assumed to have a multivariate normal distribution. According to Böckenholt (1993), Thurstone’s was one of the most influential models of individual human rankings. This area has witnessed substantial research in recent times. Gormley and Murphy (2006) embed a Plackett-Luce model in a latent space framework in their analysis of Irish election data. Biernacki and Jacques (2013) describe a probabilistic generative model for rank
data that incorporates the insertion sort algorithm, and compare it to the Mallows $\phi$ model (Mallows 1957) on several real data sets.

Typically, human rankings are assigned to a moderate number of objects or individuals, as for example the 14 candidates vying for five seats in the Meath constituency during the 2002 Irish elections discussed in Gormley and Murphy (2006). The number of assessors can also be large; there were 64,081 valid votes in the constituency that year. Our goal is markedly different: given just two long ranked lists—potentially running into the thousands or tens of thousands—how to detect the end of agreement between the two long lists and identify the point where the signal degenerates into noise? While human rankings involve a moderate number of objects, our approach considers a large number, and while the number of assessors in human rankings can be large, we consider two. An area where these tools can be applied is web search, where two search engines may rank thousands of web pages from most to least relevant in response to a user query.

Assessing partial agreement between rankings is a vital area of research with applications in practically every field of inquiry. Three resources that cover the rich variety of methods in this area are Critchlow (1980), Fligner and Verducci (1993) and Marden (1995). In many situations, ranks are the only data available from a survey or experiment. This area has also garnered considerable attention in the last decade from fields as diverse as bioinformatics and internet search.

This manuscript focuses on the question of Top-$K$ specification, which seeks to determine the stopping stage or endpoint $K$ of agreement between two assessors of objects on a long list. If the assessors independently rank the same long list of objects, the point $K$ in the list where the second assessor becomes uninformative about the
first is of considerable interest. As databases grow in size and number, practitioners are increasingly pressured to focus on that subset of the data where the two assessors show the greatest agreement, and potentially discard or underweight any information derived from the remainder of the data.

The question of Top-$K$ specification in ranking literature is a recent one and has witnessed an innovative approach by Hall and Schimek (2012). Their algorithm anchors one assessor’s ranks for a group of objects, and digitizes the second assessor’s stage-wise agreement or disagreement with the first assessor’s ranks, respectively, through a sequence of 1’s and 0’s. A type of concordance occurs when the second assessor’s rank for a given object falls within a distance $\delta$ of the first assessor’s rank for that object. $K$ is then estimated from this sequence of Bernoulli random variables using a moderate-deviation-based approach.

Several list aggregation algorithms, particularly the cross entropy Monte Carlo framework adapted by Lin and Ding (2009) to consolidate ranked data from multiple -omics platforms into one integrated list, are being incorporated by Schimek, Budinská, Kugler and Lin (2011) into their R-package TopKLists, to provide exploratory, graphical, inference, and integration tools for the study of ranked lists. The theoretical underpinnings of this methodology are discussed in detail in Schimek, Myšíčková and Budinská (2012).

Here we tackle the question of Top-$K$ specification using the forward-looking multistage ranking model framework developed by Fligner and Verducci (1988). At every stage, each assessor is assumed to select the most preferred object from the remaining objects. Hence one of the features of this model is that the stages become independent, and the probability calculations at every stage avoid the need to be conditioned
on the outcomes of prior stages, which in turn leads to mathematical tractability and closed-form solutions. Another aspect of this model is that the higher the deviation between the two assessors’ rankings for a given object, the greater the penalty assigned to the mismatch. In this manner, the model provides a nuanced approach to capturing the stage-wise disagreement between the two assessors.

Chapter 2 describes the Hall–Schimek algorithm in detail and provides a framework against which to compare our algorithm. The application of their algorithm in a rank aggregation tool is also briefly described. Chapter 3 presents the model framework of our moving average maximum likelihood estimator (MAMLE) in detail. The theoretical underpinnings of the MAMLE and the use of a stopping rule to determine the endpoint of agreement between two long lists are also provided in this chapter. Simulations that show the MAMLE’s behavior under various assumptions of window widths and list lengths are examined, as well as its performance in the presence of random shocks. Finally its performance is compared to that of the Hall–Schimek algorithm on repeated simulations from a Thurstone framework.

The tau-path algorithm (Yu et al. 2011) seeks to extract highly-associated subsets from a population by successively reordering the population from the most strongly associated data pairs to the least. Chapter 4 shows how our Top-$K$ method can be applied as a stopping rule to the tau-path algorithm to determine the exact subset size where agreement ends. The application of this stopping rule to gene-chemical data from the National Cancer Institute is also studied here.

Chapter 5 describes the application of the MAMLE to the study of trends in the naming of newborns in the United States based on an in-depth study of the most popular baby names compiled by the Social Security Administration for over
a hundred years. Separate analyses are conducted for male and female names, and similarities and differences in naming trends across the years are presented. The assignment of penalties is also adapted here to allow for the possibility that new names enter the list of most popular names every year.

Finally, Chapter 6 summarizes our conclusions and describes future research and possible extensions, both in the theory and in applications.
Chapter 2: Background and Framework for Comparison

The stopping rule proposed by Hall and Schimek (2012) considers a sequence \( \{I_j\} \) of Bernoulli variables, equal in length to the number of objects in the long list ranked by two independent assessors. The random variable \( I_j \) takes the value 1 if the ranks assigned by the two assessors to object \( j \) are no more than distance \( \delta \) from each other, and 0 otherwise. \( \delta \) is thus a stage-wise proximity measure between the two assessors. For example, if \( \delta \) is taken to be 50, the object ranked 51 by the first assessor will return a value of \( I_{50} = 1 \) if the second assessor assigns to the same object any rank in the set \( \{1, 2, \ldots, 101\} \). A list with a high level of initial agreement between the two assessors will therefore have a preponderance of 1’s in the early stages, which gradually devolves into a greater proportion of 0’s in the later stages as agreement fades and noise takes over.

The Hall–Schimek algorithm iteratively extracts rolling pilot samples of length \( \nu \) from the \( \{I_j\} \) and compares the estimated probability of success from each sample to a cutoff value \( z_{\nu} + 1/2 \), where \( z_{\nu} \equiv (C\nu^{-1}\log\nu)^{1/2} \) and \( C \) is a threshold parameter. To prevent the possibility that the algorithm stops too early due to random noise, as for example in a patch of random 0’s surrounded by a majority of 1’s, the algorithm alternately retraces and advances through the list, starting the next iteration at a stage \( r\nu \) distant from the endpoint of the current iteration. Exceeding the cutoff value
$z_\nu$ signifies the presence of an adequate proportion of 1's in the rolling pilot samples, and provides impetus to proceed further down the list. The algorithm continues in this manner until one of three end conditions is met, whereupon it provides an estimator of the endpoint between signal and noise, and terminates.

2.1 Moderate Deviation-Based Inference for Random Degeneration in Paired Rank Lists

Originally developed in Hall and Schimek (2008) and explored in detail in Hall and Schimek (2012), this research involves the application of a moderate deviation-based approach on Bernoulli random variables to consolidate two lists, and to determine the point in the sequence of rankings where the two lists are extremely long, and eventually become uninformative about each other. The approach is described as exploratory rather than precise, and has as one of its goals the reduction in computational complexity that accompanies these questions, especially in the case where partial rankings are present. The end goal to determine the point in the sequence of rankings where the signal degenerates into noise is of considerable interest in our research.

2.1.1 Problem Statement and Methodology

Two assessors rank a specific attribute of $N$ distinct objects independently of each other, from 1 to $N$, without allowing for ties. The indicator for the object ranked $j$ by the first assessor is denoted $I_j$ and assigned the value 1 if the second assessor is within a distance $\delta$ from $j$, and assigned the value 0 otherwise. $\delta$ is a tuning parameter and its value will depend on the data and the experiment being conducted. The two lists of ranks have therefore been translated into a single sequence of Bernoulli random
variables with 1 representing agreement (up to the distance $\delta$) and 0 representing disagreement between the two assessors. The Bernoulli random variables are also assumed to be independent, or $m$-dependent, in which latter case the results continue to hold with a slight change to the cutoff value $z_\nu$.

Defining $p_j = P(I_j = 1)$, the goal is to estimate the point at which the signal turns into noise; in other words, to estimate the stopping stage $j_0 \geq 2$ satisfying the conditions, $p_j \geq \frac{1}{2}$ for $1 \leq j \leq j_0 - 1$, $p_{j_0 - 1} > \frac{1}{2}$, and $p_j = \frac{1}{2}$, for $j \geq j_0$. We note that this $j_0$ corresponds to the $K$ in the expression Top-$K$ that we use everywhere in this manuscript, and what we ultimately seek is the estimator $\hat{j}_0$ of $j_0$.

### 2.1.2 Estimating the Stopping Stage

The algorithm to estimate $j_0$ requires establishing a series of test stages, $s_1, s_2, \ldots$, with stage $s_k$ terminating at a distance $J_{s_k}$ into the sequence $I_1, I_2, \ldots, I_N$. At every stage a pilot sample of size $\nu$ is captured; to the right of $J_{s_{k-1}} - r\nu$ if $k$ is odd, or to the left of $J_{s_{k-1}} + r\nu$ if $k$ is even. The alternating left and right averages are taken to help the algorithm overcome local patches of noise in the $I_j$’s without coming to a stop. The following quantities are constructed:

$$\hat{p}_j^+ = \frac{1}{\nu} \sum_{t=j}^{j+\nu-1} I_t \quad \text{and} \quad \hat{p}_j^- = \frac{1}{\nu} \sum_{t=j-\nu+1}^{j} I_t.$$  

The hypothesis test to be conducted is $H_0 : p_k = \frac{1}{2}$ for $\nu$ consecutive values of $k$, versus $H_1 : p_k > \frac{1}{2}$ for at least one value of $k$. Setting $z_\nu \equiv (C\nu^{-1}\log \nu)^{1/2}$ as a moderate-deviation bound, $H_0$ is rejected if and only if either $\hat{p}_j^+ - \frac{1}{2} > z_\nu$ or $\hat{p}_j^- - \frac{1}{2} > z_\nu$.

The algorithm is terminated when either

1. The algorithm enters an infinite loop between adjacent stages $s_{2k-1}$ and $s_{2k}$, or
2. For some $k$, $J_{s_{2k+1}} \leq J_{s_{2k-1}}$, or

3. $J_{s_{2k}} - r\nu \leq 1$, occurs,

and the estimator of $j_0$ is calculated to be $\hat{j}_0 = J_{s_{2k-1}} + \frac{1}{2}\nu$. If the algorithm does not terminate, then $\hat{j}_0 := N$.

The paper also discusses adjustments that need to be made for incomplete rankings and for the possibility where one assessor ranks a few objects so highly that they throw off all subsequent rankings, leading to a very early stop to the algorithm. Furthermore, several technical parameters and the one tuning parameter $\nu$ need to be set before executing the algorithm, based on the nature of the data being analyzed.

2.1.3 Theorems

The paper also states three theorems, of which only the first two are quoted here.

**Theorem 1.** For each integer $k$, $J_{s_{2k-1}} + \nu - 1 \leq J_{s_{2k}}$. The values of $J_{s_{2k-1}}$, for $k \geq 1$, form a strictly monotone increasing sequence, until the algorithm terminates.

That is, for each $k \geq 1$, if the algorithm has not terminated by the end of stage $s_{2k-1}$, then $J_{s_{2k+1}} > J_{s_{2k-1}}$.

**Theorem 2.** Assume that the data $I_1, \ldots, I_N$ are generated by our mathematical model; that $C$, in the definition of $z_\nu$ satisfies $C > \frac{1}{2}$; that $\nu = \nu(n) = o(n)$; and that, for constants $0 < B_1 < 1$ and $B_2 > 0$, $\nu \geq B_2 n^{B_1}$. Then, if $B$ is so large that $2B_1(B - C^{1/2})^2 > 1$, and with $j_1 = j_1(B, \nu)$, we have $|\hat{j}_0 - j_0| = O_P\{\min(j_0 - j_1, \nu)\}$.

2.2 The Consolidation of Ranked Lists

Schimek, Myšičková and Budinská (2010) extend the results of Hall and Schimek (2008) to $\ell > 2$ lists. The end goal is to create a smaller set $\tau^*$ of ranked objects
from the input lists by first using a moderate deviation based approach and then a cross entropy Monte Carlo approach. This approach allows for idiosyncratic and incomplete rankings, and handles input lists that may contain different objects. \( k^* \) is the position up to which the rankings have high conformity. As with the earlier paper, the emphasis is on mathematical simplicity and a reduction of computational complexity.

2.2.1 Description of the Algorithm

The algorithm consists of two steps.

- Step 1 is the calculation of the lengths \( \hat{k}_l \) of the truncated lists \( O'_l \) created from all the \( \binom{\ell}{2} \) pairwise combinations of the \( \ell \) input lists \( O_l \), using the Hall and Schimek (2008) algorithm. \( \hat{k}_l \) represents the point in the sequence at which the signal from that pair of input lists degenerates into noise. The overall Top \( k \) list length, across all input lists, is then calculated to be \( \hat{k}^* = \max_l(\hat{k}_l) \).

- Step 2 is the consolidation of the various lists using an optimization criterion that is dependent on the distance measure used (whether Kendall’s Tau or Spearman’s Footrule or another measure). More precisely, the goal of Step 2 is to determine the list \( \tau^* \) such that

\[
\tau^* = \arg \min_{\tau} \left\{ \sum_{l=1}^{\ell} w_l d(\tau, \tau_l), \tau \in O' \right\},
\]

where the weights \( w_l \) can be used to favor some input lists over others, \( d(\cdot, \cdot) \) is the chosen distance measure, and \( O' = \bigcup O'_l \). Since one goal of this paper is to reduce computational complexity, the authors adopt the cross entropy Markov Chain proposed by Lin and Ding (2009), which sidesteps the need to process all possible permutations.
2.2.2 Experiments and Conclusions

The algorithm was tested on simulated microarray data at various settings of the tuning parameters. Kendall’s Tau and Spearman’s Footrule were both adopted as distance measures. For both measures, the results faithfully recreated the separation between the true top-ranked objects and the remaining objects on the lists. An evaluation of the performance of the algorithms on colorectal cancer microarray data also resulted in full agreement with current biological knowledge of the disease. The authors hope to apply the methodology in a consolidation of web search engine results.

The authors are further satisfied that a) the estimate for $k^*$ is provided by the data rather than pre-specified, which makes the algorithm more flexible and data-driven, and b) the past preference for needing permutations is relaxed due to this work since the distance measures have been extended to complete the partially ranked lists, as motivated by Fagin, Kumar and Sivakumar (2003).

Schimek (2011) also integrates the moderate deviation-based approach with the Lin and Ding (2009) CEMC algorithm, using Kendall’s tau measure as the optimization criterion. Gene expression data from the real-time quantitative reverse transcription polymerase chain reaction (RT-qPCR) technology were used. The moderate deviation-based approach, combined with the CEMC algorithm, performed on par with the familiar de Borda (1781) and Markov Chain techniques.
Chapter 3: The Moving Average Maximum Likelihood Estimator (MAMLE)

In this chapter we lay out the mathematical framework for the moving average maximum likelihood estimator (MAMLE), which is the focus of our research, and examine its performance under various simulations. A comparison to the Hall–Schimek algorithm is also conducted.

Section 3.1 describes the model framework and the role of the truncated geometric distribution in the assignment of probabilities to the stage-wise disagreements between the two assessors’ ranks. The determination of the MAMLE and the explicit stopping rule are given in Section 3.2. This stopping rule applies to a setting that is distinct from the customary changepoint scenarios. http://www.changepoint.info/publications.html provides a comprehensive list of publications in this area.

Section 3.3 examines the behavior of the MAMLEs in multiple settings, and compares the new stopping rule with that of Hall and Schimek (2012). First, in Subsection 3.3.1, MAMLE behavior for different window widths and list lengths is examined in the presence of steep and shallow cliff events in assessor agreement. The impact of random shocks on the descent form is also studied here. In Subsection 3.3.2, the two stopping rules are compared on repeated simulations from Thurstone models (Marden 1995), allowing us to examine some effects of nonindependent stages in a
system with a well-defined endpoint of agreement. Lastly, in Section 3.4, the strengths and weaknesses of the two stopping rules are discussed.

3.1 Forward Multistage Model for Rankings

A ranking or permutation of $n$ distinct objects is a vector of length $n$, with each component corresponding to an object, and the value of each component being the rank of that object, namely the quantity $(1 + \text{the number of other objects that are considered superior})$, where superiority may be determined in either a qualitative or quantitative sense. We use

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

to denote this ranking or permutation.

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

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

We now consider the situation where the $n$ objects are ordered sequentially according to two independent processes. The processes may be the qualitative ranking schemes of two judges who evaluate a shortlist of books that are finalists in a prestigious competition. Another example is two technologies that independently measure the severity of a disease in a group of patients, and rank the patients based on their risk level. It is possible that the two ranking processes are initially governed by common parameters and, as one moves further down the list, the ranks start to diverge from each other, and ultimately, at point $K$, become completely uninformative about
each other. We are interested in determining the value of \( K \) where the two ranking schemes become uninformative about each other.

Our approach is to fix one of the two assessors’ rankings as the reference ranking or ground truth, and evaluate the deviation of the other assessor’s ranks—the generated or observed ranking \( \pi \)—from the reference ranking, using the Fligner and Verducci (1988) multistage ranking approach as follows. For illustrative purposes, we describe the mathematics of the first stage and then generalize to the other stages, which are all assumed to be independent.

Stage 1: Here all \( n \) objects are available. The second assessor selects the \((1 + v)\)th best object overall, as specified by \( \pi^{-1} \), and incurs the penalty \( V_1 = v \), with truncated geometric probability

\[
P(V_1 = v) = \left( \frac{1 - r_1}{1 - r_1^n} \right) r_1^v, \quad v = 0, \ldots, n - 1, \quad 0 < r_1 < 1.
\] (3.1)

Stage \( j \) \((j = 2, \ldots, n - 1)\): In stage \( j \), \( n - j + 1 \) objects are available. The second assessor selects the \((1 + v)\)th best available object, as specified by \( \pi^{-1} \), and incurs a penalty \( V_j = v \), with truncated geometric probability

\[
P(V_j = v) = \left( \frac{1 - r_j}{1 - r_j^{n-j+1}} \right) r_j^v, \quad v = 0, \ldots, n - j, \quad 0 < r_j < 1.
\] (3.2)

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 is equivalent to the removal of all skill from the second assessor with respect to the reference ranking. In other words, the second assessor is assumed to select an object from the remaining objects in a random manner. Thus \( r_j \) represents a form of disagreement at stage \( j \) between the two assessors, with deterministic, perfect agreement attained as \( r_j \to 0 \).
We assume independent choices at each stage of the ranking process, and so the \( \{V_j \mid j = 1, \ldots, n-1\} \) are independent. \( \{V_1, \ldots, V_{n-1}\} \) is the \textit{discordance} or \textit{penalty vector} between the reference ranking and the generated ranking. Since the probabilities in (3.1) and (3.2) are decreasing functions of \( v \), the model assigns lower probabilities to larger departures from the reference ranking at every stage.

For mathematical convenience, we apply the transformation

\[
\theta_j = -\log r_j, \ j = 1, \ldots, n-1,
\]  

and follow the behavior of the \( \{\theta_j\} \). The condition \( r_j \to 1 \), which leads to the limiting uniform distribution for the \( V_j \)'s, is equivalent to the condition \( \theta_j \to 0 \).

The original problem statement of determining the value of \( K \) where the two ranking schemes become uninformative about each other is equivalent to determining the value of \( K \) for which \( \theta_K > 0 \), and \( \theta_j = 0 \) for all \( j > K \). Informally, we seek the final stage where the second assessor exhibits some level of agreement with the reference ranking, before devolving into uniformly random selections of the remaining unranked objects. Once past this stage \( K \), the two assessors become uninformative about each other, as suggested by the empirical work in Section 3.3.

### 3.2 Parameter Estimation and the Stopping Rule

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

Recall that the \( r_j, j = 1, \ldots, n - 1 \), are stage-wise measures of the second assessor’s agreement with the reference ranking. 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 assessors. We now determine the MAMLE of each \( r_j \), making the reasonable assumption that the \( r_j \)'s vary gradually over most of the range, but not over the changepoint to noise. Specifically, to determine the MAMLE \( \hat{r}_j \) for a given stage \( j \), we assume a common value \( r \) for all the unknown \( r_i \) in the backward-looking window \( j - w + 1 \leq i \leq j \) of width \( w \). In this manner we calculate a set of MAMLEs, \( \hat{r}_j \), and note that these MAMLEs use overlapping rank data as the window moves forward through the stages.

### 3.2.1 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

\[
L(r) = P(V_{j-w+1} = v_{j-w+1}) \times \cdots \times P(V_j = v_j) = \left(\frac{1 - r}{1 - r^{n-(j-w)+1}}\right)^{v_{j-w+1}} \times \cdots \times \left(\frac{1 - r}{1 - r^{n-j+1}}\right)^{v_{j}}.
\]

The local log likelihood is therefore

\[
\log L(r) = w \log(1 - r) + (\log r) \sum_{i=j-w+1}^{j} v_i - \sum_{k=n-j+1}^{n-j+w} \log(1 - r^k). \quad (3.4)
\]
Differentiating (3.4) with respect to \( r \) gives the MAMLE \( \hat{r}_j \) as the solution to the equation 
\[
\overline{V}_j = g_j(r_j),
\]
where
\[
\overline{V}_j = \frac{1}{w} \sum_{i=j-w+1}^{j} V_i,
\] (3.5)
is the mean penalty for the window, and
\[
g_j(r) = \frac{r}{1-r} - \frac{1}{w} \sum_{k=n-j+1}^{n-j+w} \frac{kt^k}{1-t^k}
\]
is an increasing function of \( r \). To the extent that \( \{r_{j-w+1}, \ldots, r_j\} \) are actually increasing and not constant, the estimator \( \hat{r}_j \) will be negatively biased.

### 3.2.2 The Stopping Rule

The calculations in Subsection 3.2.1 result in MAMLEs for the stage-wise \( r_j \)'s, which in turn can be used to compute the stage-wise MAMLEs \( \{\hat{\theta}_j\} \) by applying the transformation (3.3). To determine the rejection bound of the latter MAMLEs, we generate a large number of simulations from the multistage model assuming that all the \( \theta_j \)'s are 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

\[
\hat{K} = \text{the earliest stage at which } \hat{\theta}_{\hat{K}+w} > q(\hat{K}), \text{ and } \hat{\theta}_j > q(j) \text{ for at most } \alpha \text{ percent of the remaining } j > \hat{K} + w. \quad (3.6)
\]

The reason for moving the region of \( \theta_j \)'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.

### 3.3 Descent Models

A detailed discussion of the MAMLE approach including an examination of the various properties of the estimator, and its performance in several simulated settings are presented in Subsection 3.3.1. The simulations replicate typical forms of agreement between two assessors, and also allow for the possibility of non-linear descents, where new entrants into the list of objects ranked by the second assessor cause instantaneous shocks to the linear descent form of the true $\theta$ curve. Subsection 3.3.2 then compares the two stopping rules on simulations generated under the framework of Thurstone ranking models, which provides a setup of nonindependent stages where the endpoint of agreement is known. As the application that runs the Hall
and Schimek (2012) methodology was under development at the time of this submission, we coded the algorithm ourselves to facilitate comparison of the stopping rules. Appendix B provides our code of the Hall–Schimek algorithm.

3.3.1 MAMLE Behavior under Various Multistage Ranking Models

In this subsection, the behavior of the MAMLEs is studied under different window widths, list lengths, cliff events, and shocks. Beginning with $n = 200$ objects, the true agreement between the two assessors is assumed to decline gradually and linearly for about one-half of the stages, whereupon there is either a steep or shallow cliff event caused by an instantaneous drop in agreement between the assessors. The stages following the cliff event assume that the observed ranking is completely random. For our purposes, identifying the shape of the descent curve and accurately picking the location of the cliff event are important.

The solid black line in the two exhibits of Figure 3.1 represents the true $\{\theta_j\}$ used in the simulation. The linear descent over the first 100 stages is meant to represent the decreasing stage-wise agreement between the two assessors. The cliff event at stage 101 signifies the abrupt disagreement between the second assessor and the preference ranking, after which point the second assessor ranks the remaining objects at random, which action is represented by the horizontal line. The two graphs show the computed MAMLEs with window widths of 20 and 40 stages respectively. While the dashed blue line represents the computed $\{\hat{\theta}_j\}$, the dotted red line represents the ‘reverse’ $\{\hat{\theta}_j\}$, where the MAMLEs are calculated exactly as before, but the roles of the first assessor and the second are reversed, with the second assessor supplying the reference ranking and the first assessor supplying the generated ranking.
It is clear that although the reversal of roles between the two assessors leads to a slightly different picture of the degradation of information, and that an average of the two estimators could be used for Top-$K$ detection, in practice this is not necessary. It is also evident that both window width assumptions capture the cliff event very well, with the narrower-window MAMLEs providing a more accurate picture of the cliff event than the wider window MAMLEs. Furthermore, all four MAMLE curves show a positive bias, which is attributable to the descent form. If the initial stages had the same constant value of $\theta$, this bias would go away, and does in fact become negative if the initial stages $\{\theta_j\}$ are increasing. In all cases, the effect on the stopping rule is small. Once past the cliff event, all four MAMLEs succeed in capturing the flatness of $\{\theta_j\}$ in the later stages, with the wider-window MAMLEs showing better results. It is also clear that the wider window MAMLEs show lower variability in the earlier stages.

Figure 3.2 gives the results of a simulation involving two assessor rankings and 400 stages. The top graph calculates the MAMLEs using rolling 40-stage windows, while the bottom graph uses 80-stage window widths. Here again the $\{\hat{\theta}_j\}$ and the reverse $\{\hat{\theta}_j\}$ are very similar, and all MAMLEs identify the more shallow cliff event at stage 182. The MAMLEs again show a positive bias which decreases close to the cliff event. As with the 200-stage runs, all four MAMLEs are successful in capturing the flatness of $\{\theta_j\}$ in the later stages. As expected, the wider-window MAMLEs show lower variability in the earlier stages.

Figures 3.3 and 3.4 show, respectively, the form of the MAMLE curve computed for windows of widths 10, 20, 30 and 40 stages for a list of 200 objects, and the
Figure 3.1: MAMLEs for a 200-stage simulation with a steep cliff. Top: window width 20, bottom: window width 40. Solid black line: true $\{\theta_j\}$, dashed blue line: $\{\hat{\theta}_j\}$, dotted red line: reverse $\{\hat{\theta}_j\}$. 
Figure 3.2: MAMLEs for a 400-stage simulation with a shallow cliff. Top: window width 40, bottom: window width 80. Solid black line: true \( \{\theta_j\} \), dashed blue line: \( \{\hat{\theta}_j\} \), dotted red line: reverse \( \{\hat{\theta}_j\} \).
MAMLE curve computed for windows of widths 20, 40, 60 and 80 stages for a list of 400 objects. The following are evident for the linear descent form:

1. The presence of the positive bias of $\{\hat{\theta} \}$ that increases as the window width increases,

2. Volatility decreases as the window width increases,

3. The shallow cliff event is best identified by the MAMLE created with the smallest window width, and the MAMLEs created with larger window widths miss the cliff event by increasing margins, and

4. The flat region is best picked up by the MAMLE created using the smallest window width, and the MAMLEs created using larger window widths miss the flat region by increasing margins.

It is clear from the above analysis that window width plays a critical role in the MAMLE methodology. At this juncture in our research, window widths are chosen to strike a balance between providing a clear view of the descent, and supplying an adequately smooth estimator of the stage-wise agreement between the assessors. Work on determining self-adaptive window widths to reflect the distinctive characteristics and length of the agreement process is underway.

Finally, the impact of random shocks on the assumed $\theta$ curve is discussed. It is not unusual to see new objects being introduced at various points in an existing list. The list of most popular infant names in a country could see the addition of entirely new names from characters in a hit film or television program. The list of bestselling books could witness new titles from prizewinning authors.
Figure 3.3: MAMLEs for a 200-stage simulation with window widths 10, 20, 30 and 40.
Figure 3.4: MAMLEs for a 400-stage simulation and window widths 20, 40, 60 and 80.
Figure 3.5: MAMLEs with window width 20, and Winsorized means applied to stage-wise penalties.
Figure 3.5 shows a linear descent function that experiences shocks at two stages. New objects have evidently appeared on the list at stages 50 and 75, resulting in a steep instantaneous decline in the true $\theta$ curve at these two stages. The MAMLEs were calculated with the mean penalty formula given in equation (3.5), as well as with Winsorized means that trimmed the outlier penalties for each window to equal the most extreme remaining penalties. The effects of various trim levels—10 percent, 20 percent, and the reciprocal of the window width (here, 20 stages)—on the MAMLE curve are shown. The characteristic positive bias of the MAMLE curves caused by the linear descent is removed by the two shocks. Importantly, in all four cases, the MAMLE curves identify the final cliff event at stage 100.

Overall, in its own multistage ranking model setting, the MAMLE behaves well in detecting both large and small cliff events, and is resistant to occasional shocks even without trimming.

### 3.3.2 Comparing the MAMLE and Hall–Schimek Stopping Rules under Thurstone Ranking Models

In this subsection, the stopping rules are compared in two settings where the rankings are derived from multivariate normal distributions, as proposed by Thurstone (1927). In the first setting, the underlying normal variables are independently generated; in the second setting, a strong covariance structure is assumed. Note that, even in the independent setting, the Thurstone model is quite distinct from the multistage model. In particular, the multistage model assumes that the relative ranking of available objects is independent from previous top choices. In the Thurstone model, the relative ranking of available objects is determined by the conditional distribution of order statistics from independent normals; this distribution depends strongly on the
previous top choices. Thus the Thurstone models provide a good setting to check the behavior of the stopping rules when the assumption of independent stages is violated, but the transition from informative to noninformative (remaining objects have the same mean and variance) stages remains clear.

With the first assessor’s ranks fixed, ranks are simulated for the second assessor according to the Thurstone model. Specifically, 100 simulations are generated from each of two multivariate normal distributions of dimension 400 as follows: both distributions have the same mean of form

$$\mu_{400 \times 1} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

where $\mu_1$ and $\mu_2$ are column vectors of dimension 200, with the former descending linearly from 6 to 3 over its 200 elements, and the latter being a zero vector. The covariance matrices of the two multivariate normal distributions are assumed to be $\Sigma_0 = I_{400 \times 400}$, and $\Sigma_1$ having the block diagonal form

$$\Sigma_1 = \begin{pmatrix} \Sigma^* & 0 & 0 \\ 0 & \Sigma^* & 0 \\ 0 & 0 & I \end{pmatrix}$$

where $\Sigma^* = ((\sigma_{ij})) = 1$ if $i = j$, and $\rho$ if $i \neq j$, is a 100 $\times$ 100 symmetric matrix, and $I$ is a 200 $\times$ 200 identity matrix.

The simulations generated from these two multivariate normal distributions are ranked to provide two versions of the second assessor’s list. It is evident that there is a true endpoint at stage 200, and also that there is dependence across the stages in both scenarios.

To help understand the performance of the MAMLE, we illustrate the results from a single data simulation and use window width 40 and $\alpha$ equal to 0.05. The solid black line in Figure 3.6 depicts the MAMLE curve computed for a single simulation of
Figure 3.6: MAMLE endpoint using window width 40 for Thurstone simulation with identity covariance matrix. Solid black line: MAMLE curve from Thurstone simulation, dashed red line: stage-wise 95th quantile MAMLEs under the null distribution. The dashed red line gives the stage-wise 95th quantile MAMLEs under the null distribution. The stopping rule described in Subsection 3.2.2 selects stage 188 as the endpoint for this data simulation. Similarly, the solid black line in Figure 3.7 depicts the MAMLE curve computed for a single simulation of the multivariate normal \((\mu, \Sigma_1)\) distribution. As before, the dashed red line gives the stage-wise 95th quantile MAMLEs under the null distribution. For this simulation, the stopping rule selects stage 189 as the endpoint.
Figure 3.7: MAMLE endpoint using window width 40 for Thurstone simulation with block diagonal covariance matrix. Solid black line: MAMLE curve from Thurstone simulation, dashed red line: stage-wise 95th quantile MAMLEs under the null distribution.
Next, the Hall–Schimek algorithm is illustrated for a single simulation of the multivariate normal \((\mu, \Sigma_0)\) distribution. Recall that their method first reduces the data to statistics \(I_j\) where

\[
I_j = \begin{cases} 
1, & \text{if the difference between assessor ranks for object } j \leq \delta \\
0, & \text{otherwise.}
\end{cases}
\]  

(3.7)

Figure 3.8 shows the \(\Delta\)-plot, representing the total number of 0’s in the sequence \(\{I_j\}\) calculated for different values of the distance \(\delta\). This graph is used as a tool to select a reasonable value of \(\delta\) in the determination of the endpoint.

Additional parameters for the Hall–Schimek algorithm are

\[
\begin{align*}
\nu &= \text{length } |\{I_j\}| \text{ of rolling sequences,} \\
r &= \text{multiplier of } \nu \text{ used to counteract early stopping, and} \\
C &= \text{threshold parameter based on the moderate-deviation bound.}
\end{align*}
\]

For typical values of \((\nu, r, C)\), as for example, \((50, 1.2, 0.26)\), and a value of \(\delta\) that falls just beyond a steep decline in the \(\Delta\)-plot, say, \(\delta = 38\), the endpoint from the moderate-deviation-based stopping rule is 60. However, setting both \(\delta\) and \(\nu\) equal to 60 gives a very reasonable endpoint of 191. The challenge here is to select a reasonable set of values for the parameters in order to achieve a stable endpoint.

Setting \(\delta\) equal to 50 ensures that an object ranked by the first assessor in the 51st place will contribute a 1 to the sequence \(\{I_n\}\) if the second assessor were to assign it any rank from 1 to 101, which leads to high levels of agreement in the midrange ranks.

Exactly 100 simulations from each of the two Thurstone ranking models, and 1,000 simulations from the null distribution were generated. For the latter Thurstone ranking model, \(\rho\) was set at 0.7. Endpoints based on both stopping rules were computed
Figure 3.8: Δ-plot between reference ranks and observed ranks in a Thurstone simulation with identity covariance matrix, used by the Hall–Schimek stopping rule.
### Table 3.1: Comparison of MAMLE and Hall–Schimek stopping rules under Thurstone descent models.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Unadjusted</th>
<th>Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Covariance matrix</td>
<td>Identity</td>
<td>Block diagonal</td>
</tr>
<tr>
<td>MAMLE endpoint analysis ($\alpha = 0.05$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Window width = 20</td>
<td>Initial window width = 20</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>212</td>
<td>219</td>
</tr>
<tr>
<td>Median</td>
<td>191</td>
<td>192</td>
</tr>
<tr>
<td>MAD</td>
<td>5.93</td>
<td>7.41</td>
</tr>
<tr>
<td>Window width = 40</td>
<td>Initial window width = 40</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>209</td>
<td>204</td>
</tr>
<tr>
<td>Median</td>
<td>187</td>
<td>187</td>
</tr>
<tr>
<td>MAD</td>
<td>6.67</td>
<td>5.93</td>
</tr>
<tr>
<td>Hall–Schimek endpoint analysis ($r = 1.2, C = 0.26$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\nu = 20, \delta = 50$</td>
<td>$\nu = 20, Initial \delta = 50$</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>39</td>
<td>131</td>
</tr>
<tr>
<td>Median</td>
<td>33</td>
<td>196</td>
</tr>
<tr>
<td>MAD</td>
<td>30.39</td>
<td>8.90</td>
</tr>
<tr>
<td>$\nu = 40, \delta = 50$</td>
<td>$\nu = 40, Initial \delta = 50$</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>81</td>
<td>143</td>
</tr>
<tr>
<td>Median</td>
<td>74</td>
<td>195</td>
</tr>
<tr>
<td>MAD</td>
<td>34.10</td>
<td>8.90</td>
</tr>
</tbody>
</table>
for each simulation. The results are summarized in Table 3.1. The section on the left shows unadjusted values. In the MAMLE case, this means that the diagnostics provided in Subsection 3.2.2 are ignored. In the Hall–Schimek case, this means that diagnostics such as the Δ-plot are ignored. The section on the right shows adjusted values, where all diagnostics for both stopping rules are observed. Adjusting the endpoints has the effect of additional smoothing of the MAMLE curve by increasing the window width in certain simulations. For the Hall–Schimek algorithm, the adjustment increases δ past the point of steepest decline in the Δ-plot, allowing for a looser form of agreement between the two assessors; this seems to work best with the block diagonal model.

For all cases in Table 3.1 the MAMLE stopping rule provides a more stable estimator of the endpoint than the Hall–Schimek stopping rule, as measured by the median absolute deviation from the median (MAD). However the Hall–Schimek estimator may be modestly less median-biased. Beyond the table, we have informally conducted a post hoc search of the parameter space for points where the Hall–Schimek estimator was unbiased, but the search usually led to large settings of δ and ν and the resultant variability tended to be high. Most critically, the table shows that different parameter settings for the Hall–Schimek estimator may lead to very different conclusions.

3.4 Discussion

The MAMLE methodology discussed above was developed in the free and collaborative programming environment of R, version 3.0.1 (R Core Team 2013) on the 64-bit MS Windows 7 Professional operating system of a Hewlett-Packard EliteBook
Folio 9470m. The laptop used a dual-core Intel® Core™ i5-3437U CPU, with a turbo clock speed of 2.40 GHz. Installed memory was 8 GB RAM. The computation time to compare two lists, each of length about 19,400 objects, by calculating the MAMLE curve using a window width of 40 stages, was 2.5 minutes. Computation time naturally extends with an increase in window width, but it is clear that the MAMLE algorithm can support long lists, even those that are 4-5 times the length of the test data. A key observation about this reasonable computation time is that the analysis is performed in a single pass-through of the data and no large matrices are involved.

We have proposed an approach to determine the endpoint $K$ of agreement between the ranks of two assessors who evaluate a list of objects. Our approach exploits the forward-looking multistage ranking model framework developed by Fligner and Verducci (1988), and uses a graded assignment of penalties to the mismatched ranks provided by the assessors at each stage. Simulations show that our approach is very successful in recognizing the overall shape of the parameter curve, especially cliff events and flat regions. The underlying mathematics is tractable and the algorithm is quick to code and execute, providing a new tool to track the agreement between the preferences of two assessors along a list of objects, and a technique to estimate the point of degeneration of agreement between the rankings. In the multistage model, the flexibility of staging allows the order of stages to be determined by special considerations.

We conclude this chapter with a comparison of several characteristics of the two stopping rules. It is important to note here that while we are following the published usage for $p$ in Hall and Schimek (2012), it would be more reasonable to treat $p$ as
a function of the proximity measure \( \delta \); however it is not clear that the moderate-deviation bound has been derived for this case.

- **Data Compression and Information**: The Hall–Schimek method reduces to a 1 or 0, respectively, the agreement or disagreement between the two assessor rankings of each object on the long list. In this sense information about the proximity of the assessor valuations of each object is lost, as the outcome at each stage is reduced to either a success or a failure. The MAMLE approach extracts more information at each stage by recording every possible level of disagreement between the assessor ranks through a nuanced penalty function. Mathematical tractability permitting, both the penalty function and the form of its probability distribution can be modified, leading to a rich family of models of agreement.

- **Estimation and Choice of Parameters**: Our experience with the Hall–Schimek algorithm has been that its estimation of the endpoint benefits from large choices of \( \delta \) and even \( \nu \). Parameter selection relies heavily on preliminary data analysis, consistent with the characterization of the methodology by the authors as exploratory. We have observed that improper selections of \( \delta \) and \( \nu \) can lead to substantially different endpoints. The critical parameter in the MAMLE technique is the window width \( w \). The form and variability of the plot of \( \{\hat{\theta}_j\} \) and the diagnostics provided in Subsection 3.2.2 are good starting points in the selection of the appropriate \( w \) for the data under consideration. We are also in the process of formalizing and automating the approach to determine the optimal window width.
• *Independence versus m-Dependence:* The Hall–Schimek method is designed to be robust to dependencies. The MAMLE estimators are naturally robust to dependencies, at least in the context of ranking models such as Thurstone, as evidenced by the endpoint estimates and moderate MADs in Table 3.1.

• *Random Shocks:* The Hall–Schimek approach handles the possibility of new objects being present exclusively on the second assessor’s list by introducing a new parameter $\phi$, the number of objects that may be discarded at each test stage prior to the calculation of the success probability. The refined assignment of penalties to objects at each stage of the MAMLE algorithm makes it resistant to random spikes, even without the benefit of trimming, as shown in Subsection 3.3.1.

• *Diagnostics:* The Hall–Schimek method benefits from an analysis of the $\Delta$-plot, which influences the selection of the important parameter $\delta$. The plot of $\{\hat{\theta}_j\}$ created by the MAMLE algorithm not only plays a vital role in the selection of the endpoint of association, but also provides valuable insights into the stage-wise form of agreement between the two assessors, thus lending itself to further analysis.
Chapter 4: Application to the Tau-Path Methodology

The tau-path algorithm, developed by Yu et al. (2011), is a nonparametric approach that considers the relationship between two variables, which may be strong in a subset of the population, but masked by a modest correlation over the entire population. Using the concordance matrix and successive reorderings of the observations, the tau-path method provides a nested sequence of subsets that group the observations in order from strongest association to weakest. This rearrangement provides a natural ordering for staging in the multistage model. Application of our Top-$K$ method over the tau-path ordered stages leads to a model with decreasing agreement and a precise choice of subset size as illustrated on genomic data. Sections 4.1 and 4.2 present an overview of the tau-path methodology and in-depth look into the tau-path model framework respectively. Section 4.3 describes the use of the tau-path algorithm in conjunction with our stopping rule to identify potential interactions between gene expression and drug potency over the 60 cancer cell-lines of the popular NCI-60 data warehouse maintained by the National Cancer Institute (NCI). We simulate the null distribution by assuming that the second assessor provides ranks that are completely random relative to the first assessor’s ranks, and compare ordered gene-compound pairs with this null distribution to estimate the endpoint for each pair.
4.1 Detecting the End of Agreement in Associated Subpopulations

Situations arise in natural phenomena and experiments where a strong association exists between two variables for a subset of the population, but the strength of the association weakens or even dissipates when the full population is considered. Of interest to the researcher is that subset where the relationship is the strongest, as it may provide valuable clues about the characteristics of the underlying phenomenon. The subset can often be isolated and analyzed further. Yu et al. (2011) list several reports that analyze data from the National Cancer Institute (NCI) to provide insights into the response of the human body to chemotherapy. The substantial NCI-60 dataset, available at dtp.nci.nih.gov/docs/cancer/cancer_data.html, has two components that are frequently used in the study of human chemoresistance: gene expression patterns for a panel of 60 untreated human cancer cell-lines, and drug potency values from the cancer cell-lines for over 100,000 inhibitor drugs, which were applied independently and individually. It is worth noting that both gene expression data and treatment potency levels are univariate, and were most probably compiled during different time periods. By analyzing the association between these two variables using myriad data mining techniques, researchers are in a position to pinpoint relationships that can be tested further, and possibly in a cost-effective manner. If it were possible to mathematically isolate those subsets of cell-lines where gene expression and compound activity are most associated, through an initial screening process, these subsets could be investigated for deeper insights into the nature of the disease.
Furthermore, in these times of massive data sets, extracting strongly associated sub-
sets from the population may lead to more astute findings, accelerate the discovery
process, and provide opportunities for further experimentation.

The tau-path method described next extracts this ordered, strongly associated
subset from the population, and the Top-$K$ specification is adapted to a stopping
rule that acts on this ordered subset to determine the length of association.

4.2 The Tau-Path Method

The tau-path algorithm (Yu et al. 2011) is a method of reordering a bivariate
$(X, Y)$ sample so that Kendall’s tau measure of correlation is decreasing. If con-
tributions to the correlation decrease to 0 and remain there, that is evidence that
association between the two variables is limited to a subpopulation represented by
the subsample supporting the association. Since Kendall’s tau depends only on the
ranks of the $X$ and $Y$ variables, finding the point at which contributions to the cor-
relation hit 0 is exactly the same problem as finding the stage at which the rankings
of $X$ and $Y$ become uninformative about each other. Note that stages in this case
are determined by the ordering of the sample according to the tau-path, and not by
the ordering of sample values of either $X$ or $Y$.

Kendall’s tau measure of association (Kendall and Gibbons 1990) for a pair of
random variables is a single number that encapsulates the degree of monotone asso-
ciation over all observations. It is entirely possible that the association is strong for
some subset(s) of the $n$ observation pairs, and weak for some other subset(s). Fur-
thermore, reordering the observation pairs will redistribute the association among the
newly formed subsets, even though Kendall’s tau for the full sample is unchanged.
It should therefore be possible to reorder the observations in such a manner that the sample tau coefficients, calculated for 2, 3, \ldots, \textit{n} nested and reordered observation pairs, form a monotone decreasing path toward Kendall’s tau. We follow the notation of Yu et al. (2011) below.

For independent variables \((X_1, Y_1)\) and \((X_2, Y_2)\) from a bivariate distribution, the probabilities of concordance and discordance are, respectively,

\[
p_c = P[(X_1 - X_2)(Y_1 - Y_2) > 0] \\
p_d = P[(X_1 - X_2)(Y_1 - Y_2) < 0]
\]

Kendall’s tau for the distribution is given by \(\tau = p_c - p_d\). For a random sample of \(n\) pairs \((X_1, Y_1), \ldots, (X_n, Y_n)\), an unbiased estimator of \(\tau\) is Kendall’s tau coefficient, given by

\[
T = \left(\sum_{1 \leq i < j \leq n} c_{i,j}\right) / \binom{n}{2}
\]

where

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

If \(\pi\) is a permutation \(\{\pi(1), \pi(2), \ldots, \pi(n)\}\) on \(e = (1, 2, \ldots, n)\), and \(C = C_{[i,j]}\) denotes the random concordance matrix of the original observations, then

\[
(\pi \circ C)_{[i,j]} = C_{[\pi(i), \pi(j)]}
\]

is the concordance matrix of the permuted sample. If \((\pi \circ C)^{(k)}\) denotes the leading \(k \times k\) matrix of \((\pi \circ C)\), then the average of its off-diagonal elements, given by

\[
T_k(\pi) = \left(\sum_{1 \leq i < j \leq k} C_{[\pi(i), \pi(j)]}\right) / \binom{k}{2}
\]

(4.1)
is Kendall’s tau coefficient for the first $k$ observations in the permuted sample, and the sequence $[T_2(\pi), T_3(\pi), \ldots, T_n(\pi)]$ is denoted the tau-path. The goal, then, is to determine the *sequentially maximal monotone decreasing* tau-path, given by $T_2(\pi) \geq T_3(\pi) \geq \cdots \geq T_n(\pi)$, and the corresponding nested subpopulations $S_2(\pi) \subset S_3(\pi) \subset \cdots \subset S_n(\pi)$, where $S_k(\pi) = \{\pi(1), \pi(2), \ldots, \pi(k)\}$, $k = 2, \ldots, n$. Neither the tau-path nor the permutation causing it is necessarily unique, due to the possibility of ties.

Yu et al. (2011) further provide two backward conditional search algorithms to determine the tau-path, the Fast Backward Conditional Search (Fast BCS) and the Full Backward Conditional Search (Full BCS) algorithms. The Fast BCS algorithm is shown in 4.1 and determines a locally optimal tau-path through the data. It starts with the full complement of observations and eliminates, one step at a time, the observation that is least associated with the others, where association is measured by Kendall’s tau. The algorithm resolves ties in a manner that results in local optimality.

The Full BCS algorithm is given in 4.2 and tracks down the most plausible subsets in the data. Due to the extreme computation time needed for the Full BCS algorithm, it is recommended more as an accessory to the Fast BCS algorithm, and is used to further refine the shortlist already created by the Fast BCS algorithm, rather than as a standalone algorithm.

We use the Fast BCS algorithm to determine a locally optimal tau-path through the data. The Fast BCS algorithm starts with the full complement of observations and eliminates, one step at a time, the observation that is least associated with the others, where association is measured by Kendall’s tau. The algorithm resolves ties
0. Set starting values: $\pi = e; \ i = n; \ Tie = \text{list of } n \ NULL \ vectors.$

1. Repeat Steps a, b, c and d until $T_i = 1$

   a. For each $j = 1, \ldots, i$,
      
      let $s_j = \sum_{u=1}^{i} C_{i(u),i(j)}$ be the $j$-th column sum of $(\pi \circ C)^{(i)}$
   
   b. Let $s_{[1]}^i$ be the minimum of these column sums.
   
   c. Let $v = \{ \pi(l), l = 1, \ldots, i | s_{\pi(l)}^i = s_{[1]}^i \}$
      
      IF $\pi(l)$ is NOT unique, THEN
      
      Randomly select $\pi(l)$ from $v$, transpose $\pi(i)$ and $\pi(l)$
      
      Store all such $\pi(l)$ in $Tie: Tie[i] = v$;
      ELSE $\pi(l)$ is unique, transpose $\pi(i)$ and $\pi(l)$
   
   d. For $k = n, \ldots, i + 1$,
      
      IF $\pi(i) \in Tie[k]$, THEN
      
      For $h \in \{i, k\}$, let $q_h = \sum_{a=i}^{b} C_{i(a),i(h)}$; $b = i, \ldots, k$.
      
      IF All($q_h \geq q_j$) and Any($q_h > q_j$), THEN
      
      Transpose $\pi(i)$ and $\pi(k)$
      
      Reset $i = k - 1$ and $Tie[m] = NULL, \ \forall m \leq k - 1$
      
      STOP and GO TO step a.
      
      IF No transposition happen $\forall k \geq i + 1$, set $i = i - 1$, GO TO step a.

2. Set $k = i$, $T_2 = \ldots = T_k = 1$

3. Output the tau-path statistic $\{T_2, \ldots, T_n\}$ from the final permutation $\pi$.

Figure 4.1: Fast Backward Conditional Search (Fast BCS) algorithm.
0. Set starting values:
   \( \omega_{n}[1,] = e, i = n; \)
1. Repeat Steps a, b, c and d until \( T_i = 1 \)
   a. \( i = i - 1; \)
      Set \( s_{[i]}^j = n, \omega_i = NULL, L_i = \text{num of row vectors in } \omega_{i+1}; \)
   b. For \( i = 1, \ldots, L_i, \)
      \begin{align*}
      &\text{Let } \pi_l = \omega_{i+1}[l,], \\
      &s_{[i]}^{(i, \pi_l)} = \sum_{u=1}^{i+1} C_{[\pi_l(u), \pi_l(j)]}, j = 1, \ldots, i + 1, \\
      &s_{[i]}^{(i, \pi_l)} \text{ be the minimum of these column sums,} \\
      &v^l = \{ \pi_l(j), j = 1, \ldots, i + 1 | s_{[i]}^{(i, \pi_l)} = s_{[i]}^{(i, \pi_l)} \}; \\
      &\text{IF } s_{[i]}^{(i, \pi_l)} < s_{[i]}^j, \text{ THEN} \\
      &\text{Set } s_{[i]}^j = s_{[i]}^{(i, \pi_l)} \text{ and } \omega_i = NULL, \\
      &\forall \pi_l(j) \in v^l, \text{ let } \pi_l \setminus \pi_l(j) \text{ be the next row vector of } \omega_i; \\
      &\text{ELSE IF } s_{[i]}^{(i, \pi_l)} = s_{[i]}^j, \text{ THEN} \\
      &\forall \pi_l(j) \in v^l, \text{ let } \pi_l \setminus \pi_l(j) \text{ be the next row vector of } \omega_i; \\
   c. \text{ Let } \omega_i = \text{all unique row vectors in } \omega_i; \\
   d. \text{ Calculate } T_i; \\
2. \text{ Set } k = i, T_2 = \ldots = T_k = 1 \\
3. \text{ Output } \{ \omega_k, \ldots, \omega_n \} \text{ and tau-path statistic } \{ T_2, \ldots, T_n \}.

Figure 4.2: Full Backward Conditional Search (Full BCS) algorithm.
in a manner that results in local optimality and returns a permuted sample of the observations in tau-path order.

The tau-path test has already uncovered patterns in chemogenomics and marketing data from a financial services company. Yu (2009) reports that the tau-path test outperforms Kendall’s tau test when applied to the Gumbel, Clayton and Frank copula families, and to the Gaussian copula. An additional finding was that the power of the tau-path test depended on the strength and proportion of the associated subpopulation, and not on the functional form of the copula model being used.

4.3 Detecting the End of Association Between Gene Expression and Chemical Activity Over Subpopulation of Cancer Cell-Lines

As noted in the previous section, the outputs of the tau-path algorithm include the ordered observation pairs from most to least associated, and the sequentially maximal monotone decreasing tau-path. When applied to stages in the multistage model, this ordering produces a descent model as studied in Section 3.3. The transformation of the tau-path into the penalty function used by the Top-$K$ specification is straightforward. If $A_k(\pi)$ and $D_k(\pi)$, $k = 1, \ldots, n$, are respectively the number of concordant and discordant pairs of observations in tau-path order, Equation (4.1) becomes

$$ T_k(\pi) = \left( A_k(\pi) - D_k(\pi) \right) / \binom{k}{2}, \ k = 2, \ldots, n. \tag{4.2} $$

Since $A_k(\pi) + D_k(\pi) = \binom{k}{2}$, $k = 2, \ldots, n$, Equation (4.2) gives

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

The penalty function is finally calculated as

$$ V_k(\pi) = D_k(\pi) - D_{k-1}(\pi), \ k = 3, \ldots, n, $$
and serves as an input for the Top-K specification algorithm that generates the MAMLE curve for the ordered pairs of observations.

The NCI-60 data to which this algorithm is applied is the same one studied by Yu et al. (2011), consisting of five genes ASNS, IGFBP6, LDOC1, NQO1 and PIK3R3, and five cancer-inhibiting compounds 40, 757, 1771, 2039 and 3062 from the quassinoid class of chemicals. The quassinoids are a group of complex, highly oxygenated, degraded triterpene, which have been shown to display marked antileukemic activity. All genes and compound effects are measured across 60 cancer cell-lines. We first ran all 25 gene-compound pairs through the Fast BCS algorithm and recovered both the order of observation pairs from strongest associated to weakest, and the tau-path across the observations. We then applied the outputs of the tau-path algorithm to the Top-K algorithm to calculate the MAMLE curve for each gene-combination pair.

For the MAMLE calculations, we selected a window of five stages as an optimal width for the size of the data under study, so as not to over-smooth the estimates.

In order to determine how far down the ordered list the association in the subpopulation ends, we need to generate the null distribution for the MAMLE curve. The null hypothesis presumes that the second assessor’s ranking scheme has no agreement with that of the first assessor, and therefore, that the second assessor’s ranks are completely random relative to the first. To this end we anchored the first assessor’s ranks in the order from 1 to 60, and generated 10,000 random permutations of the 60 ranks to simulate the second assessor’s choices under the null hypothesis. As with the gene-compound data, we ran these simulations through the Fast BCS algorithm to extract both the tau-path and the orderings of the observation pairs. The outputs were then run through the Top-K algorithm to produce the stage-wise MAMLE
curve for each simulation, and we once again used a window width of five stages in this computation. Finally, as noted in Section 3.2.2, we calculated the stage-wise 95th quantiles of the MAMLEs for the simulations to create the rejection region, depicted in Figure 4.3.

In our comparison of the MAMLEs for the 25 gene-compound pairs with the null distribution, our goal was to find instances where the Top-$K$ specification would provide an early degeneration point between signal and noise in the ordered observation pairs. We uncovered such an outcome in the association between gene LDOC1 and
Figure 4.4: 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.

compound 1771 (Figure 4.4). Here the MAMLEs computed for the gene-compound data exceed the quantile curve at stages 28–35, suggesting that the endpoint is best estimated by stage 28, following the stopping rule described in Section 3.2.2. This outcome is supported by computations that use a window width of ten stages, which suggest that the endpoint is stage 27. In either instance, the recommendation is therefore to focus on a subset of the full data set, which could result in more accurate analyses and a potential cost savings during subsequent experimental work.
Specification according to the tau-path ordering thus provides a precise method for estimating the size of the underlying subset of the data that supports association. The application of the Top-$K$ specification as a stopping rule to the tau-path algorithm shows that it is capable of detecting the stage where the signal turns to noise, which suggests numerous applications in an era where the growth of data is explosive.
Chapter 5: Application to the Naming of Newborns in the United States

In this framework 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 objects in the current year. This ordering starts with the most favored object of the current year and proceeds to the least favored object. At each stage, all remaining objects (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.

An innovation in this chapter is the adjustment applied to the stage-wise penalties to handle overlapping but different sets of ranked objects 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) in the particular context of baby names is provided in Section 5.1. In Section 5.2, two-part penalties are proposed to handle the case of overlapping but different lists. Simulations that mimic the baby name database are used to evaluate our model’s efficacy in Section 5.3. Section 5.4 describes the Social Security Administration baby names data set and presents the results discovered by our methodology. A concluding discussion is presented in Section 5.5.

5.1 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.

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 uninformati ve 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,

$\textit{Jacob, Ethan, Michael, Jayden, William, Alexander, Noah, Daniel, Aiden, Anthony,}
\textit{Joshua, Mason}$,

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

$\textit{Jacob, Mason, William, Jayden, Noah, Michael, Ethan, Alexander, Aiden, Daniel,}
\textit{Anthony, Matthew}$.

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, $\textit{Jacob}$, matches the most popular name in 2010, the penalty for the first stage is 0. In
the second stage, the 2011 pick of $\textit{Mason}$ was the twelfth most popular name in 2010. Since $\textit{Jacob}$ has already been accounted for in 2011, ten other more popular names from 2010 were overlooked before $\textit{Mason}$, and therefore the penalty for the second stage is 10. Similarly, the penalty for the third stage is 3 (since $\textit{Ethan, Michael}$
and $\textit{Jayden}$ were ignored before picking $\textit{William}$). Continuing in this manner, the
stage-wise penalties for the dozen names are

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

This penalty assignment can now be modeled by means of a truncated geometric
probability distribution as given in Chapter 3. Maximum likelihood estimation pro-
ceeds as before, as does the determination of the endpoint of agreement by means of
the stopping rule. As new names are likely to enter every observed year, the next section describes the assignment of a two-fold penalty to properly penalize these new entrants.

5.2 Assigning Penalties to Overlapping but Different Lists

The manuscript 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 \( n_{\text{new}} \) 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 \( (n_{\text{new}} + 1)/2 \) to account for the size of the incoming block of names.

5.3 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 5.1 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 (3.1) and (3.2) 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 5.2. Lastly, the average penalty for a given window was calculated using a Winsorized mean with trim \(1/(\text{window width})\) 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 5.1 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 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.

### 5.4 Application: Trends in Popular Baby Names in the U.S.

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 3.2.2.
5.4.1 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 applicants 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).

5.4.2 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 3.2.2, 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
Figure 5.2: 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.

rejection bound to which all the MAMLE curves derived from the baby names data were compared. In both instances, a window width of 40 was used. 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 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 5.2 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.

5.5 Discussion

As discussed in earlier chapters, estimation in the multistage ranking model has been refined 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, the concepts of position and median penalties have been incorporated which, in conjunction with Winsorized means, extend the applicability of the method to handle lists of overlapping but different objects.

Simulations show that the new approach recognizes the overall shape of the parameter curve in the context of newborn names; 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. As described in the previous subsection, the application of this method to the database of popular baby names 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 manuscript 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 objects 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.
This dissertation proposes an approach to determine the point in the rank sequence where the signal degenerates into noise, allowing for possible agreement late in the ranked list. This is a real-world problem with important ramifications. As data sets explode in size, effective techniques to streamline analyses and capture the endpoint of agreement become increasingly important. Extensions to our algorithm can have far-reaching applications, including the merging of results from several search engines, the consolidation of scores of multiple medical tests conducted on a fixed patient group, and the synthesis of patients’ drug responses in a clinical study. For example, the laboratory study of each gene-compound interaction described in Chapter 4 can cost several millions of dollars, and our algorithm, should it detect an early endpoint, could result in substantial savings of time and money, and speed up the findings and, potentially, the arrival of a cure in the market.

The comparison between the stopping rule proposed here and that conceived by Hall and Schimek (2012) clearly shows that the MAMLE stopping rule is not only more robust, but also depends on fewer tuning parameters than the Hall–Schimek stopping rule. By virtue of using the observed ranks in a more nuanced manner, i.e., by creating the possibility of multiple penalties at every stage rather than just the binary outcome of the Hall–Schimek approach, the MAMLE approach provides
a clearer picture of agreement between the reference ranks and the observed ranks at every stage. The MAMLE approach also handles well the random shocks caused by the introduction of new objects to the second assessor’s list, and even without the use of Winsorized trimming. The MAMLE methodology is quick to code and execute, and is able to handle lists of length about 20,000 objects in mere minutes on today’s laptops.

We plan to explore the possibility of developing the methodology of our locally smoothed stage-wise model even further to combine three or more lists of ranked objects, estimate parameters and study their asymptotic properties, and further investigate and refine the approach to determine the endpoint of agreement. One immediate benefit is the bundling of our stopping rule with the tau-path algorithm, to demonstrate not only how two variables are associated, but also to identify a precise stage in the sequence where agreement ends.

The adjustments to window width described in Subsection 3.2.2, and as partially implemented in Subsection 3.3.2, are first steps in developing a fully automated algorithm to self-tune the window width based on key features of the underlying data. We plan to develop an algorithm that determines, at every stage, how wide the window should be; this may be represented either as a fixed number or as a percentage of the total number of remaining stages. Another avenue for exploration is the position of the first stage for each window. At every stage, the algorithm currently uses a backward-looking window by considering the penalties from the current stage and the previous $w - 1$ stages. It will be interesting to examine the impact of windows where the current stage is instead the midpoint of the window.
The stopping rule currently provides a point estimator of the endpoint between signal and noise. Research that expands the stopping rule to instead provide an interval may yield additional insights.

Preliminary analyses of marketing and financial data suggest that our approach may be particularly suited in the context of copulas, especially those with asymmetric tails, and we intend to explore the behavior of the algorithm in these particular situations.

To make our method more relevant to rank aggregation, for example in resolving disagreements among rankings from three or more value systems, we are interested in finding not only where agreement degenerates to pure noise, but also where there is a clear lack of consensus among the rankings. In the specific case of electoral data, where rankings are available from a large number of assessors and a tentative consensus order from weakest candidate to strongest is present, this order may be exploited to determine a robust aggregation of the ranked lists, as well as an endpoint of agreement between all the assessors.

In Section 4.1, the tau-path approach was first used to reorder data from highest to lowest association, and then the MAMLE curve \( \{\hat{\theta}_j\} \) was used to identify the endpoint of association. In cases where the endpoint is the list length \( n \), the MAMLE curve may still be modified to seek a sudden drop in agreement occurring earlier in the list. This would be useful in learning whether or not an overall association was driven mainly by a subpopulation. In this way, a test for detecting whether a modest overall correlation is truly homogeneous among all subsamples, or rather it is driven almost entirely by a single subpopulation, can be developed. In the context of the multistage model, this test for concentration of association can be based on a generalized likelihood ratio test.
of non-constant decreasing \( \{ \theta_j \} \) versus constant \( \theta_j = \theta \) in all stages \( j \). A particular application is in finding the endpoint of association between levels of expression from pairs of genes in cells that are exposed to chemicals. This in turn may help identify both a gene network and the class of chemicals to which the network responds.

To the best of our knowledge, ours is the first comprehensive study of trends in the baby name 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, the fitting of distributions to 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 availability of state-level baby name data from the Social Security Administration also creates the possibility of studying state-level trends over the last century and, consequently, the geographic similarities and potentially migration of naming trends across the country.

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 also be investigated in future work.
Appendix A: Relevant MAMLE Code Extracts

Main Module for Computing the MAMLE

```r
ma.mle <- function(obsrank, refrank = seq(along = obsrank),
   n0 = length(obsrank), w = round(0.1 * n0)) {
   ma.theta <- rep(0, n0 - w)
   V <- Vscore(obsrank, refrank) # Vscore handles both full and partial
   matches between obsrank and refrank
   for (i in 1:(n0 - w)) {
     s <- i - 1
     Vbar <- mean(V[i:(i + w - 1)])
     ma.theta[i] <- theta.scale(Vbar, s = s, w = w, n0 = n0)
   }
   ma.theta <- c(rep(NA, w), ma.theta)
   ma.theta
}
```

Assignment of Penalties to the Observed Ranking

```r
Vscore <- function(x, REF) {
  N <- length(REF)
  ref <- REF
  n <- length(x)
  V <- rep(0, n)
  n.new <- sum(is.na(match(x, REF))) # calculated to assign penalty,
  and not used again
  penalty <- (n.new + 1)/2
  for (i in 1:n) { # assign the correct penalty to each term in
    the observed list
    worst <- length(ref) + penalty # length(ref) reduces over time,
    but penalty is fixed
    m <- match(x[i], ref)
    if (is.na(m)) {
      ```
\[ V[i] \leftarrow \text{worst} \]

\}

\text{else} \{
\[ V[i] \leftarrow m - 1 \]
\text{ref} \leftarrow \text{ref}[\neg m] \quad \text{# reduces reference list by one term so that penalties are correct for next iteration}
\}
\}
\text{return}(V)
\}

\textbf{Determination of the MAMLE \{\theta_j\}}

\text{theta.scale} \leftarrow \text{function}(\text{Vbar}, s, w, n0, \text{inc} = 0.001) \{
\quad r \leftarrow \text{c}(\text{inc}/10000, \text{inc}/1000, \text{inc}/100, \text{inc}/10, \text{seq}(\text{inc}, 1 - \text{inc}, \text{inc}),
\text{\quad 1 - inc}/10, 1 - \text{inc}/100, 1 - \text{inc}/1000, 1 - \text{inc}/10000,
\text{\quad 1 - inc}/1e+05)
\quad y \leftarrow \text{Lprime}(r, s, t = s + w, n0)
\quad sr \leftarrow \text{seq}(\text{along} = r)
\quad \text{under} \leftarrow y \leq \text{Vbar}
\quad j0 \leftarrow \text{max}(\text{c}(1, sr[\text{under}]))
\quad r0 \leftarrow r[j0]
\quad \text{theta} \leftarrow -\log(r0)
\quad \text{return(\text{theta})}
\}

\textbf{Computation of the Derivative of the Maximum Likelihood Estimator}

\text{Lprime} \leftarrow \text{function}(r, s = 0, t = n, n = 200) \{
\quad \text{tot} \leftarrow 0
\quad \text{for} \ (i \ \text{in} \ 0:((t - s) - 1)) \{
\text{\quad nis} \leftarrow (n - s) - i
\text{\quad odds} \leftarrow (r^{nis})/(1 - r^{nis})
\text{\quad tot} \leftarrow \text{tot} + nis * \text{odds}
\}
\quad \text{ans} \leftarrow (r/(1 - r)) - (\text{tot}/(t - s))
\quad \text{return(\text{ans})}
\}

\textbf{Computation of Stage-Wise Probabilities}

\text{probV} \leftarrow \text{function}(j, \text{theta} = 1, n, \text{mintheta} = 1e-06) \{
\quad \text{probs} \leftarrow \text{rep}(0, n - j + 1)
\quad \text{if} \ (\text{theta} > \text{mintheta}) \{
\}
r <- exp(-theta)
for (i in seq(n - j + 1)) {
    probs[i] <- r^(i - 1)
}
probs <- (1 - r)/(1 - r^(n - j + 1)) * probs
return(probs)

Simulation of the Stage-Wise Penalties

generate.Multistage <- function(theta) {
    nm <- length(theta)
    V <- rep(0, nm)
    n <- nm + 1
    items <- 1:n
    nu <- rep(0, n)
    for (j in 1:nm) {
        pr <- probV(j, theta[j], n)
        if (any(is.na(pr)))
            stop(j)
        V[j] <- sample(0:(n - j), size = 1, prob = pr)
        i <- V[j] + 1
        nu[j] <- items[i]
        items <- items[-i]
    }
    nu[n] <- items[1]
    nu0 <- order(nu)
    nu0
}

Computation of the Endpoint of Agreement

stoppingstage <- function(null.mle, data.mle, window.width, alpha = 0.05) {
    # inputs
    # null.mle = the stage-wise MLE of the simulated
    # null distribution
    # data.mle = the stage-wise MLE of the actual data
    # alpha = level of significance (in decimals, so 0.05)
    # output
    # stopstage = the earliest stage where
    # a) the data MLE exceeds the null distribution quantile and
b) the data MLE exceeds the null distribution quantile for at most alpha percent of the remaining stages

```r
stage.count <- length(data.mle)
null.quantile <- apply(null.mle, 2, quantile, probs = (1 - alpha), na.rm = T)

stop.stage <- NA

if (length(which(data.mle > null.quantile)) == 0)
    stop.state <- 0 else {
    rejection.stages <- data.frame(which(data.mle > null.quantile))
    len <- nrow(rejection.stages)
    names(rejection.stages)[1] <- "stage"
    rejection.stages$ratio <- NA
    for (i in 1: len){
        rejection.stages$ratio[i] <- (len - i) / (stage.count - i)
    }
    stop.stage.rec <- rejection.stages[rejection.stages$ratio <= alpha, ]
    stop.stage <- min(stop.stage.rec$stage) - window.width
    }
stop.stage
```
Appendix B: The Hall–Schimek Algorithm

The code below follows the algorithm detailed in Hall and Schimek (2012). It does not take into consideration the possibility that \( p \) depends on the proximity measure \( \delta \), as noted in Section 3.4.

The Moderate-Deviation-Based Approach from Hall and Schimek (2012)

```r
mod_dev <- function(nu = 6, r = 1.2, ref_rank, obs_rank, d = 50, 
               C_val = 0.3) {
  # Paper: Moderate-Deviation-Based Inference for Random Degeneration 
  # in Paired Rank Lists 
  # Authors: Hall, P. and Schimek, M. G. 
  # Year: 2012 
  # Source: Journal of the American Statistical Association, 107:498, 
  # pp. 661-672 
  # nu = size of pilot sample (r x nu = end point of calculated p-hat 
  # sequence) 
  # r = multiplier of nu (r x nu = end point of calculated 
  # p-hat sequence) 
  # ref_rank = reference rank (the anchor against which observed rank 
  # is compared) 
  # obs_rank = observed rank (provides 1-0 penalties based on 
  # differences with corresponding reference rank). Note that 
  # obs_rank should be the same length as ref_rank 
  # d = allowed discrepancy between reference rank and observed rank 
  # at a specific stage. If \(|\text{ref_rank} - \text{obs_rank}| \leq d\), then the 
  # corresponding I_val is zero 
  # C_val = controls moderate deviations, needs to exceed 1/4 (p. 663) 

  # Initialize variables
```

68
\begin{verbatim}
s <- 1
z_nu <- sqrt(C_val * log(nu) / nu)
N <- length(ref_rank)
stop_pos <- 0 # may be unnecessary
extreme_pos <- 0
center_pos <- 0
jump <- round(r * nu)
p <- rep(0, (1 + jump)) # will store (r*nu) pilot probabilities
       at each stage
J_minus2 <- J_minus1 <- J_now <- 0
loop <- T
I_val <- rep(0, N) # vector of Bernoulli penalty between reference
       ranks and observed ranks

# Calculate Bernoulli stream
for (i in 1:N) {
    I_val[i] <- (abs(obs_rank[i] - ref_rank[i]) <= d) # equals 1
       if d-close and 0 if d-apart
}

s <- 1 # not really necessary, just shows this is first iteration
# Perform iteration 1 (outside loop since takes non-repeating
       values)
center_pos <- 1 + jump
extreme_pos <- center_pos - jump
for (k in 1 : center_pos) {
    p[k] <- mean(I_val[k : (k + nu - 1)]) # nu terms averaged
       for each p
}
if (length(which(p - 0.5 <= z_nu)) == 0) {
    J_now <- center_pos
} else J_now <- min(which(p - 0.5 <= z_nu))
s <- 2 # stage
while (loop) {
    pos_neg <- (-1)^s
center_pos <- J_now
    # s even: extreme_pos > center_pos
    # s odd: extreme_pos < center_pos
    extreme_pos <- center_pos + (pos_neg * jump)
\end{verbatim}
# calculate pilot sample probabilities
for (k in 1:(1 + jump)) {
    #p[k] <- mean(I_val[(extreme_pos - (pos_neg * (k - 1))):
                  (extreme_pos - (pos_neg * (k + nu - 2)))))
    indices <- (extreme_pos - (pos_neg * (k - 1))):
                 (extreme_pos - (pos_neg * (k + nu - 2)))
    indices <- indices[indices > 0]
    p[k] <- mean(I_val[indices])
}

J_minus2 <- J_minus1
J_minus1 <- J_now

if ((s %% 2 == 1) & (length(which(p - 0.5 <= z_nu)) == 0)) {
    J_now <- center_pos
}
if ((s %% 2 == 0) & (length(which(p - 0.5 <= z_nu)) == 0)) {
    J_now <- extreme_pos
}
if ((s %% 2 == 1) & (length(which(p - 0.5 <= z_nu)) != 0)) {
    J_now <- extreme_pos + min(which(p - 0.5 <= z_nu)) - 1
}
if ((s %% 2 == 0) & (length(which(p - 0.5 <= z_nu)) != 0)) {
    J_now <- extreme_pos - min(which(p - 0.5 <= z_nu)) + 1
}

# four if-conditions to break the 'while' condition and
# end the algorithm
if ((s %% 2 == 0) & (J_now == J_minus1)) { # Condition 1
    stop_pos <- J_now
    loop <- F
    warning("Condition1")
} else if ((s %% 2 == 1) & (J_now <= J_minus2)) { # Condition 2
    stop_pos <- J_minus2
    loop <- F
    warning("Condition2")
} else if ((s %% 2 == 0) & ((J_now - jump) <= 1)) { # Condition 3
    stop_pos <- J_min...
on p. 664 of H/S
stop_pos <- J_minus1
loop <- F
warning("Condition3"
} else if ((s %% 2 == 1) & ((J_now + jump) > N)) { # Condition 4:
    if full list length is exhausted
    center_pos <- J_now
    extreme_pos <- N
    pos_neg <- 1
    for (k in 1:(1 + extreme_pos - center_pos)) {
        indices <- (extreme_pos - (pos_neg * (k - 1))):
        (extreme_pos - (pos_neg * (k + nu - 2)))
        indices <- indices[indices > 0]
        p[k] <- mean(I_val[indices])
    }
    J_minus2 <- J_minus1
    J_minus1 <- J_now
    J_now <- extreme_pos - pos_neg * min(which(p - 0.5 <= z_nu))
    stop_pos <- J_now
    loop <- F
    warning("Condition4"
} 
s <- s + 1
}
return(stop_pos + nu/2)


