EXTENSION OF KENDALL'S TAU USING RANK-ADAPTED SVD TO IDENTIFY CORRELATION AND FACTIONS AMONG RANKERS AND EQUIVALENCE CLASSES AMONG RANKED ELEMENTS.

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

---

by
Kathleen Campbell
August, 2014

Examining Committee Members:

Dr. Richard M. Heiberger, Chair Temple University Statistics Department
Dr. William Wei, Temple University Statistics Department
Dr. Pallavi Chiturri, Temple University Statistics Department
Dr. Joseph Pigeon, External Member, Villanova University Statistics Department

ABSTRACT

EXTENSION OF KENDALL'S τ USING RANK-ADAPTED SVD TO IDENTIFY CORRELATION AND FACTIONS AMONG RANKERS AND EQUIVALENCE CLASSES AMONG RANKED ELEMENTS.

Kathleen Campbell
Doctor of Philosophy

Temple University, August, 2014

Dr. Richard M. Heiberger, Chair

The practice of ranking objects, events, and people to determine relevance, importance, or competitive edge is ancient. Recently, the use of rankings has permeated into daily usage, especially in the fields of business and education. When determining the association among those creating the ranks (herein called sources), the traditional assumption is that all sources compare a list of the same items, (herein called elements). In the twenty-first century, it is rare that any two sources choose identical elements to rank. Adding to this difficulty, the number of credible sources creating and releasing rankings is increasing. In statistical literature, there is no current methodology that adequately assesses the association among multiple sources. We introduce rank-adapted singular value decomposition (R-A SVD), a new method that uses Kendall's τ as the underlying correlation method. We begin with P, a matrix of data ranks. The first step is to factor the covariance matrix K as follows:

$$K = \text{Cov}(P) = VDV'$$

Here, V is an orthonormal basis for the rows that is useful in identifying when sources agree as to the rank order and specifically which sources. D is a diagonal of eigenvalues. By analogy with singular value decomposition

(SVD), we define U^* as

$$U^* = PV'D^{-1}$$

The diagonal matrix, D, provides the factored eigenvalues in decreasing order. The largest eigenvalue is used to assess the overall association among the sources and is a conservative unbiased method comparable to Kendall's W. Anderson's test determines whether this association is significant and also identifies other significant eigenvalues produced by the covariance matrix. Using Anderson's test [9], we identify the a significantly large eigenvalues from D. When one or more eigenvalues is significant, there is evidence that the association among the sources is significant. Focusing on the a corresponding vectors of V specifically identifies which sources agree. In cases where more than one eigenvalue is significant, the a significant vectors of V provide insight into factions. When more than one set of sources is in agreement, each group of agreeing sources is considered a faction. In many cases, more than one set of sources will be in agreement with one another but not necessarily with another set of sources. Using the a significant vectors of U^* provides different but equally important results. In many cases, the elements that are being ranked can be subdivided into equivalence classes. An equivalence class is defined as subpopulations of ranked elements that are similar to one another but dissimilar from other classes. When these classes exist, U^* provides insight as to how many classes and which elements belong in each class.

In summary, the R-A SVD method gives the user the ability to assess whether there is any underlying association among multiple rank sources. It then identifies when sources agree and allows for more useful and careful interpretation when analyzing rank data.

ACKNOWLEDGEMENTS

I want to thank Dr. Heiberger for his tireless guidance, his willingness to work in an area that he did not find as fascinating as I do, and his personal investment of energy given to help me succeed. A sincere thank you to Dr. Wei and Dr. Chitturi for their availability and flexibility as well as the hours they spent reading through the document itself.

To my husband, Brian, for his understanding, love, and unwavering support even when I was ready to give up. To Brian, Kevin, and Abigail: you are the reason I finished this. To my family and friends who have always stopped what they were doing to help me along the way from babysitting to reading this manuscript to lending an ear to my ideas and concerns.

In loving memory of Thomas J. Campbell, Mary J. Campbell and Fr. William E. Atkinson who were here at the beginning of the journey but are giving guidance from a different level today.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

NEED FOR MEASURES OF ASSOCIATION AMONG RANKERS

1.1 Overview

- *U. S. News and World Report* [36] claims: "Get exclusive rankings and data on nearly 1,800 schools. Use our tools to search for your perfect fit." Meanwhile, *Forbes* [21] promises to find the college with the most value for the money invested. Individuals relying on these and other collegiate rankings ask: "Is there a best college or a group of colleges that are the best fit?" Universities relying on these rankings to increase applications ask: "How can we improve our standing for future rankings?"

- Business sustainability is measured using criteria that range from carbon disclosures to fiscal responsibility. Using ranked lists from sources such as *Newsweek* [6] and *Bloomberg* [1], can we find groups of rankings that agree on which company is truly sustainable?

- Choosing which type of car to purchase requires research because of

the vast selection of car makes and models. *J. D. Power* [4], *Consumer Reports* [2], *Edmunds* [3], and *Kelly's Blue Book* [5] each release rankings to help the buyer choose a car. These sources rate automobiles in the areas of safety, fuel economy, and resale value, to name a few. Can these rankings be used jointly to help consumers identify the best car for them?

Numerous official rankings, such as the aforementioned examples, are currently available over the Internet and in print publications. According to Kendall [30], ranking occurs when a number of items are arranged in order according to some quality that they all possess to a varying degree; these items are said to be ranked. The arrangement as a whole, in which each member has a rank, is called a ranking. Rankings are created using a range of underlying variables, many of which are quantitative. Once an overall algorithm is established, the items considered for ranking (herein referred to as elements) are entered into the algorithm. Using these results, the highest element receives a rank of 1, the next highest is ranked 2, and so forth, and the method is repeated to produce an overall ranking.

Historically, ranks were used in situations when creating a ratio-scale proved to be difficult or when the rankers (herein called sources) lacked tools and/or data to make a meaningful numeric determination. For example, determining the best figure skater involves a panel of judges who rank the same set of skaters on pre-specified skill sets, some of which are hard to quantify. These rankings then would be combined in some way that would indicate the top three. Although there remain instances (e.g., the Olympics) where this is still true, rankings are more prominently used in fields where measureable data are readily available and the interest is in a small sample obtained from a larger population. In these cases, successfully analyzing large quantities of data can be performed only by individuals acquainted with the various distributions of the variables being studied and extensive technological tools used to create a plausible algorithm. Once the data are

correctly analyzed, explaining this algorithm to associates within the company, sharing it with competitors or the general public is either unfeasible or unwise. Therefore, creating a rank scale, which allows pertinent information to be communicated without revealing proprietary equations and avoids extensive explanation, has become popular and sometimes commonplace.

1.2 Rank Applications

Rank metrics are used for numerous business applications (e.g., supply chain decisions), Internet research, and the purchase of expensive consumer goods and services (e.g., education, cars). Current literature strongly focuses on information retrieval (as performed by search engines like Google) due to the very large sample sizes that are created with each search. Little research has been published using the smaller samples (e.g., business and education applications), which are the focus of this research. From a business perspective, rankings are used on a daily basis. Inventory management is one such area. Walmart, which manages an average of *U.S.* $32 billion in inventory [15], uses an inventory management system that ranks products from the most to the least popular on a daily basis. This information in turn permits store managers to make up-to-the-minute purchasing and pricing decisions on inventory. For instance, observing that a certain toy that was full price on Black Friday sold out across a region can help managers stock appropriate quantities of the item throughout the holiday season. Instances similar to this scenario allow Walmart to track the top ten items on a daily or weekly basis instead of waiting until the season ends to properly stock the products in highest demand. Looking at a different business application, multiple sources such as *Newsweek* and *Bloomberg* provide rankings of the most sustainable companies in multiple scenarios from carbon footprint to financial reporting. Many top international firms are reporting on their

environmental and social performance with 80% of the Global Fortune 250 companies now releasing this information [12]. For organizations and customers interested in creating and supporting green businesses, it will make the rankings released more useful.

As mentioned earlier, colleges are often ranked by multiple sources (e.g., *Forbes* and *U. S. News and World Reports*) that generate 'best of' categories. Criteria used to rank educational institutions include admissions rates, SAT scores, student retention and graduation rates, and student evaluations of faculty and institution [14]. For each university, these rankings affect the number of student applications they receive, the amount of financial support they receive from alumni, the number of research grants they obtain, and the fan support garnered for athletic teams. A less favorable rank leads an institution to accept a greater percentage of its applicants, a smaller percentage of its admitted applicants matriculate, and the resulting entering class is of lower quality, as measured by its average SAT scores [34].

These rankings are intended to help consumers, namely prospective students, choose a college or university that matches their specific educational needs. Prospective students and their parents are overwhelmed by choosing a college when so many options exist. Although college visits and fairs provide some information, students are often at a loss as to where to start the search for their ideal college. With Internet and news sites bombarding students and parents with the 'best of' lists, parents and students have begun to rely heavily on these rankings. Finding a method to assess whether the rankings agree and if there are classes of good, better, and best colleges would allow students and parents to make more informed decisions when combing these rankings.

1.3 Financial Impact of Rankings

Releasing rankings is a lucrative business. Sources invest heavily in creating rankings, while organizations spend heavily to be included in these rankings. For instance, collegiate sports programs wishing to be listed within the National Collegiate Athletic Association (NCAA) sports rankings spend significantly to ensure they have the best coaches and most athletic students as part of their program. In 2010, college athletic spending at NCAA Division I schools exceeded *U.S.* $7.9 billion [23]. Likewise, determining which institutions are comparable to a university could help it identify meaningful competition. The university then can work to be the best within its class, with the long-term goal of being elevated into the next highest class and thereby attracting better students and funding. Since more sources than ever are releasing rankings, the task of finding meaningful information becomes more daunting for decision makers.

Further complicating matters is that the underlying equations used in forming most ranked lists are unknown. Google discloses that its ranking algorithm involves more than 100 factors [10], but the company further states that "due to the nature of our business and our interest in protecting the integrity of our search results, this is the only information we make available to the public about our ranking system" [26]. In a similar fashion, journals and other reliable rank sources also conceal their proprietary algorithms. However, decision makers relying on the ranked data need to determine which sources to use and whether to accept the ranking at face value. Identifying whether sources are releasing ranks that are truly associated with other sources may assist in determining a definitive best college, company, or car. When the association across the different sources is strong, classes may exist. Being able to partition the ranked elements into groupings of good, better, and best help identify these classes clearly. Armed with an understanding of these classes, an end user can make well-informed decisions without experiencing information overload.

1.4 The R-A SVD Method

First, efforts will be made to assess recent available comparative rankings to determine whether there is a reliable and measurable association. The most commonly used method, Kendall's *W*, offers information as to the existence of agreement among *all* of the sources but provides very little insight beyond overall association. The method introduced in this research, rank-adapted singular value decomposition (R-A SVD), will provide an unbiased measure of association that is comparable to Kendall's *W*. It also allows end users to identify agreement in cases where only some of the sources agree while identifying specifically which sources.

Second, R-A SVD will identify factions. When more than one set of sources is in agreement, each group of agreeing sources is considered a faction. For example, there are five sources: A, B, C, D, and E. Sources A and B agree strongly with one another, whereas sources C, D, and E also agree strongly with one another but not very strongly with A and B. In this case, there are two factions. R-A SVD allows users to identify the number of factions that exist and the exact sources that belong in each faction.

Third, once association is determined and the factions identified, R-A SVD partitions comparative elements being ranked into meaningful subpopulations called equivalence classes. Specifically, the method examines the rankings within the identified faction(s). Here, we define equivalence class as subpopulations of ranked elements that are similar to one another but dissimilar from other classes. For example, the five sources described earlier (A through E) agree, then we will focus on the rankings of this one faction. If each source agrees that the first five elements are indeed the first five elements but not necessarily agree as to the ordering, we would consider the top five to be one class and call them the 'best' set of elements within this rank system. When these classes exist, this class system helps the user to create groups of items that are the best, second best, and so forth.

Creating tools to help discern patterns among comparative rankings will

increase the information gained by those relying on the rankings especially as new rank sources enter the marketplace. Rankings are widely used by society to make decisions. From choosing inventory to final purchase, multiple rankings influence the various decisions made by managers and purchasers alike. Therefore, methods such as the ones established in this research are imperative to help the user make informed decisions.

CHAPTER 2

LITERATURE REVIEW AND A NEW APPROACH

Before discussing R-A SVD, our new method for comparing rankings, it is necessary to review applied methods of rank comparison for data sets containing identical elements: Kendall's τ, Spearman's ρ and Kendall's W. These methods provide the groundwork for assessing overall association when comparing rank sources using R-A SVD. R-A SVD represents an improvement over Kendall's W by providing a more informative method of measuring association while identifying the structure of association across sources and elements. For the current methods of rank order comparison, adaptations necessitated by ties in the data will be briefly described and cases of feasible bias are highlighted as well. Next, this chapter outlines the specific techniques required to adjust the data for non-overlapping elements and establish the data as near-metric so that R-A SVD can be applied. R-A SVD has multiple steps, the first of which provides a meaningful measure of association and which will be discussed in this chapter. The remaining steps will be outlined and discussed in Chapter 3.

2.1 Classic Methods of Rank Order Comparison

Rank comparisons are measured by considering the association or correlation of the sources ranking the data. Correlation is the degree to which two or more attributes or measurements exhibit a tendency to vary together [7]. Correlation, as it pertains to regression, has the underlying assumption that x and y are continuous variables. In the continuous case, the most common measure of association used is Pearson's correlation coefficient, r. However, rank data are not continuous even when n is large, meaning that Pearson's r cannot be used. The most common methods for assessing association when comparing rankings are Kendall's τ, Spearman's ρ and, for multiple rankings, Kendall's W; these methods will be the focus of this section.

Both Kendall's τ and Spearman's ρ are derived from what Kendall [31] referred to as a general correlation coefficient for which he used the symbol Γ. Although each rank association measure (τ and ρ) have slight variations (they each have a slightly different application and numeric outcome), they can be created from his basic definition of Γ where Γ is:

$$\Gamma = \left[\frac{\Sigma a_{ij} b_{ij}}{\sqrt{\Sigma a_{ij}^2 \Sigma b_{ij}^2}} \right] \tag{2.1}$$

In this Equation 2.1, a_{ij} is regarded as zero if $i = j$ as it is indicative of a tie in the ranking that neither changes the τ numerator nor the distances measure for ρ. When considering a_{ij}, the subscripts i and j each represent specific elements being considered by the first source a. Similarly for b_{ij}, the subscripts i and j would be the same elements being considered by the second source b.

As was discussed in Chapter 1, many current available rankings are created independently. Due to the nature of the rank data considered, instances occur when the lists compared by the sources do not have complete overlap. In these cases, a non-overlapping element will have a placeholder

Table 2.1: Initial Ranking by Each Teacher

Student	Teacher A	Teacher B
c	1	2
d	2	3
e	4	5
f	6	6
g	5	4
h	3	1

added. This process maintains the initial structure and adds ties into the data. Therefore, we will outline adaptations (necessitated by ties) for each method: τ, ρ, and W. In Section 2.2.2, the specific approach to adding a placeholder will be discussed in more detail.

2.1.1 Kendall's τ

Kendall's τ is a measure of association for ordinal rank data that considers identical sets [30]. Like Pearson's correlation coefficient, Kendall's τ estimates the overall association between two variables. Kendall's τ achieves an upper bound of +1 if the two rankings are identical and a lower bound of -1 when the two rankings are ordered inversely. All other orderings will supply a value of τ that falls between -1 and +1. A list with a higher proportion of agreeing terms will approach +1, whereas a list with a higher proportion of disagreeing terms approaches -1. When there is an equal count of agreeing and disagreeing terms, the value of τ will be zero, indicating no association [39].

The following example illustrates how τ is created. First, two sources, teachers A and B, are considered. Each teacher ranked six elements, students c through h, from best student (indicated by the value 1) to weakest student (indicated by the value 6) as is shown in Table 2.1.

To create Kendall's τ, we first examine each pairing by teacher. Table 2.2 shows all paired scores from the teachers in Table 2.1. When the ranks

from both sources (teachers A and B) are in the same order, a value of +1 is given to that pair to indicate what Kendall called concordance [30]. For example, teacher A ranked *cd* as 1^{st} and 2^{nd}, respectively, and teacher B ranked them 2^{nd} and 3^{rd}, respectively. Both rankings follow the natural order of numbers; therefore, they are concordant and are given a value of +1. If the numbers are in inverse order, a value of −1 is given to the pair to indicate discordance. When comparing students **ch**, teacher A ranked them as 1^{st} and 3^{rd}, respectively, and teacher B ranked them 2^{nd} and 1^{st}, respectively. Since the first rank is in natural order and the second is in inverse order, the pair is discordant and it receives a value of −1.

Table 2.2: Each Pair with an Associated Concordance/Discordance Value

	c	d	e	f	g
d	+1				
e	+1	+1			
f	+1	+1	+1		
g	+1	+1	−1	+1	
h	−1	−1	+1	+1	+1

For two rank sources, there are $\binom{n}{2}$ possible paired orderings or outcomes. The numerator for τ is determined by subtracting the number of discordant pairs (Q) from the number of concordant pairs (P). This difference, $P - Q$, is called S. Therefore, S is divided by the total number of possible orderings, which is shown in Equation 2.2.

$$\tau = \frac{S}{\frac{1}{2}n(n-1)} \tag{2.2}$$

When ties occur in the data, the numerator of Equation 2.2 remains the same. Ties add no value to either P or Q. The denominator is reduced, creating what we will refer to as τ_b, where the subscript b is used to show that the method is an adaptation to the standard τ equation. The new formula

follows:

$$\tau_b = \frac{S}{\sqrt{\frac{1}{2}n(n-1) - T}\sqrt{\frac{1}{2}n(n-1) - U}} \tag{2.3}$$

where $T = \frac{1}{2}\sum_t t(t-1)$ represents each set of ties for the first ranked set and $U = \frac{1}{2}\sum_u u(u-1)$ represents each set of ties for the second ranked set. This adapted formula, Equation 2.3, will be the τ used throughout this paper. When testing whether the observed correlation is significant, the distribution of S is considered due to the evidence that it is symmetric. From both the symmetry and the frequency distributions, Kendall [30] found:

- If $\frac{1}{2}n(n-1)$ is even, S takes only even values and attains a maximum frequency when $S = 0$.
- If $\frac{1}{2}n(n-1)$ is odd, S takes only odd values and finds a maximum frequency equivalent at -1 and $+1$, which is still symmetric about zero.
- The frequencies become smaller as they move further away from the maximum value, and S reaches a value of 1 at $\pm\frac{1}{2}n(n-1)$.
- As n increases, the shape of the frequency polygon approaches the normal curve with a mean of zero. When n is greater than 10, this curve of the frequency distribution provides a reasonable approximation to the normal curve. The parameter σ_s^2, which is the variance of the curve, is given by:

$$\sigma_s^2 = \frac{1}{18}n(n-1)(2n+5)$$

The proof that τ is asymptotically normal relies only on the numerator S, which regardless of sample size is symmetric about zero. Therefore, the moments about the mean of odd order vanish. Furthermore, the moments of even order tend towards those of the normal distribution. Using the moments, mean, and variance, Kendall [30] establishes a basis for approximate normality for τ.

2.1.2 Spearman's ρ

Spearman's ρ, a similar approach to Kendall's τ, focuses on the sum of squared differences between the ranks of each element instead of the ordering the elements ([40], [30]). Again, the upper bound is +1 when the rankings are identical and the lower bound is -1 when the rankings are inverted. The value zero implies no correlation. As values approach the limits, the evidence of a relationship increases.

Spearman's ρ can be derived from the generalized correlation coefficient from Section 2.1. In this case, the correlation Equation 2.1 becomes:

$$\Gamma = \frac{\Sigma(p_i - p_j)(q_i - q_j)}{\sqrt{\Sigma(p_i - p_j)^2}}$$

where $(p_i - p_j)$ is the difference of a pair of ranked elements by the first source and $(q_i - q_j)$ is the difference for the same pair ranked by the second source. This equation reduces to Spearman's ρ:

$$\rho = 1 - \frac{6S(d^2)}{n^3 - n}$$

where $S(d^2)$ is the sum of the square of the distances between the two sources ranking each element [39].

When there are ties, the method of creating ρ requires a more complex equation. The new equation is:

$$\rho_a = \frac{\frac{1}{6}(n^3 - n) - S(d^2) - T' - U'}{\sqrt{\left\{\frac{1}{6}(n^3 - n) - 2T'\right\}\left\{\frac{1}{6}(n^3 - n) - 2U'\right\}}} \tag{2.4}$$

where $T' = \frac{1}{12}\sum_t (t^3 - t)$; $U' = \frac{1}{12}\sum_u (u^3 - u)$; t represents each set of ties for the first ranked set; and u represents each set of ties for the second ranked set.

In this case, the variance formula also is adjusted for ties and it becomes:

$$\sigma_\rho^2 = \frac{1}{2n}\left\{\frac{1}{6}(n^3 - n) - 2T'\right\}$$

Again, the adapted Equation 2.4 will be the equation used when considering Spearman's ρ henceforth.

Much like Kendall's τ, a significance test for ρ exists, although it tends to normality slowly. With a large enough sample size ($n \geq 200$) and a continuity correction, a z score can be calculated. The size of the sample required to achieve normality is much larger than Kendall's τ. To test the significance of a value of ρ, there are tables available for values of n that range from 4 to 10. When a table is not available but the sample size is greater than 10, a reasonable approximation of the distribution can be modeled as a t-distribution with $n - 2$ degrees of freedom. The equation to find t is:

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}} \tag{2.5}$$

Note that Equation 2.5 is insoluble when there is perfect agreement or disagreement causing the denominator to be zero. When the sample size exceeds 200, an approximate z test is applicable where the value of z comes from: $z = \rho\sqrt{n-1}$

2.1.3 Kendall's *W*

Most commonly, the measure of correlation for more than two sources is Kendall's coefficient of concordance, called Kendall's W. Created in a way that is linearly related to the average value of Spearman's ρ, the value of W achieves an upper bound +1 when all of the sources rank the elements identically. W approaches a lower bound of zero if the sources show no pattern of agreement. In this case, it is meaningless to have a negative association. When there are three or more sources, it is impossible to have

complete disagreement among them.

Before discussing the steps for Kendall's W, a short example is presented using an $n \times m$ matrix where $n = 5$ students who are ranked by $m = 4$ teachers. In this example, the value 1 represents the highest ranking for a student while 5 represents the lowest ranking. The teachers are labeled A, B, C, D, using the subscript j with values 1 to 4. The top students are labeled e, f, g, h, k, using the subscript i with values from 1 to 5. The first step is to sum all of the teacher's rankings for each student. In Table 2.3, the sum is denoted by R_i, where $R_i = \sum_{j=1}^{m} r_{ij}$. For student e, the sum of the rankings is 8. Each sum is added together, thus creating a total value represented in the table by T, where $T = \sum_{i=1}^{n} R_i$. Next, each element sum, R_i, is squared. Finally, U is solved by taking the sum of these squares, $U = \sum_{i=1}^{n} (R_i)^2$. The sum of squared deviation, the variance of the sum of ranks (called S), can then be determined by:

$$S = \frac{nU - (T)^2}{n} \tag{2.6}$$

The maximum possible values that can be computed for S is found to be $\left(\frac{m^2(n)(n^2-1)}{12}\right)$. Therefore, W is found by:

$$W = \frac{S}{\left(\frac{m^2(n)(n^2-1)}{12}\right)} \tag{2.7}$$

where m is the number of sources and n is the number of ranked elements. In this example where $m = 4$ and $n = 5$, the maximum possible value that can be computed for the variance of the sums of the ranks (the denominator of W) is 160.

For this example,

$$W = \frac{\left[\frac{5\times 844-(60)^2}{5}\right]}{160} = 0.775$$

Using this ratio creates an equation almost equivalent to that of the explained variance in the analysis of variance formula [39]. For small n and m, tables

Table 2.3: Kendall's W Example

	Teacher					
Student	A	B	C	D	$R_i = \sum_{j=1}^{m} r_{ij}$	R_i^2
e	2	2	3	1	8	64
f	1	3	2	3	9	81
g	3	1	1	2	7	49
h	5	5	4	5	19	361
k	4	4	5	4	17	289
					$T = \sum_{i=1}^{n} R_i = 60$	$U = \sum_{i=1}^{n} R_i^2 = 844$

exist to test whether the value of W is significant. As the sample sizes increase and an exact table is not available, the chi-squared distribution provides a reasonably good approximation of the sampling distribution of the estimated W.

As with τ and ρ, an adjustment to the equation must be made when ties occur. According to Sheskin [39], most sources agree that, unless there is an excessive number of ties, the difference between W and the coefficient corrected for ties, W_c, is minimal. The correction for ties will cause a slight increase in the value of W. The only change required is to the denominator, so the new formula is:

$$W_c = \frac{12S}{m^2(n^3 - n - mT')} \tag{2.8}$$

where $T' = \sum_{i=1}^{m} \sum_{a=1}^{s} (t_{a_i}{}^3 - t_{a_i})$, and i represents each source and a represents each set of ties within that source.

As previously discussed, there is an underlying relationship between Kendall's W and Spearman's ρ. If one were to create all of the possible pairs of Spearman's ρ and average them for m separate rankings, the value attained would be ρ_{av}. This average of all the ρ values is a linear function of

W as follows: $\rho_{av} = \frac{mW-1}{m-1}$. This relationship can be equated for all sizes of m including when $m = 2$.

Kendall's W is the most common measure of association, and it looks primarily for a relationship among all of the ranked sources. R-A SVD will use Kendall's τ as a more conservative and realistic value when comparing multiple rank sources. R-A SVD allows the user to identify when there is correlation among some of the sources, or even identify when there are factions of sources that agree with one another. As such, R-A SVD provides insight into the ranked data set that is not available when using W.

2.2 Data Structure: Assumptions and Adaptations

R-A SVD resembles principal component analysis (PCA). It applies SVD to ranked data. PCA (which is discussed in more detail in Section 2.3 and 3.1) and SVD historically are applied to variables from a discrete or continuous data matrix. In contrast, rank-order data traditionally have been considered ordinal data. This section will provide evidence that Kendall's τ is near-metric, or quantitative in nature. A second related issue concerning the use of PCA and rank data involves maintaining the elements within the data. Unlike current large data sets being analyzed using PCA, ranked data (such as the data considered in this research) often have small sample sizes. With m sources considered, Fagin, Kumar, and Sivakumar's [19] method is applied to ensure all of the originally ranked elements remain unlike the typical method of deleting elements which were not given a rank by all sources. This process keeps the small data sets from becoming minute and maintains the initial structure.

2.2.1 Spearman's Footrule: a Metric Application

Unlike Spearman's ρ, which uses the sum of the squared distances of the elements across two rankings, Spearman's footrule measures the sum of the

absolute value of the distances (i.e., its value without regard to sign) between the ranks for each element ([40], [19]). The footrule was chosen because its structure greatly facilitates the study of the exact and asymptotic properties as a metric [37]. The footrule is the L_1 distance between two permutations. Fagin et al [19] define the footrule as $F(\sigma_1, \sigma_2)$ as follows:

$$F(\sigma_1, \sigma_2) = \sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)| \tag{2.9}$$

where the maximum value of $F(\sigma_1, \sigma_2)$ is $\frac{n^2}{2}$ when n is even and $\frac{(n+1)(n-1)}{2}$ when n is odd [19]. The distance between two sources ranking the same element, $\{\sigma_1(i) - \sigma_2(i)\}$, is always positive due to the absolute value in the formula. The sum of these distances is the value used in Spearman's footrule.

According to Fagin et al [19], all distance measures in a distance equivalence class have similar properties. (Here, distance equivalence class is distinct from the term equivalence class, which in this paper refers to subpopulations of ranked elements that are similar to one another but dissimilar from other classes.) Therefore, Spearman's ρ and footrule as well as Kendall's τ measure approximately the same thing. Thus, they can be clustered into one large distance equivalence class. Spearman's footrule, which is the one measure that does not rely on permutations, provides a metric solution. It is found that if any class contains a metric, then the other members of the class, if not exactly metric, could be considered near-metric. The metric described with Spearman's footrule is similar to the one created using a *polygonal inequality*, an extension of the standard triangle inequality for distances [19]. By both association and the following two properties, ρ and τ can be considered near-metric and hence all adaptations discussed herein apply:

1. *Metric boundedness property.* There is a metric d' and positive constants c_1 and c_2 such that for all x,y in the domain,
   $c_1 d'(x, y) \leq d(x, y) \leq c_2 d'(x, y)$.

2. *Relaxed polygonal inequality.* There is a constant c such that for all $n > 1$ and $x, z, x_1, ..., x_{n-1}$ in the domain, $d(x,z) \leq c(d(x,x_1) + d(x_1,x_2) + ... + d(x_{n-1},z))$.

Unlike τ and ρ, the footrule is not what Kendall considered a true correlation as one cannot derive it from his general correlation coefficient (Γ), defined as Equation 2.1. Although Kendall [30] does not advocate using Spearman's footrule as a measure of correlation, Fagin et al [19] do so because Equation 2.9 can be comparable to a distance metric. By using rankings as a meaningful metric, Fagin et al [19] create distance equivalence classes. For the purpose of their research, all distance measures in a distance class are essentially the same up to constant factors (that do not depend on the sample size, n) [19]. R-A SVD uses Kendall's τ as the underlying correlation technique. When the correlation and covariance matrices are considered, it is with the underlying understanding that the τ considered is near-metric in nature. For the purposes of this research, establishing τ as near-metric means that the once ordinal data now are considered discrete, where the ranked outcomes range from 1 to $k + 1$ (k +1 being the number used as a placeholder for ties in the data; see Section 2.2.2 for more details). The overall data are in an $n \times m$ matrix, where n will represent the number of elements being considered (note $n \geq k + 1$ and m will be the number of sources). Allowing τ to be considered near-metric establishes the basis for applications that will be done throughout our new method.

2.2.2 Adaptation for Non-Overlapping Elements

When Kendall [30] established his rank methods in the 1930s, it was assumed that there would be two or more sources looking at the exact same elements ranking them from least to greatest. In contemporary times, multiple sources publish rankings of the same (or similar) set of objects. As the populations from which the sample of ranked elements grows, the likelihood that any two lists overlap entirely decreases. Additionally, these

sources only consider a small group, the top k, to rank. For example, there are over 7000 higher educational institutions in the Unites States alone. Yet, released rankings typically rate only the schools that they believe represent the top 100. For this reason, comparing rankings becomes more complex and creates a need to adjust the data before comparisons are made.

Fagin et al [19] also discuss a method for adjusting data for non-overlapping elements. Using the equation for Spearman's footrule 2.9, Fagin et al [19] address the concept of adding a penalty when sources' lists are not identical. Namely, the number $k + 1$ is substituted for all missing data, where k is the number of elements that are being compared in the ranking. In a scenario where only the top 20 from two different rankings are being considered, a ranked element that appears in one list but not the other will receive an arbitrary penalty number in the list from which it is missing (in this case, 21). This is a penalty because each tie added slightly reduces the overall association attainable. The rationale for the penalty is that if the ranking is missing from the top k list, it must have been ranked $k + 1$ or higher. To see the effect of the penalty, Equation 2.10 shows the effect they have on the bounds of the association:

$$F^{(k+1)}(\sigma_1, \sigma_2) = 2(k - z)(k + 1) + \sum_{i \in Z} |\sigma_1(i) - \sigma_2(i)| - \sum_{i \in A} \sigma_1(i) - \sum_{i \in B} \sigma_2(i) \quad (2.10)$$

where Z is the set of overlapping elements, z is the size of Z, k is the number of ranked elements being considered, A represents the elements included solely in the first ranking and B represents the elements solely included in the second ranking. Thus, the lower bound of $F^{(k+1)}(\sigma_1, \sigma_2)$ is

$$F^{(k+1)}(\sigma_1, \sigma_2) \geq (k - z)(k - z + 1) \quad (2.11)$$

This value is normalized by dividing by the maximum value it can attain, which is $k(k + 1)$ [10]. Due to the absolute values of this metric, it will take values between 0 and 1 inclusive when it is normalized, where zero is

indicative of identical ranking.

When Equations 2.10 and 2.11 are dominated by non-overlapping elements, the range decreases but approaches 1 as the number of non-overlap increases. In the case where only half of the data overlaps and in which case $z = k/2$, the normalized lower bound becomes:

$$F^{(k+1)} = \lim_{k\to\infty} \frac{k(\frac{k}{4} + \frac{1}{2})}{k(k+1)} = 0.25$$

Therefore, the value of the $F^{(k+1)}$ will be between 0.25 and 1.

For rankings from multiple sources considered in this research, non-overlapping elements are common. In each simulation and application attempted, $k + 1$ is used as a placeholder. The placeholder assures that each element is maintained when deciding whether an overall association exists.

2.3 R-A SVD Methods for Determining Association among Multiple Sources

PCA is a multivariate procedure that aims to reduce dimensionality of multivariate data while accounting for as much variation of the original data as possible. This technique is especially useful when the variables within the data are highly correlated. Ever [18] states that PCA seeks to transform the original variables to a new set of variables that are:

- Linear combinations of the variables of the original data set
- Uncorrelated with each other
- Ordered according to the amount of variation of the original variables that they explain

Using varying methods involved with PCA, R-A SVD attempts to reduce the sources ranking the data into reasonable agreeing factions, identify when

there are multiple factions uncorrelated with one another, and identify when there are equivalence classes of elements within the factions. As noted earlier, this chapter will discuss only the first step of R-A SVD, which is determining overall association of the data.

Kaiser [29] uses the first eigenvalue of a population of data to identify the overall association of the data. Adaptations to this method are required in cases with small samples that have little or no association. In R-A SVD, the first eigenvalue of the correlation matrix will be used when calculating the correlation in a technique similar to Kaiser's method.

2.3.1 Kaiser's Proposed Method

When assessing whether, to a certain degree, there is an underlying association occurring in a system as a whole, the obvious answer is to determine the mean intercorrelation. Using PCA, the first (largest) eigenvalue captures much of the variability of the initial data structure. The largest eigenvalue of the correlation matrix is approximately a linear function of the average correlation if the original correlations are all positive [22].

Kaiser [29] proposed that with p standardized variables, a scatterplot could be created where the points would form an ellipsoid through which a line could be fit such that the mean squared deviations of the perpendicular distances from the line would be a minimum. Using Hotelling's [27] finding that a line is the first principal axis of the system and that the mean squared projection on the line is equal to the largest eigenvalue, Kaiser proposed that the average correlation can be approximated using this formula:

$$\gamma = \frac{(\lambda - 1)}{(p - 1)}$$

where γ represents the average correlation and λ represents the largest eigenvalue and p is the number of standardized variables. This method works well when all of the correlations are similar and positive. However,

Kaiser did not discuss the use of his eigenvalue procedure for averaging *sample* intercorrelations and all of his equations are expressed in population parameters.

2.3.2 Adaptations to Kaiser's Method

Kaiser's averaging procedure is biased when applied to small samples from populations having correlations near zero, and the procedure does not handle negative values efficiently ([22] and [17]). Dunlap and Graham [17] introduced a modification of the average of the correlations that substantially reduces the bias with slightly smaller standard errors, thus leading to greater efficiency for small correlations. In this modification, all elements of the correlation matrix were increased by adding one prior to finding the eigenvalue. This adjustment assures that no element in the matrix is negative. The first eigenvalue as a measure of association is valid. However, a back transformation is necessary where this new first eigenvalue is reduced by the rank (the number of ones previously added). The remaining steps follow the outline of Kaiser's method where we subtract one from this value and divide by the quantity of the rank minus one. The new average for the correlation is found:

$$r = \frac{(\lambda - m - 1)}{m - 1} \tag{2.12}$$

where r is the average correlation, λ represents the largest eigenvalue found, and m is equivalent to the rank of the matrix. According to Dunlap and Graham [17], the results of their Monte Carlo simulation showed substantial improvement to Kaiser's method and was almost synonymous with the back transformed average Fisher's z.

When looking at correlation matrix with values that are either negative or small, this method is the appropriate adaptation to Kaiser's method. The method gave similar results to Kaiser's method when the correlations are large and positive; therefore, Equation 2.12 will be used throughout this research as we will use the first eigenvalue to assess the overall association

among rankers [17].

In the case of rank data, there is yet another adjustment that needs to be considered. Our new method, $\lambda_{\{R-A\ SVD\}}$, follows the steps of the adapted Kaiser's method using Kendall's τ as underlying correlation measure.

2.4 Establishing the Validity and Need for A New Measure of Correlation

Having reviewed the current methods available, there is evidence that a new measure of overall association among the sources is needed. At the time that Kendall introduced the correlation coefficient, W, all calculations for his methods were done by hand. In Chapter 6 of his book, Kendall Kendall:1948 recognized "the most obvious procedure (to assess association among three or more sources) is to average all the possible values of τ between pairs of observers; however, this approach is tedious when m is large." Since Kendall lacked the mathematical tools to attempt the proposed method, he instead established W. In this section, Kendall's [31] proposed method of averaging all the possible τ values, herein called average τ, will be used as a baseline to compare R-A SVD's measure of correlation, $\lambda_{\{R-A\ SVD\}}$, and Kendall's W. Note that R-A SVD attains the estimate for the overall association of sources from Equation 2.12 although the largest eigenvalue, λ, is created using Kendall's τ as the underlying correlation measure.

Simulations in Section 2.4.1 show the similarities between Kendall's average τ and the first step of R-A SVD $\lambda_{\{R-A\ SVD\}}$. These simulations also identify the conservative nature of these methods as contrasted to Kendall's W.

2.4.1 Our Method: $\lambda_{\{R-A\ SVD\}}$

Having established a valid need for a new method, we have created $\lambda_{\{R-A\ SVD\}}$. The method applies SVD to rank matrix data where the measure of underlying correlation is Kendall's τ. We then apply the adaptation of Kaiser's Method (see Section 2.3.2) to the largest eigenvalue to estimate the average overall correlation of the rank matrix, which we call $\lambda_{\{R-A\ SVD\}}$, see Equation 2.13.

$$\lambda_{\{R-A\ SVD\}} = \frac{(\lambda - m - 1)}{m - 1} \tag{2.13}$$

where $\lambda_{\{R-A\ SVD\}}$ is the average correlation of the set of ranked data, λ represents the largest eigenvalue found using Kendall's τ as the correlation nmeasure but one was added to the correaltions prior to finding the eigenvalues, and m is equivalent to the rank of the matrix. Recall, that in many cases there will be small or negative assoiation among rankers. The adaptation proposed by Dunlap and Graham [17] of adding one to each correlation measure prior to finding the eigenvalues has been applied. This eigenvalue needs to be backtransformed. Hence, we subtract the rank of m from the numerator. Finally, the steps proposed by Kaiser are applied. This measure ($\lambda_{\{R-A\ SVD\}}$) is what we will be using to estimate the overall association throughout the remainder of the paper.

2.4.2 Comparing Average τ and $\lambda_{\{R-A\ SVD\}}$ to W

In order to determine whether the average τ and $\lambda_{\{R-A\ SVD\}}$ are similar and show how they compare to W, the following simulation was created to consider four values: the empirical correlation for the data, average τ, $\lambda_{\{R-A\ SVD\}}$, and Kendall's W. Note that $\lambda_{\{R-A\ SVD\}}$ is equivalent to the value of r found in Equation 2.12, although the λ used is the largest eigenvalue from the correlation matrix of the sources found which use Kendall's τ as the correlation measure. From this same matrix the average τ is found. That is, each off diagonal τ is averaged together (these are the τ measures between

Table 2.4: Comparing Rank Correlation Methods to the Empirical Method

Level of Association	Empirical	$\lambda_{\{R-A\ SVD\}}$	Avg. τ	Kendall's W
Low	0.0047	0.0017	0.0014	0.2515
Moderate	0.6637	0.5258	0.5252	0.8134
High	0.7712	0.6967	0.6963	0.9049

each pair of sources) the average value achieved is referred to as average τ which was proposed by Kendall **??**.

A model was created where $m = 4$ sources denoted as a, b, c, and d, respectively. Each source ranked the same 20 elements. In each case, the sources were created to have a specified overall association: none, moderate or high. The initial 20×4 matrix was created using a continuous distribution. As a comparable measurement for overall association, Pearson's correlation (our Empirical value) was used to analyze the continuous variables. The first column of Table 2.4, labeled *Empirical*, shows this association. Once the data matrix was created, the numbers were changed to ranks. For each column, the lowest number received a value of 1, next lowest received a value of 2, and so forth up to the highest number, which received a value of 20. Finally, the three rank correlation methods for association were calculated: average τ, $\lambda_{\{R-A\ SVD\}}$, and Kendall's W, respectively.

The correlations seen in Table 2.4 demonstrate that a better method than W is necessary. Although conservative, there is evidence that $\lambda_{\{R-A\ SVD\}}$ and average τ both assess whether there is agreement among the rankers with values close to that of the empirical values. In instances where no association exists, the empirical value, average τ, and $\lambda_{\{R-A\ SVD\}}$ are all very close to zero. In contrast, Kendall's W overestimates the association, with a value close to 0.25. It is important to note that the average τ produces similar results (up to three decimal places) to $\lambda_{\{R-A\ SVD\}}$. We simulated these tests thousands of times with similar results, leading to the conclusion that the methods of average τ and $\lambda_{\{R-A\ SVD\}}$ are comparable in addressing overall association among a set of sources.

2.5 Summation

As was demonstrated by the review of classic methods for comparing rank data, there is a need for a more conservative and less biased method of assessing correlation across multiple sources. In examining the data structure and recognizing the advancement of mathematical tools since Kendall's time, we establish a method that assesses overall association and significance. In the initial step of R-A SVD, we have established a correlation measure ($\lambda_{\{R-A\ SVD\}}$) that provides a conservative measure of association and gives results very similar to the average τ proposed by Kendall. In Chapter 3, we will look at the remaining steps of R-A SVD to see that not only does the method assess the correlation when multiple ranked sources are being considered, it also gives insight as to possible factions and equivalent classes.

CHAPTER 3

METHODS INTRODUCED

In the preceding chapter, we established a new method to determine overall association. This method uses Kendall's τ, which is considered to be near-metric (see Sections 2.2.1) and is augmented for ties (see Section 2.2.2), to establish the underlying assumptions necessary for creating the correlation matrix. The largest eigenvalue, λ_1, is used within Kaiser's adapted Equation 2.12 to estimate the overall association among multiple rank sources ([17], [29]). In addition to identifying association, this alternative to Kendall's W can be used to gain insight into the structure of the matrix of rank data. Chapter 3 will focus on discussing the remaining steps of R-A SVD, which require two methods: Anderson's test and hierarchical cluster analysis. When applied to R-A SVD, these respective methods help determine whether the overall association is significant and identify feasible equivalence classes in the data. All of the significant eigenvalues are identified using Anderson's test [9]. From these eigenvalues and corresponding eigenvectors, PCA is used to indicate factions and help identify equivalence classes within these factions. Once established, simulations will be provided to demonstrate how the method works and verify how well it captures the data structure.

3.1 R-A SVD: Methods Required to Identify Faction and Equivalence Classes

As discussed in chapter 2, R-A SVD is similar to SVD but it applies to ordinal data. Overall association is determined using the largest eigenvalue created by $\lambda_{\{R-A\ SVD\}}$ (see Section 2.3 and 2.4 for more details). Eigenvalues obtained from a matrix of a ranked data (P) provide information beyond measurable association. PCA uses all identified significant eigenvalues to evaluate the variability within the data set meaningfully. The remaining steps of R-A SVD rely on the factorization of K, the covariance matrix of P. When considering a matrix of ranked data, P, R-A SVD decomposes K into the linearly independent component, V, and the diagonal of eigenvalues, D. Multiplying P, V, and D^{-1} establishes U^*. Anderson's test [9] identifies the a significantly large eigenvalues from D, allowing for an analysis of the data in fewer dimensions. The first a columns of V indicate which sources agree and, in many cases, when factions agree. U^* identifies when equivalence classes exist using hierarchical cluster analysis and other graphics created using the significant vectors of U^*.

3.1.1 The Diagonal Matrix D

Consider any $n \times m$ matrix X that is distributed normally with a mean vector of zero and a variance array, Σ. The covariance matrix created by taking X^TX will be a positive semi-definite $m \times m$ matrix. SVD factors this matrix into the orthonormal matrix, V, and a matrix of diagonal eigenvalues, D, as seen in Equation 3.1.

$$SVD(X^TX) = VDV^T \tag{3.1}$$

For the purpose of creating R-A SVD, the matrix under consideration is slightly different than the one used for standard SVD. The original rank

matrix, P, is an $n \times m$ matrix, where n is the number of elements being ranked and m is the number of sources that have created a rank. Each column vector is approximately uniform with a minimum value of one and a maximum value of n (often called k). The rank matrix is not normal (see Section 2.1.1). Instead of taking $P^T P$ of the data, the first step is to take the covariance of P using Kendall's τ as the underlying measure of association:

$$K = \text{cov}(P) = VDV^T$$

Recall that Kendall (1938) [30] proved that τ is normal and the covariance matrix is symmetric of order $m \times m$.

The eigenvalues found in D are non-negative and arranged in decreasing order. Once D is established, Anderson's test is applied to D to determine how many significantly large eigenvalues (a) exist, if any.

3.1.2 Anderson's Test

Anderson [9, 38], Bartlett [11] and Lawley [32, 33] found that using the covariance and variance structure of the data (using the assumptions of independent elements and the asymptotic distribution of the data) permits the construction of hypothesis tests and confidence intervals. Within R-A SVD, Anderson's test will be used to determine which eigenvalues from the covariance matrix are significant using his primary hypothesis test though only focusing on the smallest eigenvalues.

For the purpose of our data, we consider N independent elements being ranked by m sources, where the distribution of the m dimensional random variables is $N(\mu, \Sigma)$ and Σ has distinct eigenvalues such that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$$

with the corresponding eigenvectors

$$\alpha_1, \ldots \alpha_m$$

Our sample estimate of Σ will be $\hat{\Sigma}$ with n degrees of freedom, where $n = N - 1$. We will then use l_1 and a_1 to represent the first sample eigenvalue and eigenvector, respectively. Both the population and sample eigenvectors have a directional cosine length of one [35].

Girshick [24, 25] and Anderson [9, 38] demonstrated that the following results hold true as n becomes large:

1. l_i is distributed independently of the elements of its associated vector a_i.

2. $\sqrt{n}(l_i - \lambda_i)$ is normally distributed with mean of zero and variance $2\lambda_i^2$ as n tends to infinity and is independently distributed of the other sample eigenvalues.

3. The elements of $\sqrt{n}(a_i - \alpha_i)$ are distributed according to the multinormal distribution with null mean vector and covariance matrix:

$$\lambda_i \sum_{h \neq i, h=1}^{m} \frac{\lambda_h}{(\lambda_h - \lambda_i)^2} \alpha_h \alpha_h'$$

4. The covariance of the rth element of a_i and the sth element of r_j is

$$-\frac{\lambda_i \lambda_j \alpha_{si} \alpha_{sj}}{n(\lambda_i - \lambda_j)^2} \quad i \neq j$$

In order to determine whether some of the eigenvalues are significantly similar, Anderson [9] created the following null hypothesis based on the covariance matrix:

$$H_0 \colon \lambda_{q+1} = \cdots = \lambda_{q+r}$$

This hypothesis tests whether r of the eigenvalues of Σ are equal. The q larger

and $m - q - r$ smaller are not. For the purpose of our study, we specifically focus on the set of smallest eigenvalues of the covariance matrix (that are feasibly equivalent) instead of working with the eigenvalues in the middle of the data. The goal is to find which small eigenvalues are significantly small enough as to be considered inconsequential. The hypothesis test starts with the smallest two eigenvalues to test for equivalence.

$$H_0 \colon \lambda_m = \lambda_{m-1}$$

If this hypothesis is found to be true, the next one considers the current equivalent eigenvalues and adds the next smallest eigenvalue to be tested.

$$H_0 \colon \lambda_m = \lambda_{m-1} = \lambda_{m-2}$$

This test continues until the hypothesis is rejected and only $m - q$ of the eigenvalues are found to be significantly similar and inconsequential to the data analysis. The alternative hypothesis (H_a) for all of these cases is that at least one of the roots is distinctly different than the others and hence is significantly larger.

The likelihood-ratio criterion leads to the test statistic:

$$\chi^2 = -n \sum \ln l_j + nr \ln \frac{\sum l_j}{r}$$

where $n = N-1$; r is the number of equivalent eigenvalues; and the summation extends over the values of j, where $j = m, m-1, m-2, \ldots, m-q$. To test the null hypothesis, this test statistic has a chi-squared distribution with degrees of freedom $\frac{1}{2}r(r+1) - 1$, where r increases by one each time the hypothesis is run.

The significant eigenvalues found using Anderson's test are the key to the R-A SVD method. These eigenvalues identify the integral parts of the factorization and help define the number of factions and identify potentially agreeing sources within each faction. It also permits partitioning of elements

into feasible equivalence classes.

3.1.3 Orthogonal V Factored from the Covariance Matrix K

As noted previously, R-A SVD initiates by factoring the covariance matrix, K, into an orthonormal matrix, V, and a matrix of diagonal eigenvalues, D. V is orthogonal by definition, meaning that $V^T V = I$, where I is the identity matrix. Specifically, V is the set of orthonormal eigenvectors of the covariance of P.

After the a significantly large eigenvalues from D are identified, the corresponding a columns of V are considered. If there is only one significant eigenvalue, the first eigenvector is used to identify which of the m sources agree as to the rank ordering of the data. When more than one eigenvalue is significant, the columns of eigenvectors can be used to break the sources down into groups of agreeing factions. The number of significantly large eigenvalues assists in determining the possible number of factions that exist. The corresponding eigenvectors identify specifically which groups are in agreement within each faction.

Since orthonormal vectors are normalized to have Euclidean length of 1 by definition [28], the sum of the squared elements of each column of V sums to 1. When some sources agree, those sources achieve values that approach $\sqrt{1/q}$, where q is the number of sources in agreement. The remaining $m - q$ values approach zero. In PCA, the method of determining which sources are important (in our case, agreement) generally looks at the coefficients provided in the column vector space. The term *reasonably large* is used to refer to the variables of interest where the larger the coefficient, the more interest given to a specific element [28]. Therefore, determining the members that are part of a specific principal component is based on approximate estimates without an exact test to prove they are significant.

To be thorough in our diagnosis of important variables when considering a column vector (herein called a component), we also consider the correla-

tions of each variable with the respective significant principal component. Although the correlations of the variables with the principal components often help to interpret the components, they measure only the univariate contribution of an element X to a faction Y [28]. Using these correlations frequently give the same results as simply considering the vector space and sorting the values from greatest to least. Together, these methods help confirm that we have captured the correct variables in our interpretation as to which elements belong within a faction.

In Section 3.2, we create multiple scenarios of factions using simulations. In the interpretation of V for each simulation, we will consider both the reasonably large coefficients from the significant column vectors and also create the correlation of the variables with each significant component. Discussion of both is used to consider which sources belong within each faction.

3.1.4 Back Transformation to Create U^* matrix

When classes of elements exist, U^* allows the user to identify how many classes and also which elements belong in each class. This class system helps the user to create groups of items that are the best, second best, and so forth. Often, these classes are interpreted within a faction.

Matrix multiplication is applied to create U^*. In standard SVD, some $n \times m$ covariance matrix (X) is factored into three components: U, V, and D. SVD can be expressed as a matrix expansion that depends on the rank r of K. There are: r positive constants λ_1, λ_2, ..., λ_r; r orthogonal $n \times 1$ unit vectors $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_r$; and r orthogonal $m \times 1$ unit vectors $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_r$, such that:

$$\mathbf{X} = \Sigma_i^n = \lambda_i \mathbf{u}_i \mathbf{v}_i^{'} = \mathbf{U}_r \lambda_r \mathbf{V}_r^{'} \tag{3.2}$$

where $\mathbf{U}_r = \mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_r$; $\mathbf{V}_r = \mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_r$; and λ_r is an $r \times r$ diagonal matrix with diagonal entries λ_i [28]. Similar to SVD, the V and λ_r values in R-A

SVD are found using the standard factorization. Then, U^* is created via back transformation, specifically, the matrix multiplication of the original matrix of ranks, $P * V * D^{-1}$.

The orthonormal U factored during SVD can be found using a similar back transformation, where $U = X * V * D^{-1}$. The difference between these two methods is that R-A SVD uses Kendall's τ as the underlying correlation method, whereas standard SVD uses Pearson's r. The resulting vectors of U^* are not orthonormal, but the scatterplot pattern for each pair of sources reveals that the data retain a very similar overall shape to the scatterplot pattern created when Pearson's method is applied.. The obvious difference is that values of U^* have more variation between each data point.

In order to conceptualize the differences, a simulation was run. For this example, a fictitious rank data set F was created. F has $n = 20$ elements being ranked by $m = 5$ sources. Once the SVD of F was used to identify the standard U of continuous data, $U^T U$ was multiplied with the expected result I. Next, the R-A SVD of F was applied creating the V and D. (For the purpose of this example, the other steps of R-A SVD were ignored.) Then, the back transformation was applied to identify U^*, where $U^* = F * VD^{-1}$. Multiplying $U^{*T} U^*$ did not produce I, as this as this matrix is not orthonormal since the data are discrete near-metric and the correlation measure is τ.

A scatterplot matrix was created (see Figure 3.1). In this matrix, each column of U was mapped to every other column of U. The corresponding xy plots can be seen in the northwest corner of the scatterplot matrix. Similarly, each column of U^* was mapped to the remaining columns of U^*. These xy plots are located in the southeast corner of the scatterplot matrix. In Figure 3.1, the northwest plots show the shape and dispersion among each pair of sources when SVD is applied. The southeast plots show the shape and dispersion using R-A SVD. The shapes for the matching pairs of sources are similar. As discussed above, the looser dispersion seen in the southeast corner indicate that R-A SVD has higher overall variability because the method being applied focuses on the data being ordinal in nature.

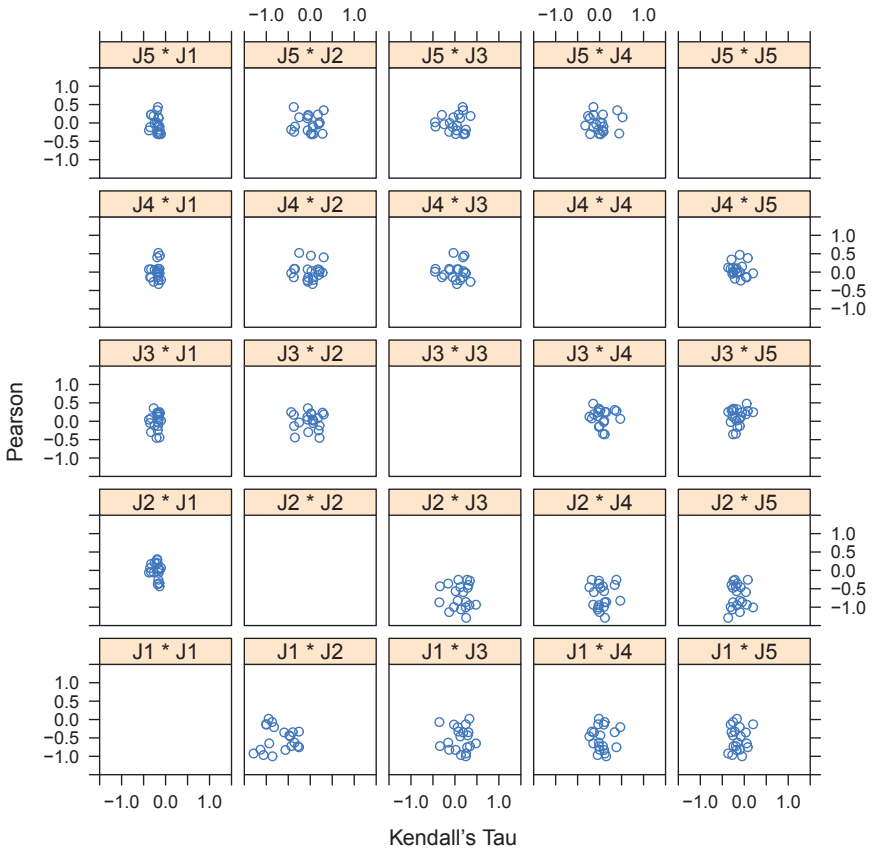

Figure 3.1: Scatterplot matrix where the northwest corner shows the U vector found using singular value decomposition (SVD) with Pearson's r as the correlation measure, while the southeast corner shows the U^* found for the same data using rank-adapted singular value decomposition (R-A SVD) with Kendall' τ as the correlation measure.

3.1.5 Hierarchical Clustering

Once the significant U^* vectors are identified, hierarchical cluster analysis will be applied to the a significant vectors of U^* to determine whether equivalence classes exist among the elements and which elements belong within each class. A hierarchical clustering of n data points is a recursive partitioning of the data into $2, 3, 4, \ldots$ and finally n, clusters. Each intermediate clustering is made more fine-grained by dividing one of its clusters [16]. Unlike other known clustering methods, *k-means* and medoids, it does not require knowledge of the number of clusters prior to partitioning. According to Dasgupta [16], these hierarchical representations of data permit the data to be understood simultaneously at many levels of granularity, and there are some simple, greedy heuristics that can be used to construct them.

For the purpose of identifying equivalence classes using the R-A SVD method, the main hierarchical cluster graphic considered is the dendrogram. The data that the dendrogram considers depend on the number of significant eigenvalues. Only the a significant corresponding eigenvectors of U^* will be considered when creating a dendrogram in which the U^* matrix is designed to estimate distances as established by the correlation metric τ. The y-axis of the dendrogram represents heights, which in turn represent the distances of each element's correlation with respect to the remaining elements. Very closely-spaced elements will resemble a set of very close branches at the end of a tree. To *prune* the tree, the height at which clear partitions of meaningful subpopulations (referred to as equivalence classes) appear is selected.

Although the dendrogram allows for clear decision-making when partitioning classes, other graphics that we introduce (where the dimension of the graphic depends on the number of significant eigenvalues found) provide additional insight into the data. These graphics, which use the dendrogram as their basis for partitioning data, illustrate two things that the dendrograms cannot. The values within the U^* matrix create distances between the elements that reflect the correlation information from the struc-

ture of the data. The first issue that our graphics address is that distances along the $x-$axis of the dendrogram are spaced evenly (not reflecting meaningful distances), whereas our graphics will allow for those distances to be captured when looking at the elements in each class and among separate classes. Second, the dendrogram does not capture the correct ordering of the classes as to best, second best, and so forth. The graphics created (using the partitioning from the dendrogram) will provide visual information for the proper ordering of the classes, especially when more than 2 classes exist. The two-dimensional graphics will maintain the structure of the originally ranked data as the classes are identified. When referring to the initial data, one can then determine if the best class is on the bottom with the remaining classes classes ascending or at the top with the remaining classes descending. However, the ordering remains constant and in one direction, unlike the dendrogram where the classes have no meaningful order.

To summarize, these graphics identify which elements belong in each equivalence class. Starting with hierarchical clustering (using dendograms as the main tool), the partitioning of the classes will be established. Then, other graphics created from the significant vectors of U^* will be color coded to reveal the same classes seen in the dendrogram. Using xy plots and three-dimensional plots (where appropriate), the number of classes and the elements belonging within each class will be identified and the distances among elements and classes will be presented. Finally, xy plots and three-dimensional plots (where appropriate) provide the means by which the classes can be ordered.

3.2 Simulations

Different scenarios were simulated to illustrate the insights gained using R-A SVD. For the sources evaluated, examples show all sources agreeing, instances where only some factions of sources agree and instances where the

ranking is purely random. For every simulation discussed in this section, the data will consist of $n = 20$ elements being ranked by $m = 5$ sources. Since no elements are missing, placeholders are not added. For each simulation, the following R-A SVD steps were performed:

1. The overall association is measured first using both $\lambda_{\{R-A\ SVD\}}$ and W.
2. The number of significant eigenvalues, a, is identified using Anderson's test.
3. The significant eigenvectors V corresponding with the a significant eigenvalues are considered to determine which sources are in agreement and, when applicable, which belong within each faction.

In many cases, equivalence classes are simulated. The final step fo R-A SVD will use U^*, hierarchical cluster analysis, and some graphics based on the significant vectors of U^* to identify these equivalence classes. Overall, these classes will be graphed, identified, and discussed.

3.2.1 Five Sources: All Agree

In this simulation, the data were created so that all five of the sources agree for the most part as to the ranking of the objects. Using random numbers, the data were simulated so that in most cases all five sources agree that the first five elements are the top five with varied ordering, thus creating an equivalence class. This process is repeated with elements 6 through 10, 11 through 15, and 16 through 20, creating four total relative equivalence classes that each contain five elements. In order to ensure that the correct number of eigenvalues was found, the simulation was run 1000 times each at levels of $\alpha = 0.10$, 0.05, and 0.01, respectively. In almost all simulations, there was only one significant eigenvalue, which is seen in Table 3.1

We now consider one replication to see what other information is gained by using the R-A SVD method. Figure 3.2 demonstrates all five sources

Table 3.1: 1000 Simulations Created Where All Sources Agree

	Number of Significant Eigenvalues					
α	0	1	2	3	4	5
.10	0	993	2	0	5	0
.05	0	1000	0	0	0	0
.01	0	999	1	0	0	0

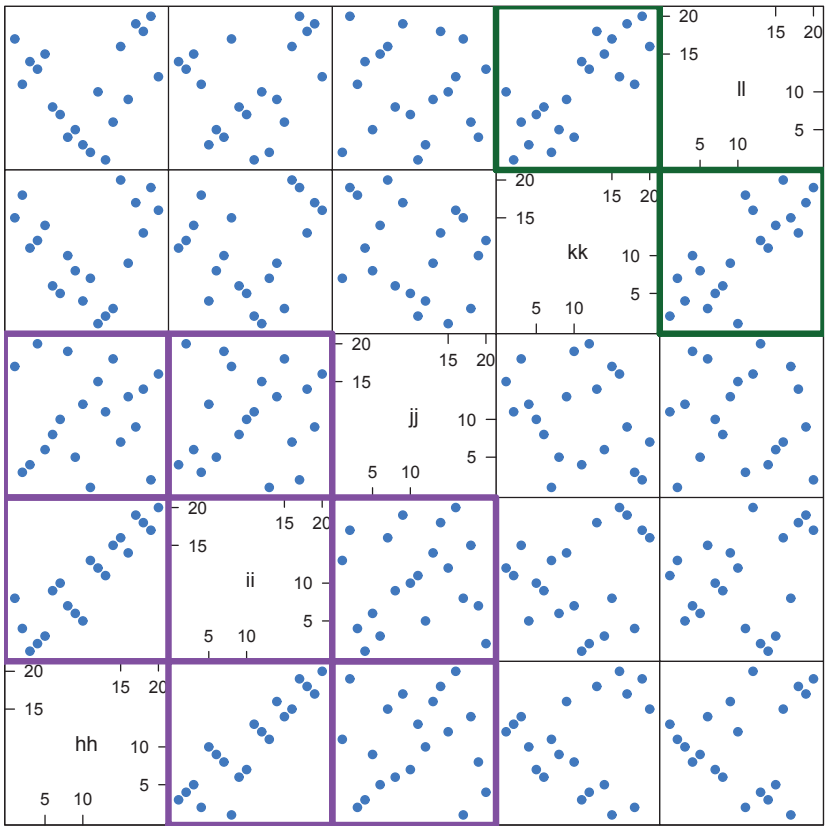

Figure 3.2: Scatterplot matrix demonstrating that all five sources agree as indicated by all individual plots having the blue dots falling along the diagonal starting at the point (0,0) and heading toward the point (26,26).

Table 3.2: Eigenvector which Corresponds to the Largest Eigenvalue from the Covariance Matrix When All Five Sources Agree

Source Number	Vector 1
1	−0.4535370
2	−0.4591496
3	−0.4557566
4	−0.4664990
5	−0.3976813
Eigenvalue	1484.57305

agreeing on the direction and association of the overall ranking, which is evidenced by the relatively linear set of correlated dots with increasing slope in each box of the scatterplot matrix. Using $\lambda_{\{R-A\ SVD\}}$, the overall association is 0.72. In contrast, Kendall's W assesses the association at 0.859. When Anderson's test is performed, there is one significant eigenvalue from D identified at $\alpha = 0.05$. Since only one eigenvalue is significant, the first vector (V_1) is used to identify which sources agree; these values are listed in Table 3.2. Notice that the absolute values of V_1 all approach $\frac{1}{\sqrt{5}} = 0.447213$, which is evidence that all five sources belong in a single faction and that they agree on strength and direction of the rankings. The sign merely indicates that they all agree on directionality of the ranking and otherwise is irrelevant.

The second method used to identify the relationship among the sources requires an examination of the correlations of each variable with any significant principal components. In this case, there is only one principal component identified by V_1. For the purpose of correlation values, all values below 0.40 are considered uncorrelated, all values that range from 0.40 to 0.60 will be considered moderately correlated, and all values that exceed 0.60 will be considered strongly correlated within the principal component. In Table 3.3, there is evidence that all five sources strongly agree and have the same directionality. Using the first vector as the principal component identifies the sources in the faction. By using the correlations of each source

Table 3.3: Correlations of Five Sources With the First Significant Principal Component

Source	Component 1
1	−0.87911
2	−0.90930
3	−0.89922
4	−0.86031
5	−0.87085

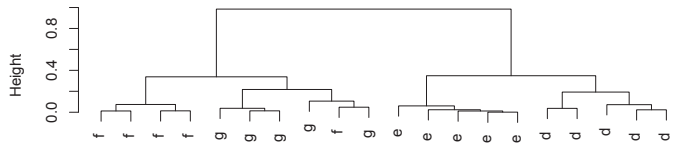

Figure 3.3: Dendrogram with four simulated equivalence classes. Note that one f value is misclassified with the g class. Key: d = elements 1 through 5; e = elements 6 through 10; f = elements 11 through 15; g = elements 16 through 20.

to this component, we confirm that the correct variables (in this case, all five sources) were captured in our interpretation.

When attempting to identify equivalence classes, the goal (similar to cluster analysis) is to find the elements with the minimum distance from one another so that they create meaningful classes with minimal variance between elements in a class. Two figures (Figures 3.3 and 3.4) are used to exemplify these classes. The initial view considered is the dendrogram (Figures 3.3) created from the first vector of U^*. Recall that the y-axis represents how close together observations were when they were merged into classes. In order to identify the equivalence classes, we aim to find the longer branches (in terms of the distances) of the classes as seen on the y-axis. Using a distance indicated at the height of 0.3, there is evidence of four classes. Of the 20 elements, 19 fall within the proper classes.

Similar information can be observed when we examine Figure 3.4. Using

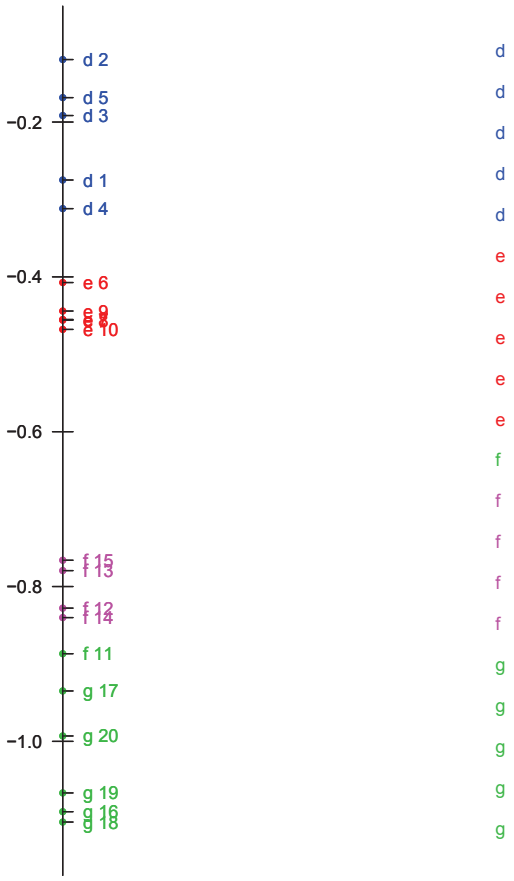

Figure 3.4: The xy plot of U^* to itself. On the right side, the distances along the y-axis represent the partitioning of the classes as identified by the dendrogram (see Figure 3.3), while the colors show the class where the elements belong. The Figure on the left represents the actual positioning of the dots which leads to overprinting, hence we reprint the names in the same order in the second column for clarity of the actual order. Note that one f value is misclassified with the g class. The letters to the far right are the initial ordering of the points according to the first vector of U^*. Key: d = ranked elements 1 through 5; e = ranked elements 6 through 10; f = ranked elements 11 through 15; g = ranked elements 16 through 20. Color code key: d = dark blue; e = red; f = magenta, g = green.

the classes identified by the dendrogram (Figure 3.3) and an xy plot created by graphing U_1^* mapped to itself (Figure 3.4), there is evidence that these four classes are again very closely identified. Because there is only one dimension that we are considering, mapping U_1^* to itself makes the graphic applied very basic. Figure 3.4 allows us to see the distances among the elements and classes (where shorter distances reflect stronger correlation), while identifying the elements within each class as well as the location of each class as far as best at the top, second best being the next class down, and so forth. Note that there is one f value that has been misclassified within the g class in both graphics.

In summation, this simulation presents the perfect situation where all sources agree as to the rankings and the sets of equivalence classes are easily identified. Using R-A SVD and some basic analysis and graphics, the sources creating a single faction and the four simulated equivalence classes are identified clearly and the structure of the initial data is maintained.

3.2.2 Five Sources with Two Factions of Three and Two, Respectively

For this simulation, the data were created so that the first three sources agree strongly with one another and the remaining two sources also agree strongly with one another but have no real relationship with the first three sources. Again, the elements were partitioned in groups of five similar to the previous simulation. In order to ensure the correct number of eigenvalues were identified, the simulation was run 1000 times each at α levels of 0.10, 0.05, and 0.01, respectively. When the α level was 0.10, many of the simulations had two significant eigenvalues. At the α level of 0.05, there was some variation with many simulations identifying two significant eigenvalues. As the level of α decreased to 0.01, the number of significant eigenvalues fluctuated greatly, which can be seen on Table 3.4.

We consider a single replication in order to identify the factions and

Table 3.4: 1000 Simulations Created Where Two Factions Exist

	Number of Significant Eigenvalues					
α	0	1	2	3	4	5
0.10	0	24	964	0	12	0
0.05	0	119	881	0	0	0
0.01	18	412	569	0	1	0

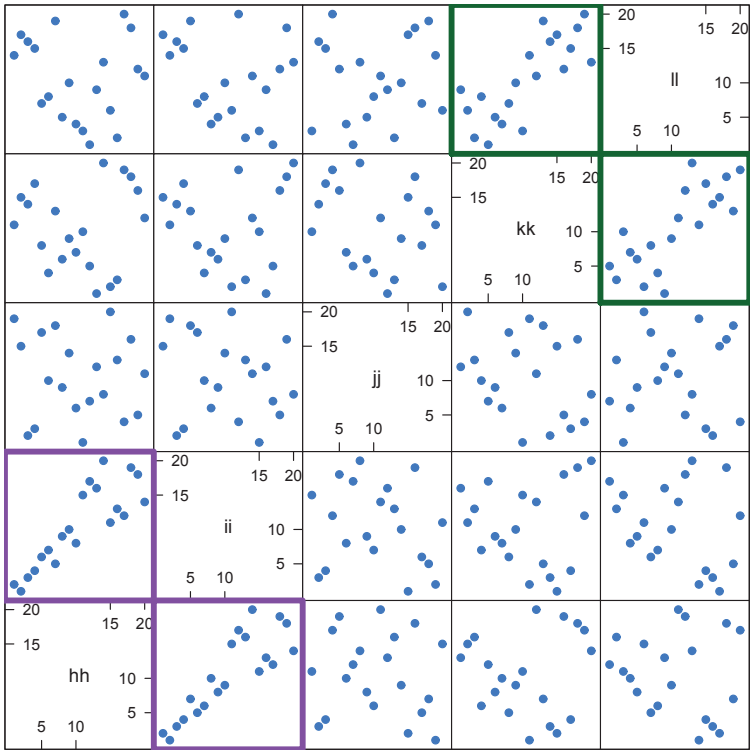

Figure 3.5: Scatterplot matrix of simulated data where the data were created to demonstrate a scenario where two separate factions exist. The scatterplot has two regions, a region with purple outlined boxes which shows the agreement among the first three sources as a relative straight line from the (0,0) location heading to the (26,26) location in each box. The remaining two sources are outlined in green, which shows the agreement of the fourth and fifth source as they also form a relative line along the diagonal from (0,0) towards (26,26). The remaining boxes show random noise and lack of association.

equivalence classes. For this replication, two of the five eigenvalues from the covariance matrix were significant at $\alpha = 0.05$ and 0.10 using Anderson's test. The overall association for $\lambda_{\{R-A\ SVD\}}$ is 0.5132 and is 0.46 for W, which is a rare instance where W is lower than our measure of association is possibly masking the factions as Kendall was concerned it might. That is, Kendall believed that solely relying on W for assessing overall association might mask factions within the data. He describes the following situation. [31]

> to take the extreme case, suppose ten observers are in complete agreement about a ranking of six (elements) giving W a value of 1. Suppose that 10 further observers are also in complete agreement, rank the (elements) in the inverse order but also having a W value of 1. The sum of the rankings of the 20 observers would produce a value of $W = 0$. In this case, one might conclude that no association exists among the rank sources whereas the reality is that one set of observers rankings has masked that of the other.

As illustrated in this simulation (see the original data in Figure 3.5), R-A SVD reveals true significance in cases where the association might be masked when using W. This association is identified despite the low value of the overall association.

Using Anderson's test, two eigenvalues are found to be significant at $\alpha = 0.05$ and 0.10. Next, the corresponding vectors of V are considered. The agreeing sources within each faction are found in Table 3.5. The column labeled vector 1 provides evidence that the first three sources strongly agree about both direction and strength of the association, because each value approaches $\frac{1}{\sqrt{3}} = 0.57773$. The fourth and fifth sources in vector 1 show minimal and no relationship, respectively. Similarly, the column labeled vector 2 provides evidence that the fourth and fifth sources very strongly agree on direction and association of the rankings because they approach $\frac{1}{\sqrt{2}} = 0.707$. Thus, the sources create a second faction. Vector 2 also indicates that the first faction has no agreement with the second faction as the other three values approach zero.

Table 3.5: Eigenvectors Corresponding to the Largest Eigenvalue from Covariance Matrix When All Five Sources Agree

Source Number	Vector 1	Vector 2
1	−0.56759605	−0.02412028
2	−0.57218534	−0.11825882
3	−0.57466340	−0.05092063
4	0.14081556	−0.68781696
5	0.01927742	−0.71396637
Eigenvalues	923.31	605.14

Table 3.6: Correlations of Sources with Each of the First Two Principal Components Which Are Significant

Source	Component 1	Component 2
1	−0.89292	0.03011
2	−0.91368	0.03619
3	−0.87897	0.13785
4	−0.02136	0.87549
5	0.23828	0.83896

Next, we consider the correlations of the sources within each principal component. In Table 3.6, the column labeled component 1 further strengthens the claim that the first three sources agree strongly and share directionality (all values exceed 0.60), thus creating the first faction (the remaining sources in component 1 are less than 0.40). The second faction also is identified clearly when considering the column for component 2. Again, the signs of the variables are only important for establishing agreement in directionality. Recall that for the purpose of correlation values, all values below 0.40 indicate no correlation and all values that range from 0.40 to 0.60 will be considered moderately correlated. Values that exceed 0.60 will be considered strongly correlated within the principal component.

In creating these simulated data, classes of five elements (1 through 5, 6 through 10, 11 through 15, and 16 through 20 were labeled d, e, f, and g, respectively) were placed once again within the data using a similar technique to the initial simulation. The first three sources agree on the top

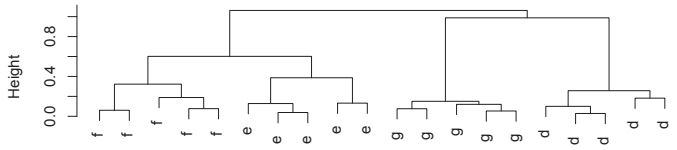

Figure 3.6: The dendrogram of the two vectors considered from U^*. The four classes are identified using a height between 0.40 and 0.60 on the y-axis. Key: d = ranked elements 1 through 5; e = ranked elements 6 through 10; f = ranked elements 11 through 15; g = ranked elements 16 through 20.

five, although in varying order. Similarly, the last two sources agree that these five elements are in a class, although not specifically in the top five. Because there are two significant eigenvalues, two vectors of U^* are used when creating a dendrogram to identify the equivalence classes. Figure 3.6 provides evidence of partitioning the data into four classes where all four classes are easily identified using a height on the y-axis of between 0.40 and 0.60. Here, all of the letters fall within their expected equivalence class.

Finally, the first two vectors of U^* are mapped to one another in Figure 3.7, allowing for a two-dimensional view that creates a clearer image of the partitioning of these classes. Using color coding from the classes identified in the dendrogram, it is evident that each class was identified correctly. All of the elements fall identifiably into the correct class.

Overall, the R-A SVD method successfully identified the initial structure of the data by dividing the sources into two factions that agreed internally but not with the opposing faction. It then correctly identified the sources and classes, thus establishing that this method works well when using clean data.

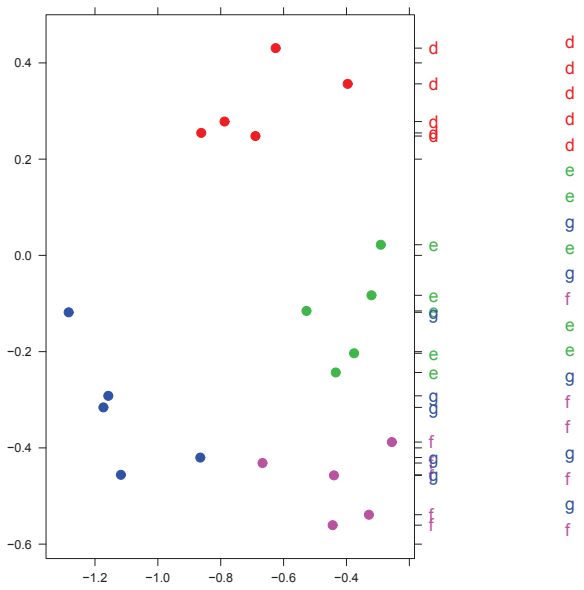

Figure 3.7: The *xy* plot of the second vector of U^* mapped to the first. There are four classes (d, e, f, g) with five elements each, as identified by the plot of the U^* vector in two dimensions. Color code key: d = dark blue; e = red; f = green; and g = light blue.

Table 3.7: 1000 Simulations Created Where Two Factions Exist as Well as a Source of Noise

	Number of Significant Eigenvalues					
α	0	1	2	3	4	5
0.10	280	47	432	221	16	4
0.05	510	44	354	100	12	0
0.01	869	30	91	9	1	0

3.2.3 Five Sources Split into Factions of Two and Two with a Source of Random Noise

In this simulation, we consider the realistic case in which there might be some sources in agreement as well as a source that is completely random. To simulate the data, the first two sources agree fairly strongly on the $n = 20$ elements being ranked, the third source ranks the 20 elements completely randomly, and the remaining two sources agree with one another but have little association with the first faction. Figure 3.8 provides evidence of these two sets of agreeing factions. It can be observed that both the first two sources agree (outlined in purple) and the last two sources agree (outlined in green). The rest of the boxes all demonstrate relative scatter as to the points within the boxes. This pattern indicates that there is evidence of two, one, and two sources in agreement, respectively, which corresponds with how the data were created. In order to be certain that the tests would identify these relationships, 1000 simulations were performed at α levels of 0.10, 0.05, 0.01, which can be seen in Table 3.7. These results indicate two factions are typically found when the α level is 0.10. However, the number of significant eigenvalues varied depending on the level of α. Namely, zero, two, and three significant eigenvalues were present with high probabilities.

Although these data are messier in nature, they allow for initial inferences for identifying significance at $\alpha = 0.10$, with some evidence of identifying the significance at an α level of 0.05 as well. Unlike W, anytime significance is found, regardless of the correlation value, R-A SVD provides further infor-

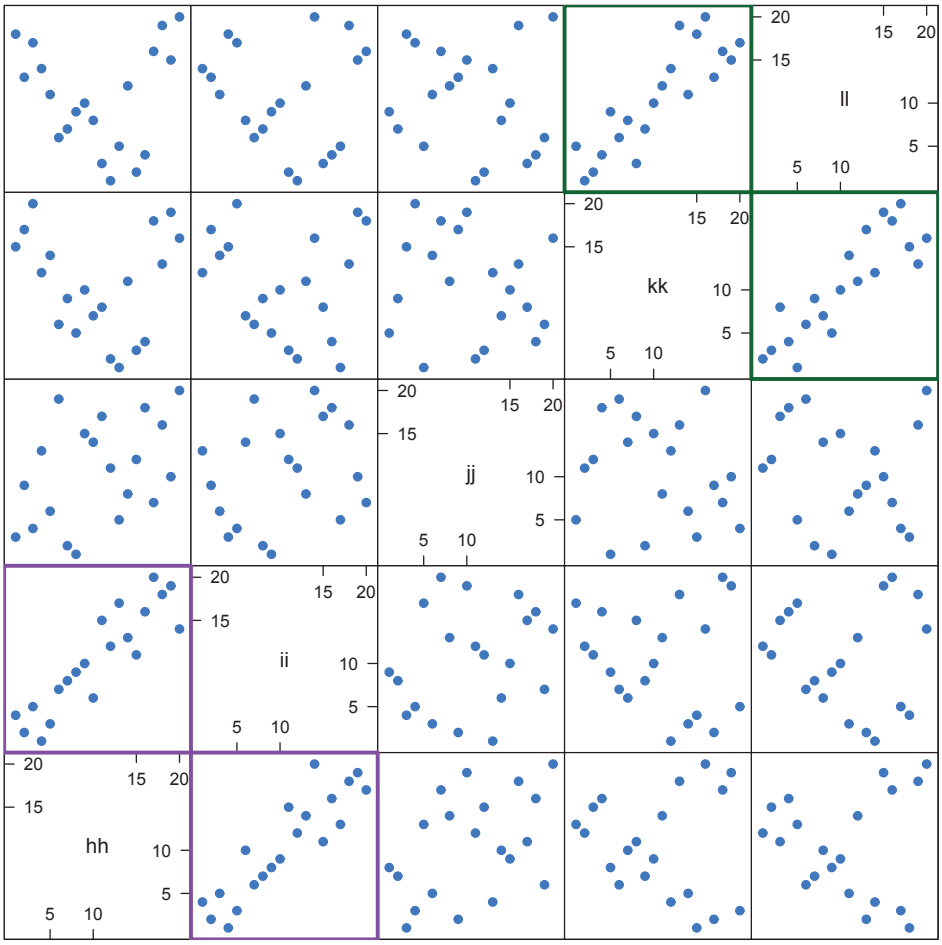

Figure 3.8: Scatterplot matrix of five sources ranking twenty items where there are two factions of two sources that agree within each faction (highlighted in purple and green, respectively) but have no association between factions. There is also a fifth source which is just random noise.

mation about the structure of the data set by investigating the corresponding vectors.

In running this simulation, the most common number of eigenvalues found to be significant were split between zero or two with the possibility of three eigenvalues. In order to analyze the meaning of the output, we will consider a single replication from each significance scenario (zero, two, and three).

Zero Significant Eigenvalues

A single replication was run with the output revealing no significant eigenvalues at an α level of 0.10, even though two significant eigenvalues could be found at an α level of 0.13. Since the cut off is 0.10, no eigenvalues are significant, indicating that $\lambda_{\{R-A\ SVD\}}$ of 0.0476 is insignificant. W seems to overestimate the association with a value of 0.238. When no significance is found, no other steps of R-A SAVD are performed.

Two Significant Eigenvalues

A different replication was performed with the output revealing two significant eigenvalues at both $\alpha = 0.05$ and 0.10. In this case, $\lambda_{\{R-A\ SVD\}}$ is 0.1890 and W is 0.421. Since, two significant eigenvalues are found, deeper analysis uncovers information about the structure of the data.

Initially, the first two vectors of V are considered to identify factions as well as to indicate which sources belong within each faction. In Table 3.8, sources 1 and 2 in the column labeled vector 1 strongly agree about the direction and association with values of 0.65513 and 0.63904, respectively. This association approaches $\frac{1}{\sqrt{2}} = 0.707$, but there is also a moderate amount of association evident from the third source, the one created randomly, which has a value of 0.39402. However, the last two sources supply no evidence of association with any other measure within vector 1 as they approach zero. Looking at the column labeled vector 2, there is evidence that the fourth

Table 3.8: Corresponding Eigenvectors of V

Source Number	Vector 1	Vector 2
1	0.65513	0.02970
2	0.63904	−0.01692
3	0.39402	0.053920
4	−0.077327	0.70299
5	0.03454	0.708323
Eigenvalues	821.63309	554.04672

Table 3.9: Correlations of Sources with Each of the First Two Principal Components

Source	Component 1	Component 2
1	−0.89715	0.03809
2	−0.87511	−0.02169
3	−0.53958	0.069154
4	−0.10589	0.90162
5	−0.04730	0.90846

and fifth sources very strongly agree on direction and association of the rankings with values of 0.70299 and 0.708323, respectively, and thus create a second faction. These values also approach $\frac{1}{\sqrt{2}} = 0.707$. In contrast, all of the other values in vector 2 approach zero, indicating that the first faction has no agreement with the second faction.

Next, we consider the correlations of the sources within each principal component. In Table 3.9 , the column for component 1 further strengthens the claim that the first two sources are in agreement in strength and directionality with values that far exceed a correlation of 0.60 (thus creating the first faction), but the third source again gives some indication of association though the value has a moderate strength falling between 0.40 and 0.60. The second faction consisting of source 4 and source 5 are identified clearly in component 2 as strongly correlated to each other, with no other sources revealing any correlation.

From Figure 3.8 there is evidence that only the first two sources are in

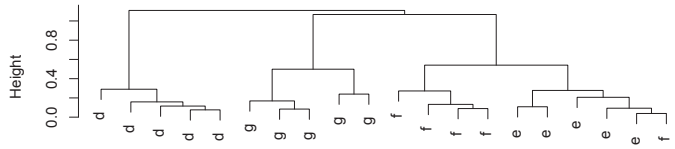

Figure 3.9: Dendrogram of the corresponding two significant vectors of U^*. Although four classes were created (using a height on the y-axis = 0.50) these classes are maintained with one element from f falling into the e equivalence class.

agreement, although the values in V_1 indicate that the third source (the source created with random noise) might sometimes fall within a faction for which it is not intended. The strength of this agreement is weaker, but the evidence suggests keeping this possibility in mind when considering the first faction. Both Tables 3.8 and 3.9 indicate that the second faction consists of only sources 4 and 5.

Next, consideration is given to the two corresponding vectors of U^* to identify whether the elements within the two factions provide evidence of equivalence classes. Similar to previous simulations, the initial agreeing sources (the first two sources) were each created in sets of five elements (items ranked 1 through 5, 6 through 10, 11 through 15, and 16 through 20 were labeled d, e, f, and g, respectively). While source three was completely random, the remaining two sources also considered the same four classes, although ranking them with different orders. When the dendrogram is reviewed (Figure 3.9), it reveals (using a height between 0.40 and 0.55 on the y-axis) that all four classes are identified. However, one f value falls within the e class.

Then, the first U^* vector is mapped to the second U^* to determine whether the underlying structure is maintained. When examining Figure 3.10, the xy plot provides evidence of the initial structure and shows similar partitioning but with more meaningful distance as the elements are partitioned along

Table 3.10: Corresponding Eigenvectors of V

Source Number	Vector 1	Vector 2	Vector 3
1	−0.57170	0.40195	0.10228
2	−0.48247	0.52516	−0.05709
3	−0.04401	0.00936	−0.99040
4	0.40908	0.58558	0.04386
5	0.52068	0.46867	−0.05877
Eigenvalue	745.2786	589.4765	349.5785

both the x-axis and y-axis. That is, if one were to only consider the labeling of the elements on the right side, the actual separation of the equivalence classes is evidenced by the distinct separation of the blue class on the bottom left from the pink and green elements on the bottom right. Although one f does fall into the e class, most of the initial data structure is maintained.

Three Significant Eigenvalues

The final replication of this section is considered with the output revealing three significant eigenvalues at both $\alpha = 0.05$ and 0.10. In this case, $\lambda_{\{R-A\ SVD\}}$ is 0.12149 and W is 0.326. Although three significant eigenvalues were found, it is notable that the evidence supplies similar results to the ones uncovered with the previous simulation with two significant eigenvalues even though the numbers seem very different.

Initially, the first three vectors of V are considered to identify factions as well as which sources belong within each faction. In Table 3.10, sources 1 and 2 in the column labeled vector 1 have very strong agreement about direction and association with values of −0.57170 and −0.48247, respectively, while sources 4 and 5 show strong agreement with one another as evidenced by the values of 0.40908 and 0.52068, respectively. Notably, sources 4 and 5 have opposing directionality to sources 1 and 2. In this case, the sign is important because it indicates that those with the same signs agree with one another while those with opposing signs are in disagreement. Since there

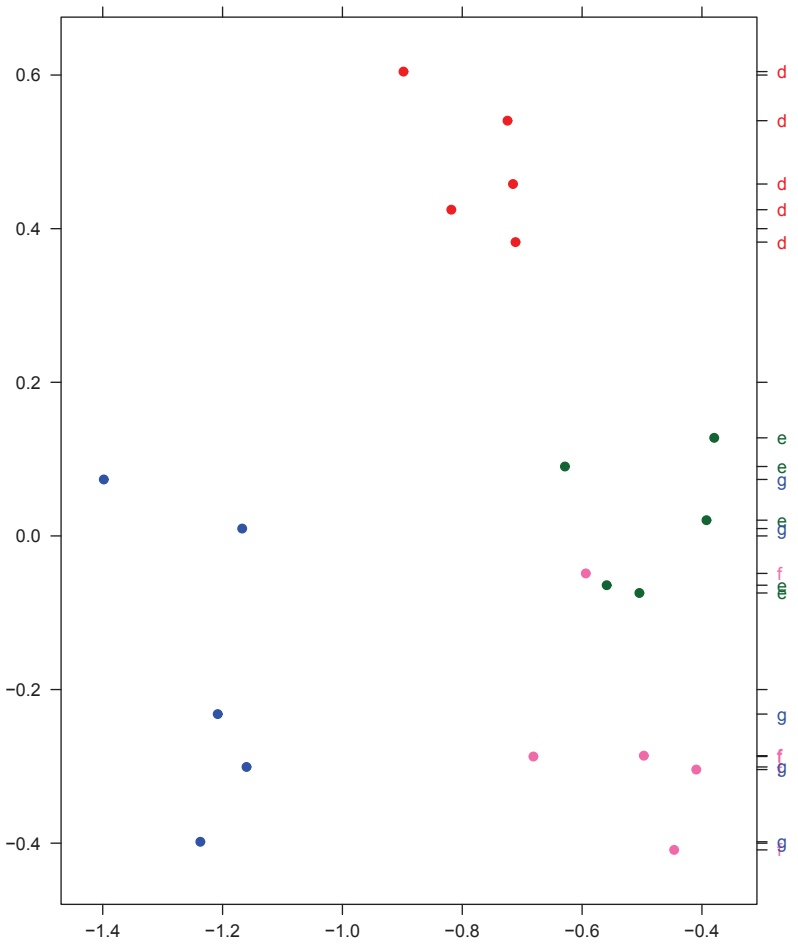

Figure 3.10: The xy plot of the second vector of U^* mapped to the first. Four classes were originally created. All four cases (d, e, f, g) are identifiable but possible overlap is evidenced as to the misclassification of the dendrogram (see Figure 3.9) for the one f that falls into the e class. Notice that using both the x and y distances helps identify partitioning of the equivalence classes. Color code key: d = red; e = green; f = hot pink; g = blue

Table 3.11: Correlations of Sources with Each of the First Two Principal Components

Source	Component 1	Component 2	Component 3
1	0.7834	−0.5089	0.1028
2	0.6611	−0.6649	−0.5074
3	0.0603	−0.0119	−0.9959
4	−0.5606	−0.7414	0.0441
5	−0.7135	−0.5934	−0.0591

are four sources being considered, the association levels approach $\frac{1}{\sqrt{4}} = 0.50$, while the third source approaches zero. All of this information is gleaned from the first vector. The column labeled vector 2 reveals that source 3 has no commonality with the other four sources, although some associations (in the column all agreeing) exist among sources 1, 2, 4, and 5. The column labeled vector 3 continues to highlight this trend, as source 3 approaches a value of 1 while all the other sources approach zero. To summarize, vectors 2 and 3 reveal the same information from opposing views, while vector 1 supports the view that there are two sets of factions that have strong relationships within their faction and opposing relationships between the two factions.

Next, we consider the correlations of the sources within each principal component. Looking at the column for component 1 in Table 3.11 provides a similar viewpoint as the one found looking at vector 1 in Table 3.10. The column for component 1 shows that there are two small factions that are in agreement. The first faction is sources 1 and 2, the second faction is 4 and 5. Disagreement among the factions is seen when observing that the signs of the first two sources are positive while the signs of the last two sources are negative. The second and third components show that there is some association among sources 1, 2, 4, and 5, while source 3 does not have any relationship with any other source.

Then, three corresponding vectors of U^* are reviewed to determine whether the elements within the two factions (with one random noise source) provide evidence of equivalence classes. Similar to previous simulations,

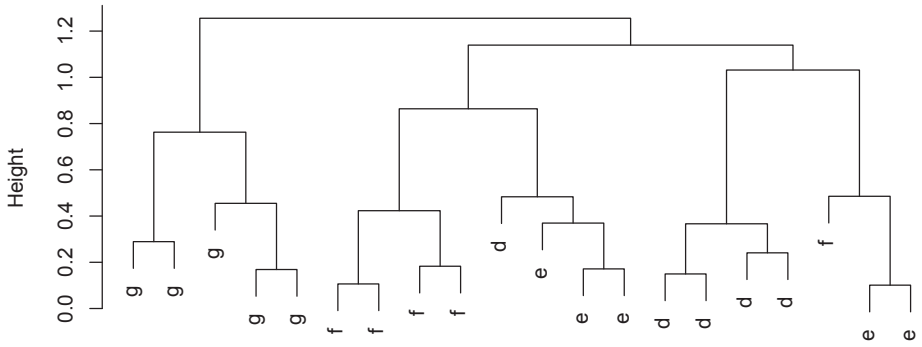

Figure 3.11: Dendrogram of the three significant vectors of U^*. Four classes were originally created. The initial structure is not well maintained. Using a height between 0.85 and 1.05 on the y-axis, two of the classes are identified (letters g and f), but the other two classes overlap excessively. Key: d = ranked elements 1 through 5; e = ranked elements 6 through 10; f = ranked elements 11 through 15; g = ranked elements 16 through 20.

the initial agreeing sources were each created in sets of five elements. For the first two sources, the elements 1 through 5, 6 through 10, 11 through 15, and 16 through 20 were labeled d, e, f and g, respectively. Although the ranking provided by source three was completely random, the remaining two sources also considered the same four classes, albeit ranking them with different orders. Initially, the dendrogram is considered using the three corresponding U^* vectors. Then, each of the first three U^* vectors is mapped to one another to see if any information can be gleaned as to the underlying structure.

In the dendrogram (Figure 3.11), there is evidence of the partitioning of the elements into four classes (y heights between 0.85 and 1.05). All of the elements for g and four of the elements are from f fall within their proper class. The other two classes provide evidence of noise (each class consists of a mix of d and e elements with a single f thrown in). The noise factions are created of one small faction (3 elements) and one very large faction.

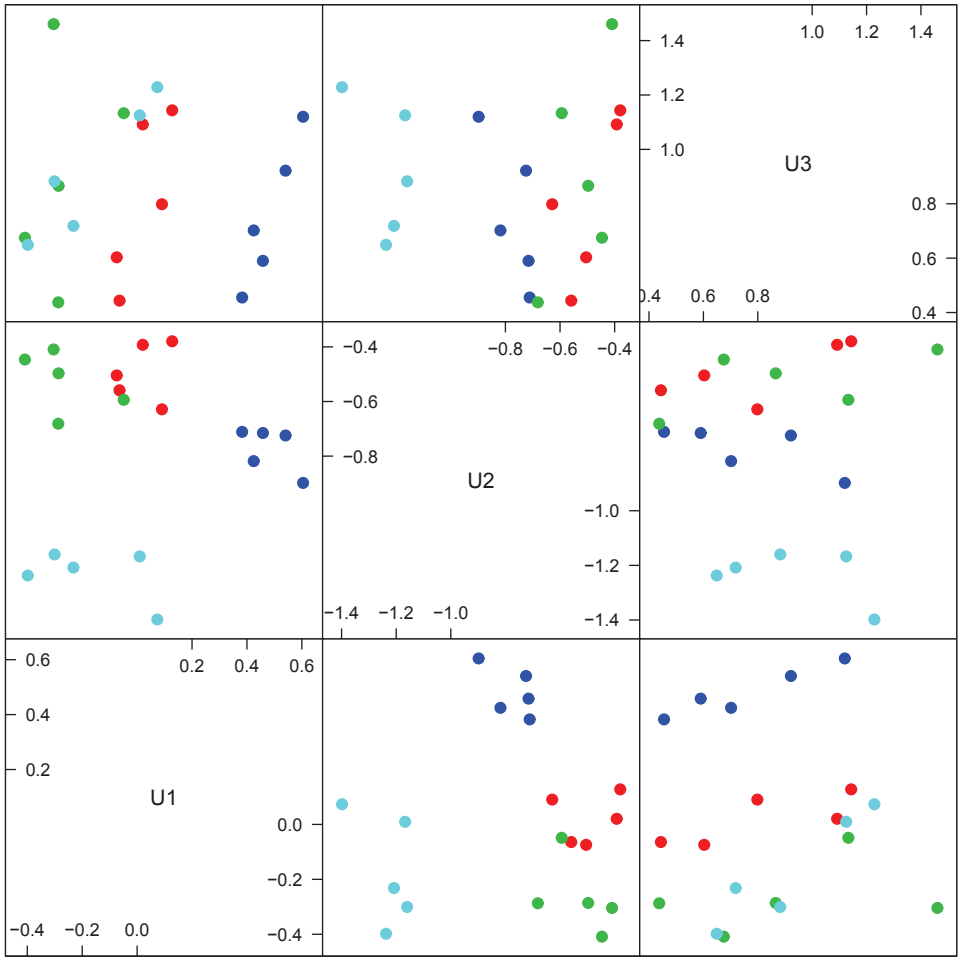

Figure 3.12: Scatterplot matrix of the first three vectors of U^* mapped to one another to provide insight as to equivalence classes. Four classes (d, e, f, g) were originally created. Color code key: d = blue; e = red; f = green; h = light blue.

On the xy plot (Figure 3.12), U_1 mapped to U_2 reveals a clean view of the proper partitioning of the data. However, when U_1 is mapped to U_3 and U_2 is mapped to U_3, there is mostly noise with one or two classes possible class separating out in each box. In reality, the third dimension seems mostly to identify the noise and is less able to identify clear partitioning of equivalence classes.

Because there are three significant eigenvalues under consideration, two separate graphics are reviewed: the scatterplot matrix examining each vector (Figure 3.12) and a three-dimensional graphic of the points (Figure 3.13). Of note, the three-dimensional graphic can be rotated for a better view of the data. A single rotation was captured for this discussion. The classes seem to be relatively evident from the specific view captured. When the rotation uses U_3 as a major axis of rotation, however, the same messy elements seen in Figure 3.11 and 3.12 reappear.

Conclusion for Five Sources Split into Two Small Factions and Added Noise

Of the simulations discussed thus far, the data set here represents a more realistic situation where the data are not created as perfectly. The first thing to note is that there are factions and equivalence classes in the initial structure. Because noise was added to the data, there will be instances even at an α level equal to 0.10 that will conclude no real relationship exists. Allowing for slightly higher levels of α may need to be considered when the goal is to truly understand the structure of the underlying data. For our purposes, we chose to rely on standard values and so there are times where the significance of the data set would have been missed. When looking at the other common scenarios (two and three significant eigenvalues identified), there is evidence that the overall conclusions are relatively similar. Even when the third eigenvalue has been identified, it adds little value to the overall picture of the data except for the opportunity to see the equivalence classes

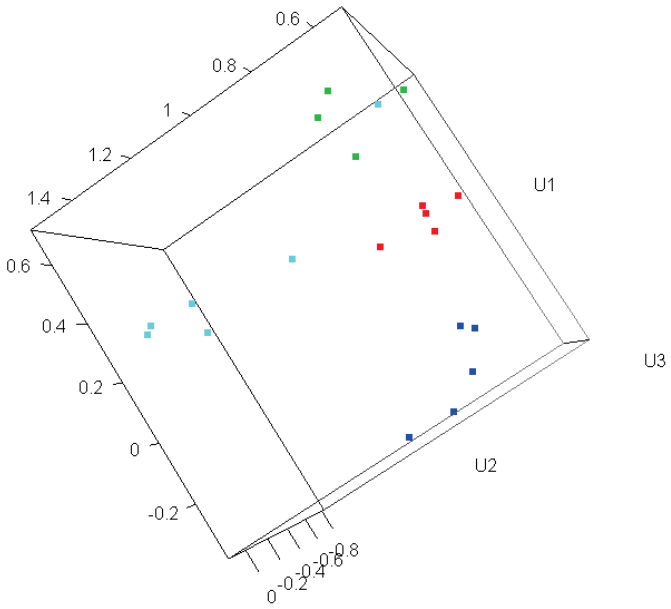

Figure 3.13: Single shot of a three-dimensional graphic looking at the three vectors of U^*. Using three dimensions to partition the classes provides a better view of the resulting equivalence classes. Dark blue = d, red = e, green = f and light blue =g.

in three dimensions. In reality, the two significant eigenvalues captured the same information. Overall, when significance was found (at $\alpha = 0.10$, this was 72% of the time), the final conclusions about the initial structure were identified and for the most part were accurate. The main issues observed here was that low α levels resulted in the noise variable overwhelming the initial structure and instances where the noise variable attached itself to one of the factions where a relationship was not meant to exist.

3.2.4 Five Sources Where All Ranking Is Random

For this scenario, the simulated data were created so that there are $m = 5$ sources with $n = 20$ elements being ranked randomly. Every time the simulation was performed at any of the commonly accepted α levels (0.10, 0.05, 0.01), no significant eigenvalues were found (see Table 3.12). Even with an α level as high as 0.30, only 9 of 1000 cases were found where any eigenvalues were considered significant. With no significant eigenvalues, there is no significant correlation. Therefore, there is no reason to examine the vectors created using R-A SVD because neither factions nor equivalence classes exist. When a single replication was considered, the $\lambda_{\{R-A\ SVD\}}$ was -0.0085, which is close enough to zero as to be considered insignificant. Kendall's W overestimates the correlation with a value of 0.183 as would be expected when the data are noise. A single replication was considered. When Anderson's test is performed, no significant eigenvalues are found even with an α level as high as 0.50. These results indicate that this method provides evidence of no association when the data structure has random association. As such, the use of V and U^* is unnecessary and provides no insight.

Table 3.12: 1000 Simulations Created Where No Sources Agree

	Number of Significant Eigenvalues					
α	0	1	2	3	4	5
0.30	991	0	0	0	9	0
0.10	1000	0	0	0	0	0
0.05	1000	0	0	0	0	0
0.01	1000	0	0	0	0	0

3.3 Conclusions

R-A SVD is a new measure of correlation among multiple rank sources. Unlike the commonly used method (W), R-A SVD measures overall association (see Section 2.3) and provides insight into the structure of the ranked data, allowing one to understand when sources have similar viewpoints (factions) and identifying when elements fall into meaningful classes. With one significant eigenvalue, the overall correlation is shown to be significant and the ability to look into the structure provides a one-dimensional view of the possible factions and clusters. With each additional significant eigenvalue, a new dimension is added and provides a more in-depth view into the structure of the ranked data.

CHAPTER 4

APPLYING R-A SVD TO 2011 NCAA FOOTBALL RANKINGS WHEN COMPARING RANKED SOURCES

In chapter 4, R-A SVD will be applied to collected rank data to ascertain the use and validity of the methods outlined in this research. These steps will be applied to four different data sets that include: ranking data for the top 25 NCAA football teams for the 2011 season, the top 50 inventions since the wheel as collected by the monthly periodical, *The Atlantic*, corporate sustainability with respect to ranking the most sustainable corporations and the top twenty M.B.A. in the northeast region as reported by the *U. S. News and World Report* from 2006 through 2014. For the NCAA data, two analyses will be performed: a comparison of the rankings by source and a look at each rank source over time. Although they use the same data set, each will be listed within separate sections. For all of the analysis, there will be a brief introduction to the specific data set considered. Then, all of the steps of the R-ASVD are applied (see Section 2.3 and Section 3.1 for a discussion of these steps). Overall, these data represent the first set of

applications where we gain insight as to how well the R-A SVD identifies underlying patterns, factions, and equivalence classes using collected data.

4.1 NCAA Football Rankings (2011 Season)

Significant spending on college athletics is often justified on the grounds that athletic success attracts students and raises donations. There are claims that winning football games enhances a school's reputation, increases the number of applications of in-state students, reduces acceptance rates, and raises average incoming SAT scores [8]. Therefore, our decision to analyze the ranking of the top 25 college football teams recognizes the potential fiscal impact for understanding whether the underlying rankings are associated and whether equivalence classes exist.

4.1.1 Overview of the 2011 NCAA Football Season Data Set Analysis Using R-A SVD

In this section, we provide information regarding the sources that have created the rankings and the timeline for which the data were collected. A set of raw data are shown to illustrate of what is being analyzed. The graphics that are used throughout this chapter are identified. The correlation techniques and the step-by-step process (as it will be applied to all data sets) are discussed.

Sources Ranking NCAA Football Teams

For the purposes of this research, we focused on four different organizations that rank college football. The first ranking source is the Associated Press College Poll (AP Poll), which provides the rankings for each week of the top 25 NCAA teams in college football. The rankings are compiled by polling sportswriters across the nation. Each voter provides his or her

own ranking of the top 25 teams, where a team receives 25 points for a first place vote, 24 points for a second place vote, and so on down to 1 point for a 25^{th} place vote. The individual rankings are then combined to produce the national ranking by summing all of the scores and giving the team with the most points a rank of 1, the next highest a rank of 2, until the top 25 have been found. The second ranking source is the *USA Today*'s Coaches Poll (Coaches), which is compiled in a manner similar to the AP Poll but instead asks a board of 59 head coaches from Division I football bowl subdivision institutions to vote for the top 25 teams. The third ranking source is the Bowl Championship Series (BCS), which is only released for weeks 7 through 14, and relies on a combination of polls and computer selection methods to determine relative rankings. Of note, this source includes data from the previous two ranking methods and therefore has an inherently dependent relationship. The final ranking source considered is the Harris Interactive College Football Poll (Harris), which is available only for weeks 6 to 14. Harris Interactive, a market research company that specializes in Internet research, compiles the rankings.

Timeframe for Comparing Rankings from All Sources Using R-A SVD

Although there are 15 weeks in a standard college football season, the purpose of examining these data is to compare multiple sources ranking the same set of elements. In these examples, it is assumed that the data will be highly correlated sets of ranks produced with a great many overlapping elements among the sources. For the first analysis of the 2011 NCAA football ranking data set using R-A SVD, only weeks 7 through 14 were considered since it is the only timeframe during which all four sources released rankings. We chose to look at the 2011 season because the data were recent and readily available. Of note, the rankings change during the season due to the nature of the game. As top-seeded teams lose and unseeded teams unexpectedly win, some new teams join the list, some current teams fall off the list and

other teams exchange places. Even though the ranking is of the top 25 teams by week, there are a total of 43 teams ranked within the polls between weeks 7 and 14 in 2011. A second analysis which considers analyzing the data by each individual source over the weeks that rankings were released can be found in Section 5.1.

Modified xy Plot

During this analysis of the NCAA data set, a graphic that we will refer to when looking at comparative sources ranking a similar set of elements is a modified xy plot. The purpose of the modified xy plot is to visualize overall association among the different sources. First, an xy plot for each source is created by modeling data from one source to each of the other sources (including itself). As an example using week 7 of the NCAA data set, column 1 of Figure 4.1 shows each individual sources mapped to the AP Poll including the AP Poll mapped to itself, thus producing four xy plots. At the bottom of Column 1, all four plots are drawn so that they overlap, thus resulting in one graphic containing the data from each xy plot that allows multiple sources ranking the same data to be compared simultaneously. To see that the results are similar regardless of the source to which the mapping is being done, this process was repeated for each source (i. e., modeling each source against BCS, then Coaches, and, finally, Harris) and can be seen in columns 2 through 4.

Figure 4.1, which was created using the initial rankings for week 7 (see Table 4.1), compare the top 25 ranked collegiate football teams during week 7 of the 2011 season. On Table 4.1, the order for each rank goes from 1 to 26. Although this case looks at the top 25 rankings in college football, the number 26 is being used as a placeholder when non-overlapping elements exist (per Section 2.2.2); a total of 27 football programs are included in this top 25. The graph created to visualize association can be found at the bottom row of column 1 of Figure 4.1 where we consider the elements falling along

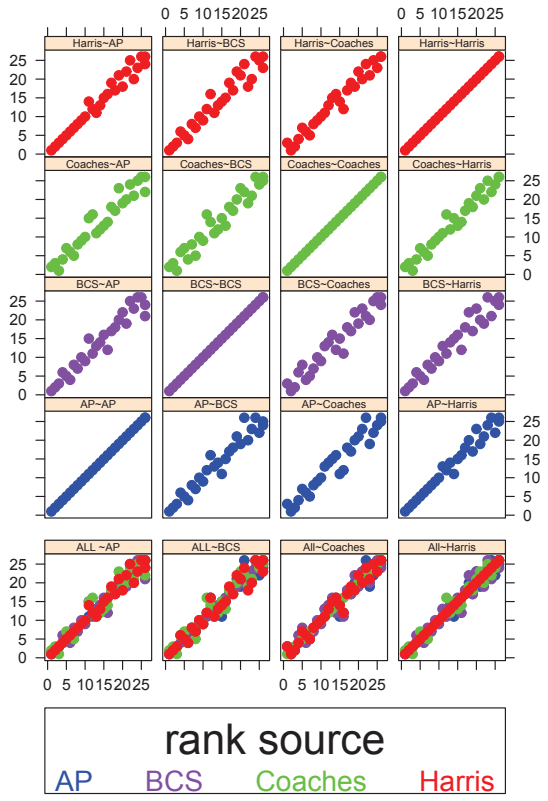

Figure 4.1: The modified *xy* plot created using the raw data from week 7. In the first column, the four *xy*plots shows each source as modeled by the *APPoll* with distinct colors. The bottom graphic in column 1 shows all four sources overlapping where the colors from each individual plot have been retained for week 7. The same pattern is followed in the remaining columns where all sources are mapped to BCS, Coaches, and Harris, respectively, at the bottom of columns 2, 3, and 4. Color code key: blue = AP Poll; red = Bowl Championship Series (BCS); green = *USA Today*'s Coaches Poll (Coaches); magenta = Harris Interactive (Harris).

Table 4.1: Week 7 Polled Rankings by Source 2011 NCAA Data

Team	AP	BCS	Coaches	Harris
LSU	1	1	2	1
Alabama	2	2	3	2
Oklahoma	3	3	1	3
Wisconsin	4	6	4	4
Boise State	5	5	7	5
Oklahoma State	6	4	6	6
Stanford	7	8	5	7
Clemson	8	7	8	8
Oregon	9	10	9	9
Arkansas	10	9	10	10
West Virginia	11	15	15	14
Kansas State	12	11	16	12
Nebraska	13	13	11	11
South Carolina	14	14	12	13
Michigan State	15	16	13	15
Virginia Tech	16	12	14	16
Texas A & M	17	17	18	19
Michigan	18	18	17	17
Auburn	19	20	23	21
Georgia Tech	20	22	19	18
Houston	21	19	20	22
Washington	22	25	24	25
Illinois	23	23	21	20
Arizona State	24	26	25	23
Georgia	25	26	26	26
Penn State	26	21	22	24
Texas	26	24	26	26

the diagonal running from (0,0) to (26,26). The more closely the elements fall on or near the diagonal, the higher we would expect the measure of overall association to be. Since all four plots look very similar, we will solely use all sources mapped to the AP Poll throughout the remainder of this analysis.

4.1.2 Overall Information of Weeks 7 through 14 of the 2011 NCAA Football Season

For each week, the same steps, tools and graphics of R-A SVD are applied to the data set. This section provides the results obtained from calculating the various correlation methods ($\lambda_{\{R-A\ SVD\}}$ and W), the modified xy plots produced from the raw data (all sources mapped to the AP Poll), and the number of significant eigenvalues found using Anderson's test for each week. These results will be referred to when discussing specific findings for each week. In order to concisely see that using R-A SVD to assess association is a better method than W as well as provide a snapshot of the rest of the findings, the specific results for each week will be explained in Sections 4.1.3 through 4.1.6. Sections 4.1.3 to 4.1.5 will discuss only three cases in detail (respectively, weeks 7, 8 and 10) since the rest of the findings are very similar in nature. The significant findings from the remaining weeks will be discussed briefly in Section 4.1.6.

Correlation Methods Compared, xy Plots and Significant Eigenvalues by Week

Table 4.2 provides the correlation averages by method for each of the eight weeks studied for the 2011 football season in the first three columns. Average τ, which was discussed in Section 2.4, is listed in column 1 to reinforce how closely it is related to the value found using R-A SVD ($\lambda_{\{R-A\ SVD\}}$) in column 2. Kendall's W is provided in the third column. In the fourth column, the number of significant eigenvalues found for each of the eight

Table 4.2: Average Correlation Methods by Week NCAA 2011 Football Rankings

Week	Average τ	$\lambda_{\{R-A\ SVD\}}$	Kendall's W	# of Eigenvalues
7	0.8759	0.8796	0.9786	1
8	0.9004	0.9006	0.9810	2
9	0.8971	0.8972	0.9814	1
10	0.8786	0.8788	0.9702	3
11	0.8575	0.8576	0.9607	1
12	0.8733	0.8736	0.9553	3
13	0.8421	0.8424	0.9267	1^*
14	0.8600	0.8607	0.9318	3

weeks is listed.

Figure 4.2, a lattice of modified xy plots, shows the strength of the association of the sources when modeled versus the AP Poll by week. Notice the lines inside the graphs are the partitionings provided by the dendrograms (not shown here). These data represent each week, from week 7 to week 14, in progression. When examining the correlation values from Table 4.2 and Figure 4.2 together, there is overwhelming evidence that Kendall's W overestimates the association among the sources, whereas using R-A SVD yields more realistic average correlation values (as judged by the scattered patterns on Figure 4.2). If the association were as high as the values attained by Kendall's W, one would expect to see a nearly perfect straight line along the diagonals of each plot.

Each week considered for this study provided at a minimum one significant eigenvalue. Of note, the single eigenvalue for week 13 is represented as 1^*. In all cases except this instance, the number of significant eigenvalues were identical at α levels of both 0.05 and 0.10. In week 13, only one significant value was found at an α level of 0.05, while three significant eigenvalues were identified at an α level of 0.10. Most likely, these results are due to red and green dot located very far above the $x-$axis value around the ranking of 10 in Week 13 (Figure 4.2). Please refer to Section 4.1.6 for more detail.

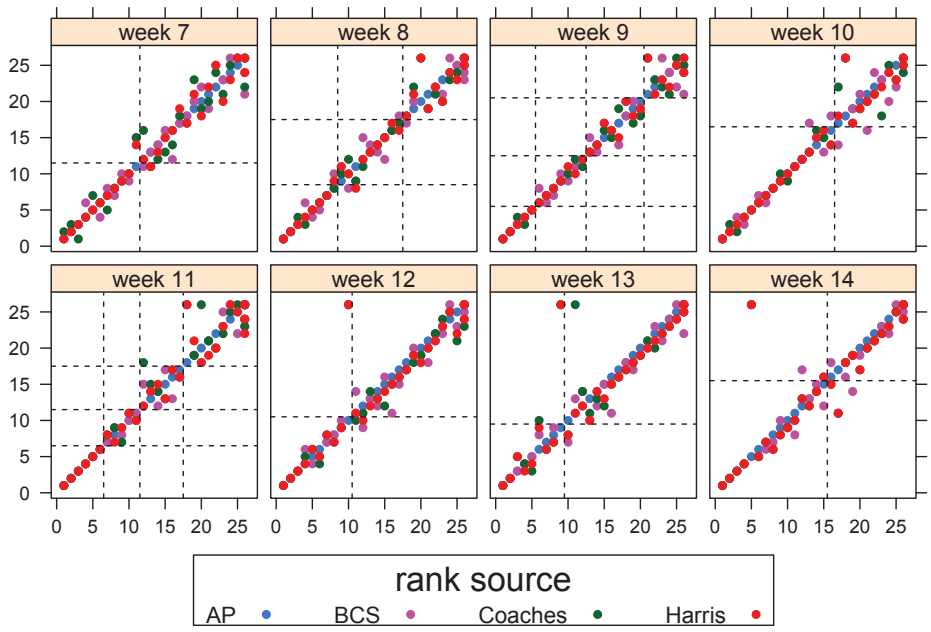

Figure 4.2: Modified xy plots for each week for weeks 7 through 14 that show correlation and feasible partitioning, which were chosen in each week by dendrograms (not shown) created using the significant vectors of U^* as defined in each specific week discussed below. Color code key: blue = AP Poll; red = Bowl Championship Series (BCS); green = *USA Today*'s Coaches Poll (Coaches); magenta = Harris Interactive (Harris).

Table 4.3: First Vector of V from Week 7 of the NCAA Football Rankings

Source	Vector 1
AP	0.5031
BCS	0.4915
Coaches	0.4978
Harris	0.5075
Eigenvalue	547.34655

4.1.3 Week 7 Analyzed Using R-A SVD

The raw data in Table 4.1 shows the 27 teams being ranked by the four sources during week 7. The overall association for $\lambda_{\{R-A\ SVD\}}$ is 0.8796, whereas W has a value of 0.9786. Referring to week 7 on Figure 4.2, this plot indicates a strong but not perfect overall association among the four ranked sources. In order to determine whether the R-A SVD association is significant, Anderson's test is performed at an α level of 0.05, and one significant eigenvalue is identified.

With only one possible faction of sources identified, V_1 is considered to determine which sources belong within this faction. Table 4.3 indicates that all four sources strongly agree as to the top 25 rankings in week 7 of the 2011 season, which is evidenced by each value closely approaching $\sqrt{1/q}$ where $q = 4$ agreeing sources. To further show that these four sources are in agreement, we also considered the correlations within the first principal component. As stated in Chapter 3, determining which sources are part of a specific principal component is based on approximate estimates without an exact test to prove they are significant. On Table 4.4, all of the values strongly agree in both strength (values exceed 0.60) and direction (all values are positive).

Finally, we will determine whether equivalence classes exist within this data set. The values within the U^* matrix create distances between the elements that reflect the correlation information from the structure of the data. Since this data set has a single significant eigenvalue, the associated

Table 4.4: Correlations of Four Rank Sources in Week 7 with the First Principal Components

Source	Component 1
AP	0.95967
BCS	0.93770
Coaches	0.94958
Harris	0.96805

graphics can only be considered in a single dimension. The dendrogram (Figure 4.3), which was created using U_1^*, provides evidence of either two classes (using a y-axis height of .0.55 or higher) or three classes (using y-axis values between 0.35 and 0.55). For this data set, we chose three classes and used the height between 0.35 and 0.55.

Next, the xy plot (Figure 4.4) is created by mapping U_1^* to itself. Color coding is added to match the partitioning from the dendrogram. Larger spaces between points (representing the teams) show the distance between classes and the partitioning. Both the dendrogram and the xy plot (Figure 4.4) identically identify these classes. The distances observed on the $y-$axis of the xy plot suggests that three classes are reasonable. Additionally, U^* provides the ordering of the initial data on the xy plot (this is obscured on the dendrogram). On Figure 4.4, the teams in blue represent the top 6 teams ranked by each source. In contrast, they appear in the center of the dendrogram (Figure 4.3). Next, the second class of teams ranked by the four sources (teams 7 through 16) appear in red. The lowest ranked teams (green) appear at the bottom.

Overall, the rank data set for week 7 demonstrates association among the four rank sources when R-A SVD was applied, which is confirmed by finding the measure of overall association significant. Both the vectors and principle components (Tables 4.3 and 4.4) provide evidence that there is a great deal of agreement among all four sources, thus confirming there is one faction during week 7. However, it should be noted that the agreement is not perfect as evidenced by the xy plot for week 7 in Figure 4.2. Using the first

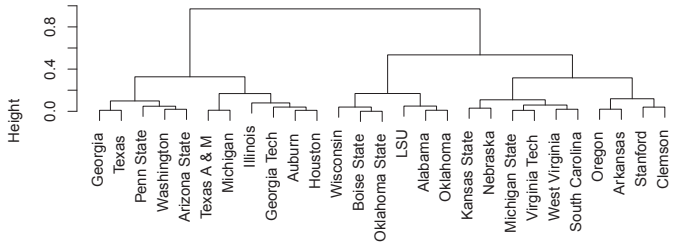

Figure 4.3: Classic dendrogram using the rankings from week 7 to identify proper partitioning. Three equivalence classes are considered using a height on the $y-$axis between 0.35 and 0.55.

vector of U^* provides evidence of at least three equivalence classes among the ranked elements where the classes are identified with the dendrogram in (Figure 4.3), but the actual ordering and distances of each class are seen in the xy plot of the first vector of $U*$ mapped to itself (Figure 4.4).

4.1.4 Week 8 Analyzed Using R-A SVD

The next week from the 2011 season to be factored using R-A SVD is week 8. In this week, 28 teams were ranked, with a placeholder value of 26 used for non-overlapping elements. Similar to week 7, there is a strong overall association among the four ranked sources. The calculated value of $\lambda_{\{R-A\ SVD\}}$ is 0.9006, while W overestimates the relationship at 0.9810. Such a high W value implies an almost perfect association, which would have resulted in most points aligning perfectly on the diagonal of week 8's modified xy plot (Figure 4.2). Instead, we see slightly scattered points. When Anderson's test is performed ($\alpha = 0.05$) to determine whether the association is significant, two significant eigenvalues are found, which implies that factions of sources and/or equivalence classes may exist.

Next, the first two vectors of V are considered to identify which sources

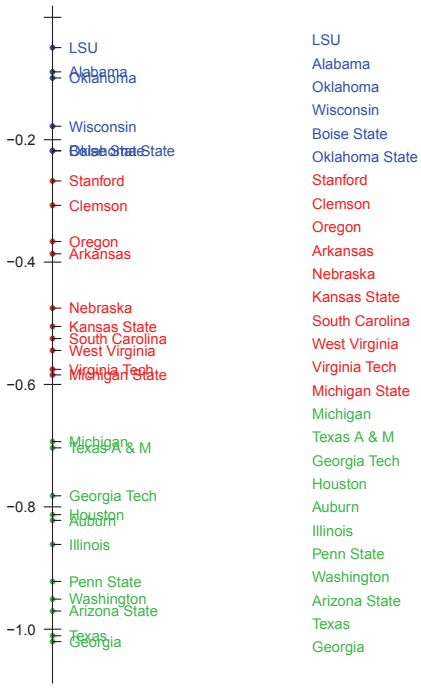

Figure 4.4: An *xyplot* of week 7 creating by mapping the first vector of U^* to itself. The top class of teams (1 through 6) are in blue, the next best class of teams (7 through 16) are in red and the bottom class of teams (17 through 26) are in green. The Figure on the left represents the actual positioning of the dots which leads to overprinting, hence we reprint the names in the same order in the second column for clarity of the actual order.

Table 4.5: First Two Vectors of V from the Week 8 College Football Rankings

Source	Vector 1	Vector 2
AP	−0.4982	−0.4176
BCS	−0.4879	0.8413
Coaches	−0.5049	−0.3374
Harris	−0.5088	−0.0632
Eigenvalues	2776.66862	121.85799

belong within each faction. There are surprising results in Table 4.5, which require a more in-depth look at the data set to understand the anomalies found. V_1 implies that all four sources exist within one faction and that all four sources strongly agree as to the top 25 rankings. V_2 provides evidence that the second source, BCS, may have some issues in regards to the strength with which it agrees to the other sources. Notice in V_1 that each value closely approaches $\sqrt{1/q}$ where $q = 4$ agreeing sources. Yet, V_2 identifies a strong positive BCS while indicating that the values for two of the sources (AP Poll and Coaches) have inverse signs (indicating opposition) to BCS and moderately agreement with one another. V_2 also shows that the ranking from Harris seems to be noise. These results require an investigation into BCS to uncover what may be causing this significant difference.

To confirm that the first vector of V indeed indicates that all four sources are in agreement, the correlations within the first principal component were calculated. Looking at the first column of Table 4.6 confirms that the four sources agree strongly within the first component, as they all have values exceeding 0.60. The second column does not indicate any real correlation (all value are below 0.40) for any of the sources within the component. Of these values, BCS is the strongest and has a positive value. The remaining sources are much weaker and are negative; Harris approaches zero. As has been stated before, most of the variation of the initial data structure comes from the first eigenvalue (Section 2.3.1). Although the second principal component is significant (but not as significant as the first component considering

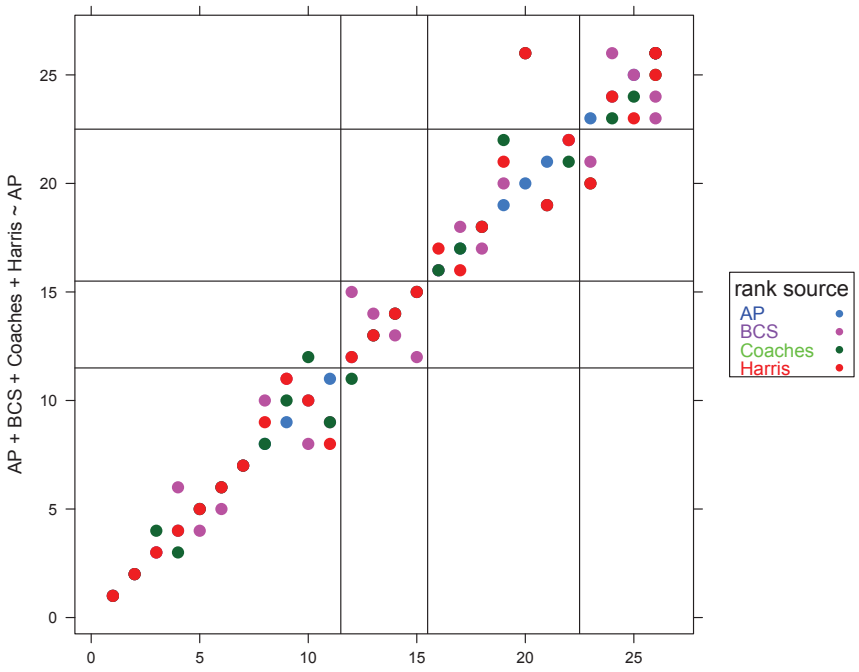

Figure 4.5: The modified *xy* plot of all sources plotted against the AP Poll during week 8. Notice that the perpendicular lines (magenta) along the diagonal indicate the disagreement with specific ordering by BCS. Color code key: blue = Associated Press (AP) Poll; magenta = Bowl Championship Series (BCS); green = *USA Today*'s Coaches Poll (Coaches); red = Harris Interactive (Harris).

Table 4.6: Correlations of Four Rank Sources in Week 8 Considering the First Two Principal Components

Source	Component 1	Component 2
AP	0.95856	−0.16833
BCS	0.93886	0.33912
Coaches	0.97141	−0.13599
Harris	0.97892	−0.02546

the eigenvalues found in the bottom of Table 4.5), it does not provide nearly as much information, which is the reason for the weaker correlations.

Before we consider the final step, we will review the modified xy plot of each source versus the AP Poll (Figure 4.5). This graphic reveals multiple locations where BCS agrees to the inclusion of football teams within a specific range of rankings but orders the teams inversely to every other rank source. In Figure 4.5, there are various areas where this type of disagreement occurs, but what can be seen succinctly is that the magenta points representing BCS often form a perpendicular hash mark across the diagonal as the line increases from 0 to 26. At two specific locations, between rankings 11 and 15 and between 23 and above, there is evidence that the other three sources are in tight agreement while BCS is not. This disagreement is visualized on the graphic by the perpendicular line of magenta points making an X.

Finally, we will determine whether equivalence classes exist within the data set for week 8. Since two significant eigenvalues exist, the dendrogram (Figure 4.6) and xy plot (Figure 4.7) are created using the first two vectors of U^*. Using a $y-$axis height between 0.45 and 0.60, three equivalence classes are identified on the dendrogram. The classes identified from the dendrogram (Figure 4.6) are then color coded on the xy plot (Figure 4.7), which identifies similar partitioning of equivalence classes (here, groups of teams as identified by the color coding). The xy plot also reveal the ordering of these classes. The top 8 teams are represented by the black at the top of the figure. The next best class, consisting of the next 9 teams can be seen in red, and the lowest class consisting of the remaining 11 teams is identified

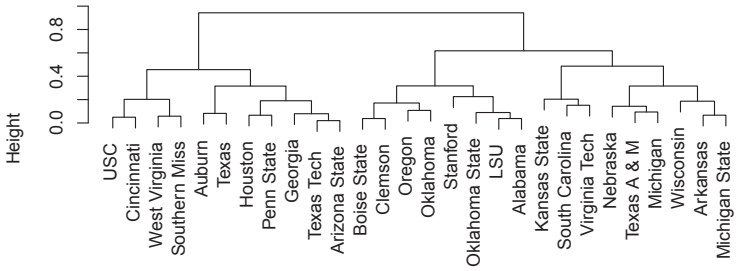

Figure 4.6: Classic dendrogram using the rankings from week 8 to identify proper partitioning. Three equivalence classes are considered using a height on the y-axis between 0.45 and 0.60.

in green.

As a ranked data set, week 8 has evident association among the four sources, which is confirmed by the significant measure of overall association. Table 4.5 and Table 4.6 provide evidence that there is a great deal of agreement among all four sources (creating a single faction). However, there are some unexpected results, possibly influenced by the BCS poll. Although the BCS Poll agrees to the overall rankings, it provides inverse orderings within smaller groups when compared to the remaining three sources. Overall, the method appears to be affected by the slightly rebellious BCS (hence, the second significant eigenvalue). As indicated by the agreement among all the sources and the identification of reasonable equivalence classes, it appears that the results are not overly affected by the second eigenvalue.

4.1.5 Week 10 Analyzed Using R-A SVD

In week 10, 27 teams were ranked; a placeholder value of 26 was added when non-overlapping elements occurred. When week 10 is factored using R-A SVD, the overall association among the four ranked sources as measured by $\lambda_{\{R-A\ SVD\}}$ is 0.8788 and by W is 0.9702. The association expressed by the

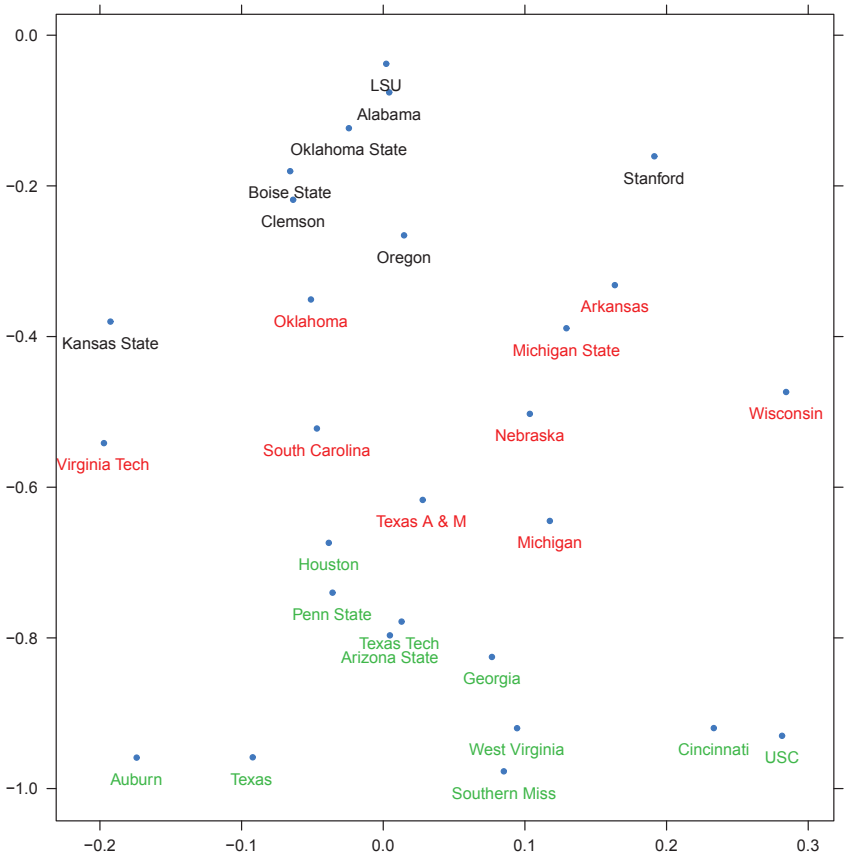

Figure 4.7: An *xyplot* of week 8 created by mapping the second vector of U^* to the first. The top class of teams (1 through 8) are in black, the next best class of teams (9 through 17) are in red and the bottom class of teams (18 through 28) are in green.

Table 4.7: First Three Vectors of V for the Week 10 College Football Rankings

Source	Vector 1	Vector 2	Vector 3
AP	−0.5010	0.0848	0.7841
BCS	−0.4861	−0.8395	−0.2392
Coaches	−0.4991	0.4746	−0.5725
Harris	−0.5134	0.2507	0.0178
Eigenvalues	2545.92024	132.31149	91.10027

scattered points for week 10 on Figure 4.2 do not appear to be close to a perfect straight line as would be evident with a strength of 0.9702, indicating that W overestimates the relationship. In order to determine whether this association is significant, Anderson's test was performed at an α level of 0.05. In this case, three significant eigenvalues are found, which implies that factions of sources and/or equivalence classes may exist. As was seen in week 8, the presence of three eigenvalues potentially may have more to do with the equivalence classes than the sources.

Next, R-A SVD considers the first three vectors of V to help identify which sources belong within each faction. The results, which are very similar to those found in week 8, are listed in Table 4.7. Much like weeks 7 and 8, of the week 10 data indicates a very strong association among all four sources with values that approach $\sqrt{1/q}$ where $q = 4$. According to V_1, all four sources also agree on directionality. However, V_2 again identifies BCS strongly and shows that the values for the other three rank sources are weaker, with the exception of the Coaches Poll which strongly disagrees with BCS. Since each of the other three sources have inverse signs to BCS, there may be contention between BCS and the other sources. Similarly, V_3 identifies the AP Poll strongly. Again, the other three rank sources are weaker. Both BCS and Coaches have inverse signs to the AP Poll, with Coaches seeming to disagree strongly. It is important to note that the first vector corresponds to a much larger eigenvalue, as can be seen in the bottom row of Table 4.7, and hence provides most of the useful information.

Table 4.8: Correlations of Four Rank Sources in Week 10 Considering the First Three Principal Components

Source	Component 1	Component 2	Component 3
AP	0.95555	0.03685	0.28285
BCS	0.92697	−0.36497	−0.08628
Coaches	0.95189	0.20634	−0.20652
Harris	0.97907	0.10898	0.00642

To affirm these discussion points, the correlations within the three significant principal components are calculated (Table 4.8). The first column continues to indicate that the four sources strongly agree within the first component. The second column does not indicate any strong correlation for any of the sources within the second component. Of these sources, the strongest is BCS, which has a negative value. Coaches has an opposing positive value. The values associated with the other sources are much smaller. The third component shows a similar disagreement between Coaches and the AP Poll. Although weak, the values for these two sources are similar but with inverse signs; the remaining sources are inconsequential. Importantly, the second and third columns have no values exceeding 0.40, meaning there is not much evidence of correlation within the second and third components.

Finally, we will determine whether equivalence classes exist within the data set for week 10. Since there are three significant eigenvalues of U^* under consideration, a dendrogram (Figure 4.8), scatterplot matrix (Figure 4.9), and a three-dimensional scatterplot (Figure 4.10) will be discussed. As in the previous cases, the dendrogram is used to decide the number of classes and the elements in each class. If a straight line is chosen at a height of 0.95 on the y-axis of Figure 4.8, the two classes (one weaker and one more cohesive) and a singular straying point (USC) are revealed. When referring to the raw data, it is discovered that USC was ranked 18 by the AP Poll but was not considered significant by the other sources. The information regarding *USC* combined with the behavior of the graphic leads to an initial indication that the results of R-A SVD (specifically, the number of eigenvalues) may

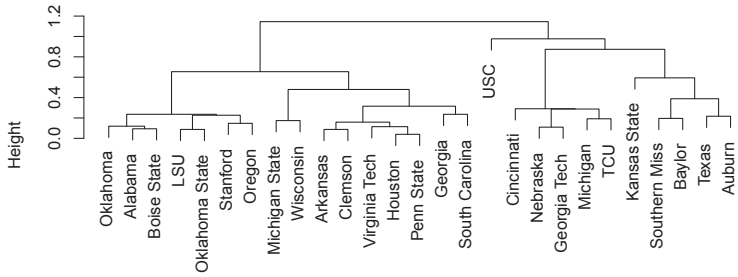

Figure 4.8: A dendrogram showing two classes and a singular straying point (*USC*) when looking at the rankings for week 10.

be affected when an element is only considered by a single source and is quickly identified when considering the U^* graphics.

The equivalence classes found by the dendrogram also are identified by the scatterplot matrix (Figure 4.9) of the significant U^* vectors mapped to each other, again demonstrating that U^* graphics readily identify a single element ranked by one source. Figure 4.9 uses three colors to help identify the two classes and the point that does not belong with the others. When looking at the comparison of U_1 by U_2, the separation of the blue and red dots is evident while the single green dot looks as though it could belong in the center of red class. When considering U_1 by U_3,the separation of the blue and red classes is mostly evident with no overlap. Here, the green point becomes more of an outlier, possibly similar to an influential observation as might be found in a regression analysis. Recall that an influential observation can completely change the results of an analysis and is identified when Cook's distance metric is applied and a value over 1 is attained. We recognize that a similar test may be needed when single elements are only ranked by a single source influencing the data and plan to explore this in future work.

When considering U_2 by U_3, the blue and red clusters are not as clearly partitioned, while the green point remains far out of the picture (possibly

influencing the data). The larger class on the left in Figure 4.8 represents the top 16 ranked teams, which are the blue points in Figure 4.9. The smaller class and USC on the right of the dendrogram are the red points (the next best class ranked 17 through 26) and USC is the lone green point (which is excluded). Together, these two graphics indicate that at least two reasonable equivalence classes can be captured from the first three vectors of U^*.

Finally, the three-dimensional graphic (Figure 4.10) clearly shows the partitioning of the classes as well as the single straying point identified by both Figure 4.9 and Figure 4.8 (again, demonstrating that U^* graphics readily identify a single element ranked by only one source). When seen live, the axis can be rotated to show that these classes exist in most views of the data. The snapshot seen in Figure 4.10 only captures the third dimension slightly. The blue, red, and green points continue to show two separated classes and a stray point (USC).

Week 10 with USC Eliminated

In order to see the effect of eliminating the one ranked element that was only ranked by a single source in week 10, USC was dropped from the data then the analysis was rerun. The correlation value for $\lambda_{\{R-A\ SVD\}}$ is 0.879, while W is 0.97. When Anderson's test is run, there are still three significant eigenvalues at both αlevels of 0.05 and 0.10. The corresponding three vectors of V provide almost identical results (Table 4.9), as do the correlations found within the principal components (Table 4.10).

The major differences that can be observed when USC is dropped from the data come from the equivalence classes identified from the first three vectors of U^*. Starting with the dendrogram (Figure 4.11), three classes are easily identified at heights ranging between 0.65 and 0.95. The three classes are color coded so that they are visible on Figure 4.12. The scatterplot matrix of each vector of U^* mapped to one another (Figure 4.12) reveals that the classes are quite clearly partitioned. The three-dimensional plot (Figure 4.13)

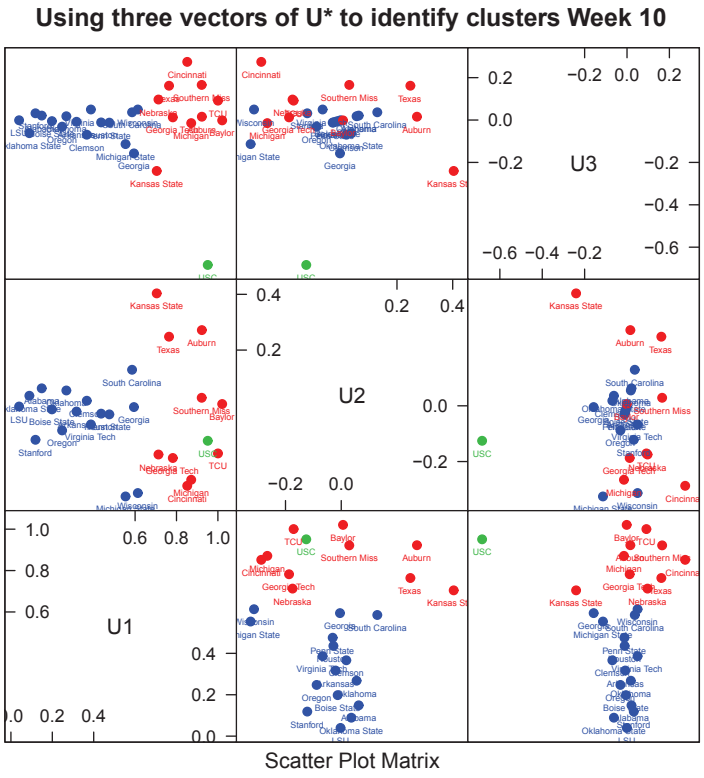

Figure 4.9: A scatterplot matrix representing the first three vectors of U^* that assists in identifying two classes and a stray point in the week 10 data. U_1 mapped to U_2 and U_1 mapped to U_3 both identify the partitioning, unlike U_2 mapped to U_3. The outlier (green dot), USC, is only isolated when U_3 is mapped to the remaining vectors. The top class of teams (1 through 16) is in blue, the next best class of teams (17 through 26) are in red, excluding USC in green.

Table 4.9: First Three Vectors of V for the Week 10 College Football Rankings with USC Eliminated

Source	Vector 1	Vector 2	Vector 3
AP	−0.5105	0.0374	0.5435
BCS	−0.4822	−0.8252	−0.2777
Coaches	−0.4957	0.5038	−0.6902
Harris	−0.5110	0.2527	0.3887
Eigenvalues	2395.80161	129.46412	58.95646

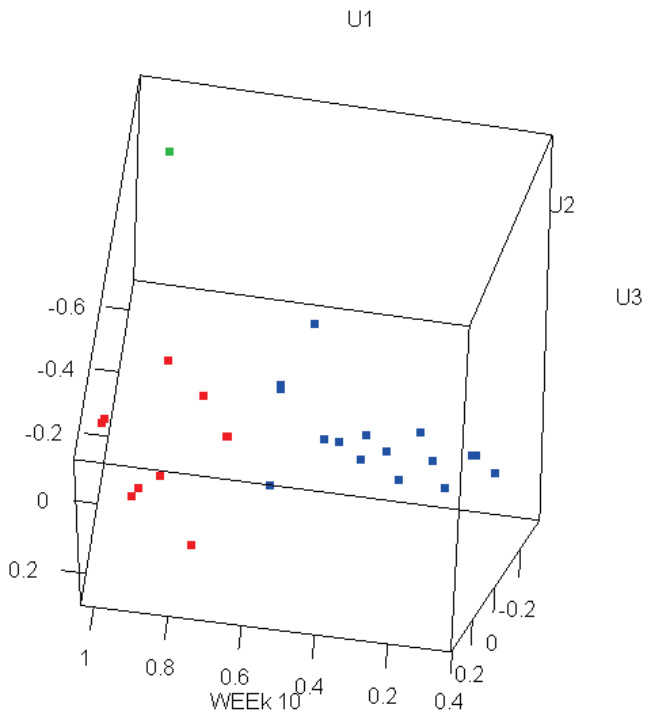

Figure 4.10: A still frame of the three-dimensional plot generated using the same color scheme as the scatterplot matrix and dendrogram to show the identified classes. The top class of teams (1 through 16) is in blue, the next best class of teams (17 through 26) are in red, excluding USC in green.

Table 4.10: Correlations of Four Rank Sources (with USC Eliminated) in Week 10 Considering the First Three Principal Components

Source	Component 1	Component 2	Component 3
AP	0.9815	0.0167	0.1639
BCS	0.9258	−0.3682	−0.0836
Coaches	0.9517	0.2248	−0.2079
Harris	0.9811	0.1128	0.1171

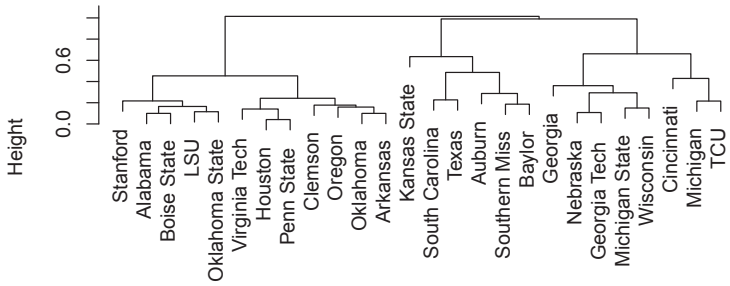

Figure 4.11: New dendrogram of the week 10 data after USC is dropped from the data. Three classes are evident using a height on the y-axis between 0.65 and 0.95.

allows for a better view of the absolute differentiation of the classes. From the initial data, the top 12 teams (identified in dark blue points) are clearly the best class. The other two classes are partitioned properly, although the ordering is not as evident.

Removing USC does not change the data or the initial discussion as to the correlation and factions. What the feasible influential point (USC) does affect is the proper partitioning of the classes. When included in the data, USC fell directly between two of the classes, making a large messy faction. When USC was removed, these two classes are clearly and easily partitioned.

Conclusion of Week 10 Data

Overall, this data set provides very similar insights to those gained analyzing weeks 7 and 8. By identifying three significant eigenvalues, R-A SVD allows for an in-depth look into the data, which identifies a single faction with two equivalence classes and a single straying point. When this point is removed, the information is similar although the classes are more clearly identified. It should be noted that the partitioned equivalence classes become more visible as more dimensions are available to view them. The first

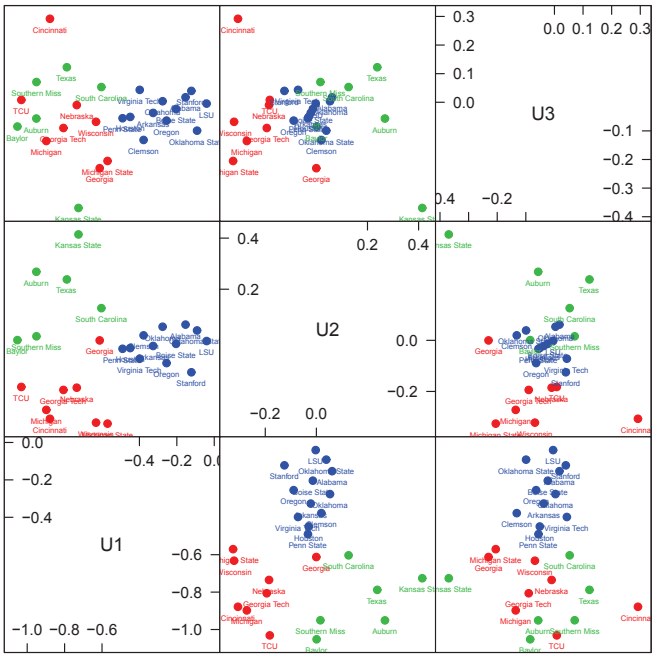

Figure 4.12: Scatterplot of the first three vectors of U^* mapped to one another, with USC removed from the data set. Each plot shows the three classes. When looking at U_1 mapped to the other vectors, the classes are clearly partitioned for the most part. There is some overlap when U^2 is mapped to U^3. Top 12 teams are dark blue, the other two classes (red and green) and clearly spearated but the ordering of second best and worst is not as evident when looking back at the raw data.

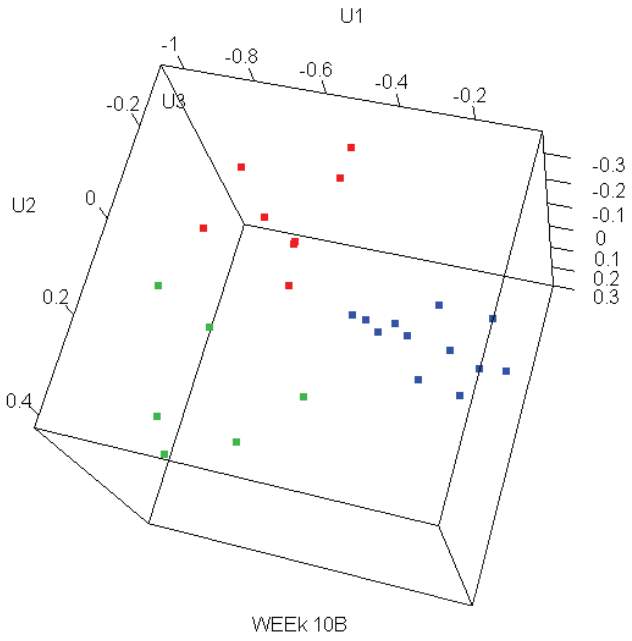

Figure 4.13: A still frame of the three-dimensional plot generated using the same color scheme as the scatterplot matrix and dendrogram to show the identified classes clearly partitioned with USC removed from the data set. The top class of teams (1 through 12) is in blue, the next set of classes of teams (17 through 25) are in red and green.

anomaly of a single school ranked by a single source also appeared here. This influential observation, which is also observed in the other weeks with three significant eigenvalues, will be discussed in more detail shortly. Depending on the number of influential observations observed, it is feasible it could elevate the number of significant eigenvalue found using Anderson's test, although that is not the exact case with this data.

4.1.6 Analyzing the Remaining Weeks Using R-A SVD

As was noted earlier, the findings for the remaining weeks of the study were very similar to weeks 7, 8 and 10. Therefore, we will briefly discuss the significant results from the remaining weeks, noting when these similarities do occur. Since weeks 9, 11 and 13 all identify a single eigenvalue, their findings will be reviewed as a group. Similarly, weeks 12 and 14 both have three significant eigenvalues and will be discussed as a group.

Weeks 9, 11 and 13 Analyzed Using R-A SVD

Much like week 7, weeks 9, 11, and 13 all identify a single significant eigenvalue. Therefore, only the first vector of V will be examined to identify the strength and association of the rank sources. In Table 4.11, V_1 indicates that all four sources show strong agreement by approaching $\sqrt{1/q}$ where $q = 4$, and the signs indicate the agreement in the direction of the rankings. To confirm these findings, the correlation for each source with the first principal component is considered (Table 4.12). These values indicate that all of the sources are very strongly correlated within the first principal component. From Tables 4.11 and 4.12, the first vector and the first principal component indicate that the results obtained during weeks 9, 11, and 13 provide similar insights to the results found during week 7, which identifies one faction with all sources in relative agreement as to the overall association.

Next, we will determine the feasible equivalence classes of the elements within the rank data. As always, the number of significant eigenvalues

Table 4.11: The First Vector of V from Weeks 9, 11, and 13

Source	Vector 1 (wk9)	Vector 1(wk 11)	Vector 1(wk 13)
AP	−0.4996	−0.5013	−0.4873
BCS	−0.4907	−0.4987	−0.5022
Coaches	−0.4997	−0.4936	−0.4922
Harris	−0.5098	−0.5063	−0.5189
Eigenvalues	2584.458	2851.232	2641.051

Table 4.12: The Correlation of Each Source with the First Principal Component During Weeks 9, 11, and 13

Source	PC 1 (wk9)	PC 1(wk11)	PC 1(wk13)
AP	−0.9600	−0.9464	−0.9144
BCS	−0.9429	−0.9415	−0.9423
Coaches	−0.9601	−0.9365	−0.9252
Harris	−0.9795	−0.9558	−0.9738

determines the dimensions in which these classes can be viewed. Since only one significant eigenvalue is identified for weeks 9, 11, and 13, the graphics are one-dimensional. Figure 4.14 shows the dendrograms (created by the first vector of U^* for each of the three weeks). For each dendrogram, the height at which a line is chosen for each dendrogram (which determines the number of equivalence classes and the teams within each class) varied but only slightly as the y-axis height of 0.30 across all the dendrograms could provide the classes for each week. For weeks 9, 11, and 13, the number of classes is determined to be four, four, and three, respectively.

Using the partitioning from the dendrograms, the color-coded xyplots (Figures 4.15, 4.16, and 4.17) reveal the same classes as seen on the respective dendrogram and additionally lists the team and their classes in their natural order (which is obscured on the dendrograms). The classes were chosen to once again show how well the method works. From here, a deeper discussion needs to be undertaken one week at a time but yields a discussion similar to the one found in week 7. Because tie breaker data was needed, these three weeks each have their own xy plot as seen in

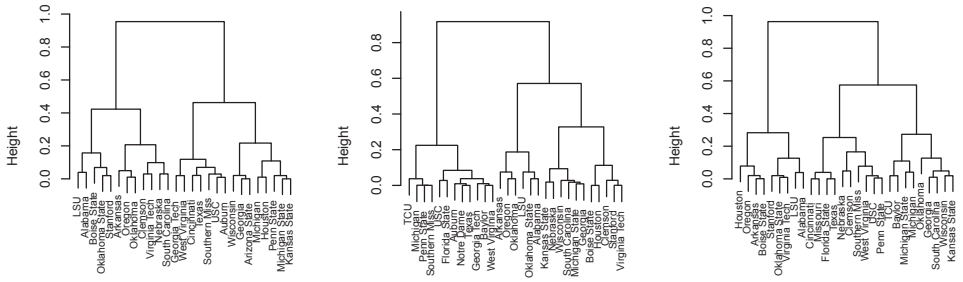

Figure 4.14: Set of dendrograms for weeks 9, 11, and 13 that represent the first vector of U^*, which is used to identify feasible equivalence classes.

Figures(4.15, 4.16, 4.17). Each figure views the first vector of U^* mapped to itself where each column shows plots of weeks 9, 11, and 13. Since only one dimension is really necessary to view this data, the graphics has been minimized horizontally, using the vertical height to help identify the distances of the classes that have been partitioned. On the y-axis, the elements (teams) included in each week's ranking can be identified on the right side of the graphic (due to overlapping values, a tie-breaker listing the teams in order is on the right side of each figure). For week 9, there is evidence of the four classes identified by the dendrogram (Figure 4.14), although the distances revealed seem to indicate at least one more partitioning could exist just above Wisconsin. Week 11 also shows three sizeable gaps, which are similar to the four equivalence classes identified by the dendrogram. Again, it appears that additional partitioning could be possible. Since the goal is to keep the classes from becoming minute, the classes are not further partitioned. Week 13 shows partitioning for at least three classes as well. In all cases, the top teams are labeled in light blue, the next best class is labeled in red, the third best class is labeled in green, and, when four classes exist, the fourth best class is labeled in dark blue.

Overall in this section, the tables and the analyses show that the results by week are very similar when one significant eigenvalue is found. That

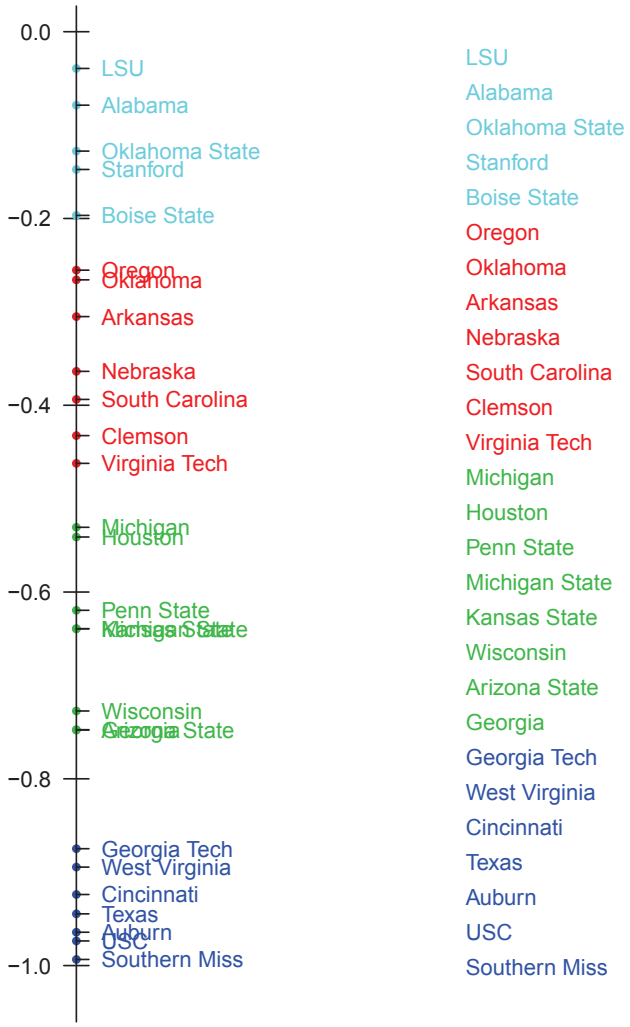

Figure 4.15: The xy plots for week 9, which shows the first vector of U^* mapped to itself. There are four classes clearly identified, which are color coded from the identified partitions from the dendrogram. The Figure on the left represents the actual positioning of the dots which leads to overprinting, hence we reprint the names in the same order in the second column for clarity of the actual order. Color code key: light blue = top ranked teams; red = second highest ranked teams; green = third highest ranked teams; dark blue = lowest ranked teams.

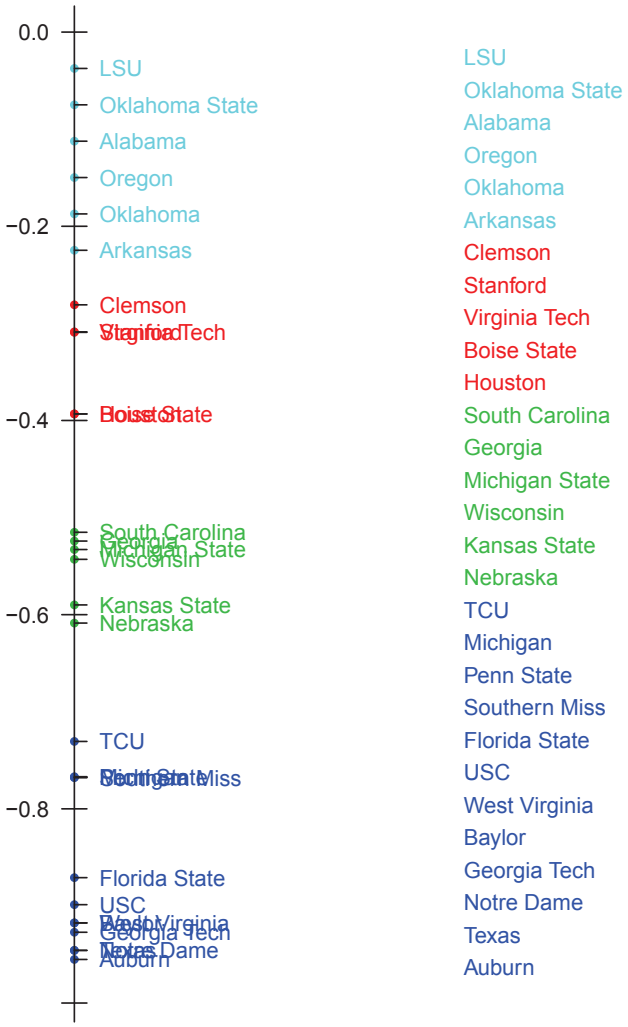

Figure 4.16: The xy plots for week 11, which shows the first vector of U^* mapped to itself. There are four classes clearly identified, which are color coded from the identified partitions from the dendrogram. The Figure on the left represents the actual positioning of the dots which leads to overprinting, hence we reprint the names in the same order in the second column for clarity of the actual order. Color code key: light blue = top ranked teams; red = second highest ranked teams; green = third highest ranked teams; dark blue = lowest ranked teams.

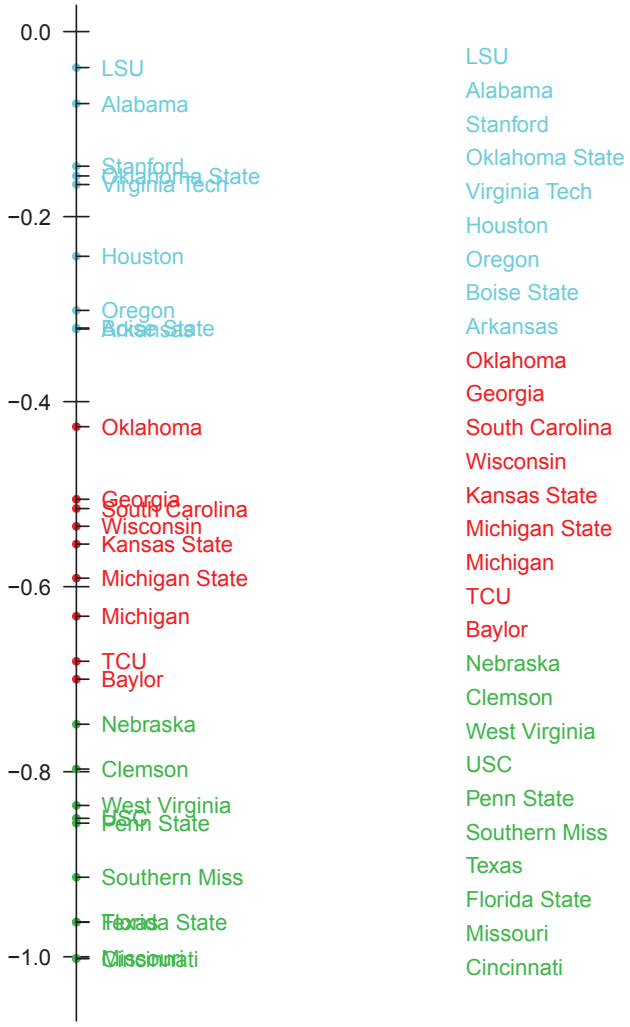

Figure 4.17: The xy plots for week 13, which shows the first vector of U^* mapped to itself. There are three classes clearly identified, which are color coded from the identified partitions from the dendrogram. The Figure on the left represents the actual positioning of the dots which leads to overprinting, hence we reprint the names in the same order in the second column for clarity of the actual order. Color code key: light blue = top ranked teams; red = second highest ranked teams; green = lowest ranked teams.

Table 4.13: The Three Significant Vectors of V from Weeks 12 and 14

Source	V1 (12)	V2 (12)	V3 (12)	V1 (14)	V2 (14)	V3 (14)
AP	−0.4874	0.7992	0.3448	−0.4743	0.8639	0.1691
BCS	−0.4913	−0.5835	0.6415	−0.4900	−0.4187	0.7646
Coaches	−0.5075	−0.1163	−0.5750	−0.5172	−0.1978	−0.4398
Harris	−0.5133	−0.08534	−0.3728	−0.5172	−0.1978	−0.4398
Eigenvalues	2716.2	143.5	110.0	2688.4	189.7	121.9

is, the single faction is identified for each week. Although the partitionings for the equivalence classes vary with the weekly rankings, they are easily identified using the first vector of U^* with both the dendrogram and the xy plot, bearing in mind that the xy plot provides the ordering of the classes as well as the meaningful distances between elements and classes.

Weeks 12 and 14 Analyzed Using R-A SVD

Much like week 10, three significant eigenvalues are found for weeks 12 and 14. For these weeks, the first three vectors of V are considered to identify the strength and association of the rank sources. Similar to week 10 (and every other week of these data), V_1 indicates that all four sources show strong agreement by approaching $\sqrt{1/q}$ where $q = 4$, and the signs indicate the agreement in the direction (Table 4.13). In both weeks, V_2 has strong inverse values when comparing the sources for the AP Poll to the BCS rankings. Harris and Coaches both have low values that do not show real support for either AP or BCS; however, both Harris and Coaches have signs that show agreement with the AP Poll. V_3 shows BCS strongly opposing Harris and Coaches. In this instance, the AP Poll moderately agrees with BCS in week 12 and weakly agrees with BCS in week 14. Note that Harris and Coaches have identical rankings for all three vectors in week 14 because their choices for ranking were identical.

To confirm the findings seen with the vectors, the first three principal components are considered in Table 4.14. Upon examining the correlations

Table 4.14: The Correlation of Each Source with the First Three Principal Components During Weeks 12 and 14

Source	PC1 (12)	PC2 (12)	PC3 (12)	PC1 (14)	PC2 (14)	PC3 (14)
AP	−0.928	0.350	0.132	−0.898	0.435	0.068
BCS	−0.935	−0.255	0.246	−0.928	−0.211	0.308
Coaches	−0.966	−0.051	−0.220	−0.979	−0.099	−0.177
Harris	−0.977	−0.037	−0.143	−0.979	−0.099	−0.177

for weeks 12 and 14, there is similar evidence of agreement and strength of correlation with each source within the first principal component. The correlations within the second and third principal component are very weak (below 0.40) for every source. The numbers are too weak to have real meaning but match the results discussed for V_2 and V_3.

Since three significant eigenvalues were identified for weeks 12 and 14, three different graphics can be used to view the feasible equivalence classes: a dendrogram, a scatterplot matrix of the first three vectors of U^* mapped to one another, and a three-dimensional scatterplot. Both dendrograms in Figure 4.18 show similar patterns. Similar to the results seen in week 10, a single straying point (USC) is isolated and two other feasible classes exist. These classes are used to identify the partitioning of the equivalence classes and the elements within each class for the scatterplot matrices of the three significant U^* vectors. When considering the scatterplot matrices for weeks 12 and 14, Figures(4.19 and 4.20) both provide a similar overview of two classes in all cases except when considering U_2^* mapped to U_3^*, which does not have the same clear separation. The green point in both cases represents USC, which was ranked 10 in week 12 and 5 in week 14 by AP Poll was but not ranked by any other source.

The final graphic is a three-dimensional plot that allows a more in-depth view of the partitioned classes. Both weeks show similar results. Overall, the three-dimensional plot allows a better view, especially when working with an interactive tool and not just the snapshots found in Figures 4.21

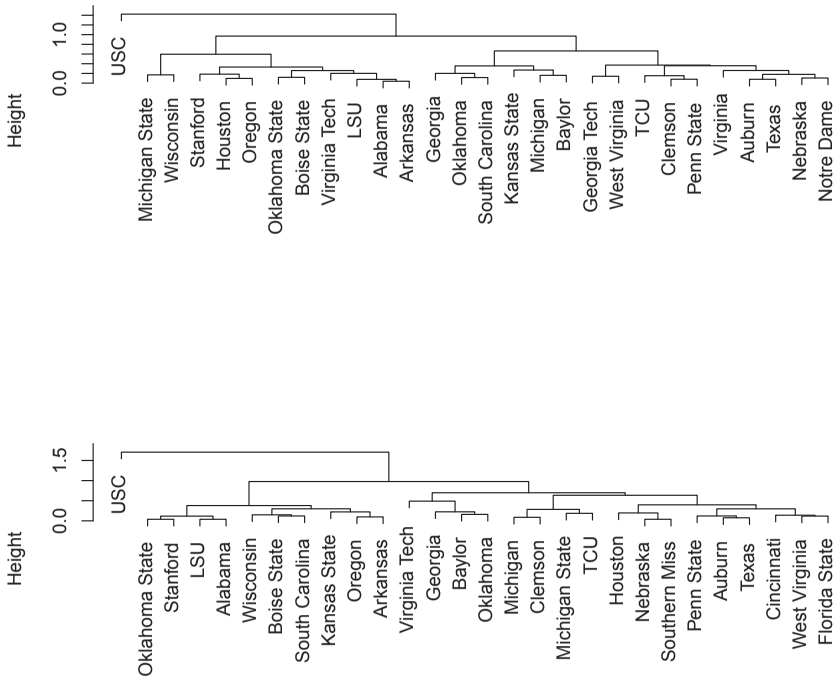

Figure 4.18: Dendrograms of weeks 12 and 14 showing two feasible classes in each and one point that does not fit with the rest (USC).

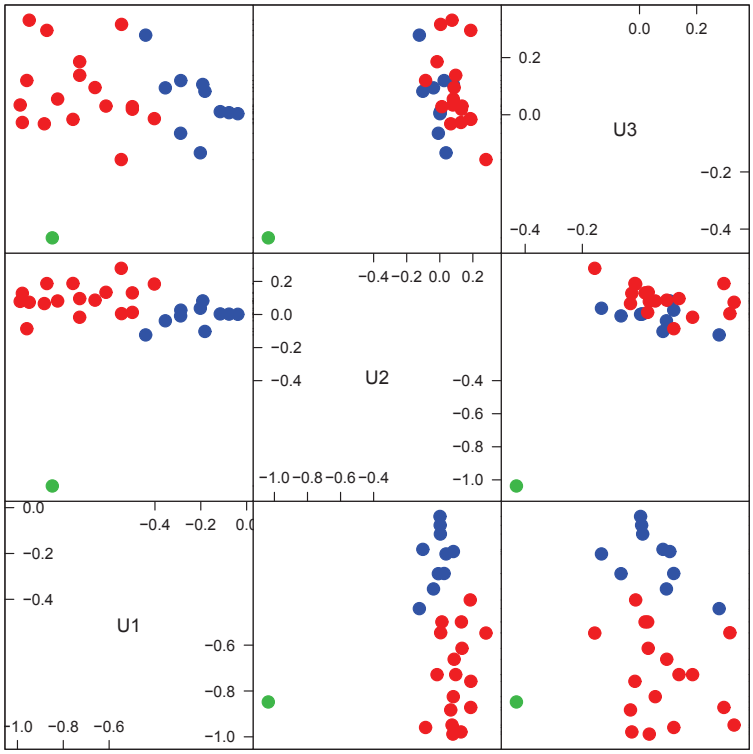

Figure 4.19: Scatterplot matrix of week 12 showing two feasible classes and one point that does not fit with the rest. The blue points represent the top 11 ranked teams excluding USC in green; the red points represent the remaining ranked teams.

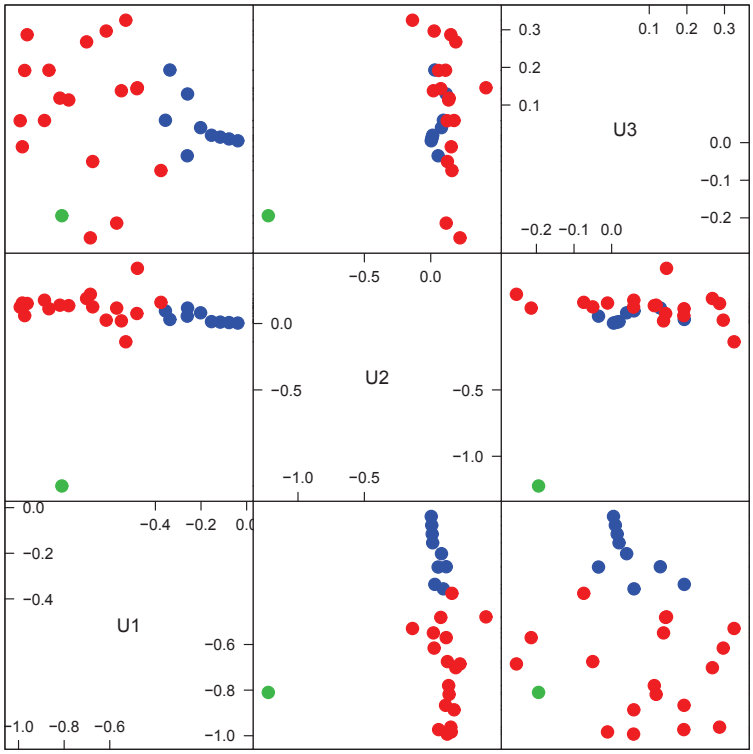

Figure 4.20: Scatterplot matrix of week 14 showing two feasible classes and one point that does not fit with the rest. The blue points represent the top 15 ranked teams excluding USC in green; the red points represent the remaining ranked teams.

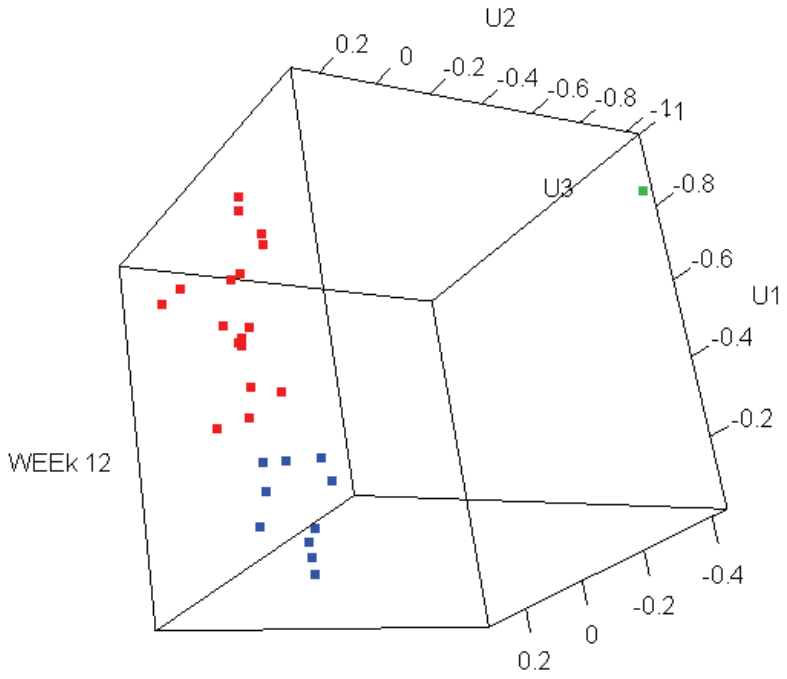

Figure 4.21: Three-dimensional graph of the two feasible equivalence classes found in week 12. The dark blue points represent the top class, while the red points represent the bottom class. The outlier (USC) is green.

and 4.22. In these cases (as with week 10), the blue dots represent the top class, the red dots represent the bottom class and USC (in green) represents the point being identified as ranked highly by only one source. Because this point has become so important, further partitioning of the data becomes unlikely.

To be thorough, *USC* was removed from the data set for both weeks 12 and 14 to see what effect it had on the data. For week 12, the number of significant eigenvalues was reduced to 2 and the number of identifiable equivalence classes becomes four when looking at the dendrogam and xy plot (not shown). For week 14, there is no evident change to the eigenvalues, but three classes are clearly identifiable in all three plots: dendrogram,

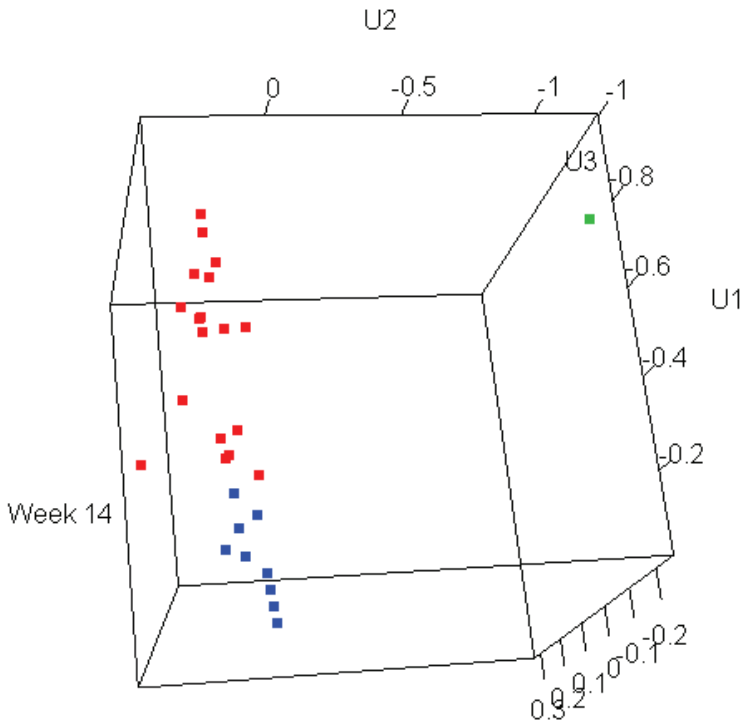

Figure 4.22: Three-dimensional graph of the two feasible equivalence classes and the influential point (USC = green) found in week 14. The dark blue points represent the top class, while the red points represent the bottom class.

scatterplot, and three-dimensional plot (not shown). All of the other results were so similar that no discussion is necessary.

Overall, when three significant eigenvalues are found, the results of the initial faction still exist. But, the relationships of the sources need to be closely considered as they may provide insight to anomalies within the overall ranking such as influential elements. With more significant eigenvalues, a better view of the partitioning becomes clearer as the graphics produced take on more dimensions. The three-dimensional plots created isolate the outlying point and also shows the partitioning of the classes along multiple axes.

4.1.7 Conclusions Drawn from Comparing Sources Using NCAA Ranking Data

Overall, each week had varying relationships with one major common factor, namely that all four sources agreed about the overall association regardless of how many significant eigenvalues were found. R-A SVD makes it possible to gain insight into rank data beginning with correlation, which provides a measure of association. Next, the use of the eigenvalues and eigenvectors allow for deeper analysis of the ranking sources and the ranked elements. Over the eight weeks of collected data, multiple interesting results were uncovered and discussed. Most notably, we observed that multiple eigenvalues were significant (something that presented itself within the analysis). When one or more specific points are highly classified by only a single source, these influential points are easily identified with all the graphics. The R-A SVD method, including its associated graphics, help to identify factions, classes, and the feasibility of influential elements. In future work, we will consider simulating instances with sets of elements with singular rankings in an attempt to identify the overall effect on the R-A SVD method.

CHAPTER 5

ADDITIONAL APPLICATION OF R-A SVD

5.1 NCAA Data Following Individual Sources across the 2011 Football Season

During the 2011 college football season, the AP Poll considered a total of 41 teams when creating its weekly top 25 rankings. Using the raw data from the AP Poll, a scatterplot matrix (Figure 5.1) is created by mapping the weekly ranking to every other week. For the most part, each individual week is very strongly associated with both the previous and next continuous week in time. The only week that varies strongly from this overall time association is the relationship between weeks 4 and 5 and possibly weeks 8 and 9. Notably, the greater amount of time in between weeks being compared, the more varied the relationship is between those weeks, which can be seen by observing the scatterplots heading away from the diagonal.

To better understand these data, we apply the R-A SVD method. The correlation value for $\lambda_{\{R-A\ SVD\}}$ is 0.5638 and for W is 0.662. Anderson's test indicates that this correlation is significant and that three significant eigenvalues can be identified at an α level of 0.05. The first three vectors

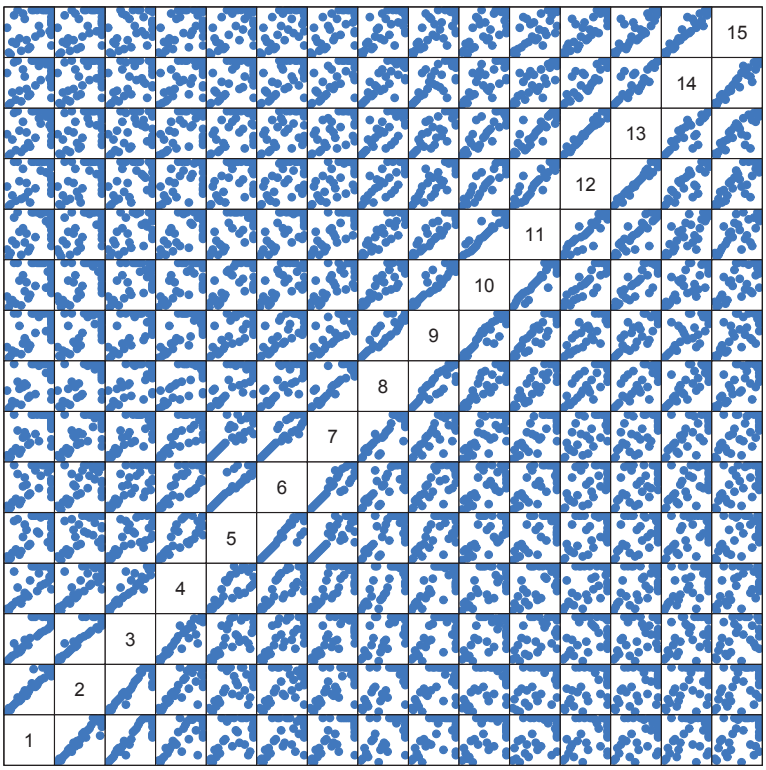

Figure 5.1: Scatterplot matrix of all fifteen weeks as ranked by the AP Poll. This plot is very noisy but shows that the variability of the association spreads out as the weeks move further from one another (evidenced by the scattered points in the boxes moving away from the diagonal).

of V from the factorization are listed in Table 5.1. In column 1, the values for V_1 indicate that all of the sources agree in direction (sign) and approach a value of 0.2582 (as determined by calculating $\sqrt{1/q}$, where $q = 15$ for the 15 weeks of collected data). However, a closer look at V_1 reveals that the first two weeks are slower to approach this value and weeks 3 through 5 are only moderately closer. In column 2, V_2 provides a changing view over time (where time is weeks 1 through 15). In week 1, there is a strong negative value. By week 7, the values approach zero. After this point, the sign changes and the values continue increasing now with positive values up to week 15. Specifically, V_2 shows that the first five weeks are in opposition to the last six weeks of the season, possibly demonstrating the transitions in the rankings (reflecting the effects of wins and losses throughout the football season) over time. For V_3 (using 0.3 as a cutoff value), the most evident inference is that the first two weeks, while in agreement, have some opposing view to the set of weeks (5, 6, and 7) in the middle of the rank season. In order to validate the findings of the three significant vectors of V, the correlations within each of the first three principal components are calculated (Table 5.2). While the first column indicates that all of the sources agree to the overall association, the first two weeks have weaker correlation with values below 0.70. Weeks 3 and 4 are not much stronger, as these values do not exceed 0.73. As time passes, the associations increase in value to more consistent strengths of approximately 0.80 and above. The second column confirms the information found in V_2, namely that the first five weeks form a faction although the values weaken over time. This faction is in complete opposition (as seen by the inverse signs) to the last six weeks of the season, which form a second faction. The third component has very weak values according to the cutoff value of 0.40, but the pattern of the relationships resemble those identified by V_3.

The final step of R-A SVD examines the corresponding three significant vectors of U^*, which will be used to create the associated graphics. Similar to previous cases, the dendrogram (Figure 5.2) is used to determine the

Table 5.1: The First Three Vectors of the AP Poll When Considering the Rankings over the 15 Weeks of the 2011 College Football Season

Week	Vector 1	Vector 2	Vector 3
1	−0.2017	−0.3656	0.3913
2	−0.2124	−0.3902	0.3123
3	−0.2398	−0.3680	0.1681
4	−0.2438	−0.3434	0.0018
5	−0.2422	−0.2549	−0.3845
6	−0.2651	−0.1594	−0.4121
7	−0.2700	−0.0751	−0.3724
8	−0.2762	0.1227	−0.1583
9	−0.2810	0.1523	−0.2074
10	−0.2686	0.2194	−0.1215
11	−0.2764	0.2400	0.1303
12	−0.2653	0.2783	0.1395
13	−0.2769	0.2515	0.2407
14	−0.2749	0.2140	0.1911
15	−0.2625	0.1850	0.2259
Eigenvalues	12506.9986	3177.8279	1630.0243

Table 5.2: Correlations for Each Week with the First Three Principal Components of the AP Poll Considering the Rankings over the 15 weeks of the 2011 College Football Season

Week	Component 1	Component 2	Component 3
1	−0.6030	−0.5508	0.4222
2	−0.6347	−0.5878	0.3369
3	−0.7168	−0.5544	0.1814
4	−0.7288	−0.5173	0.0020
5	−0.7238	−0.3841	−0.4149
6	−0.7923	−0.2402	−0.4447
7	−0.8071	−0.1132	−0.4018
8	−0.8256	0.1849	−0.1708
9	−0.8399	0.2295	−0.2238
10	−0.8029	0.3306	−0.1311
11	−0.8262	0.3616	0.1406
12	−0.7931	0.4194	0.1505
13	−0.8277	0.3789	0.2597
14	−0.8218	0.3224	0.2063
15	−0.7845	0.2787	0.2438

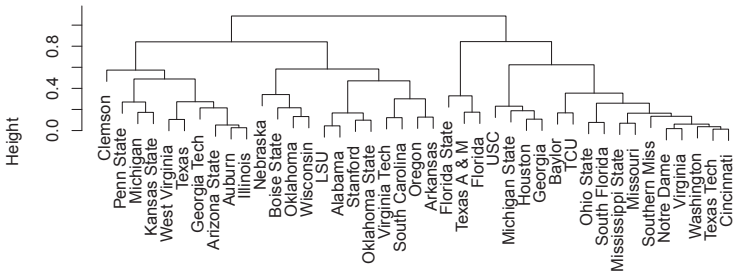

Figure 5.2: A dendrogram created using the first three vectors of U^*. Using a y-axis height between 0.55 and 0.85, four classes are identified.

number of equivalence classes and the items that belong within each class. If a height above 0.85 on the y-axis were used, only two classes would be identified. The next lower height (between 0.55 and 0.85) provides a more likely partitioning. One point that seems to stand alone (with a height close to 0.55) is Clemson, which would have been in a class by itself had a lower height been chosen. For this analysis, four classes are identified and will be considered through the remaining analysis. Of these classes, one is very small, with only three schools (Florida State, Texas A & M, and Florida). These schools all start being ranked highly by the AP Poll but fall out of contention between weeks 6 and 10; they are labeled red on the remaining two graphics.

On the scatterplot of the first three vectors of U^* mapped to one another (Figure 5.3), there is evidence of the four classes identified from the dendrogram (Figure 5.2). However, the partitioning is somewhat messy due to the overlap among the green and light blue points in most comparisons. There is evidence in each graphic that the red and the dark blue points are clearly partitioned from one another and from the remaining classes. As these points examine the teams over time, there is some sensitivity in that a single win or loss can make two points diverge from a previous ranking. On

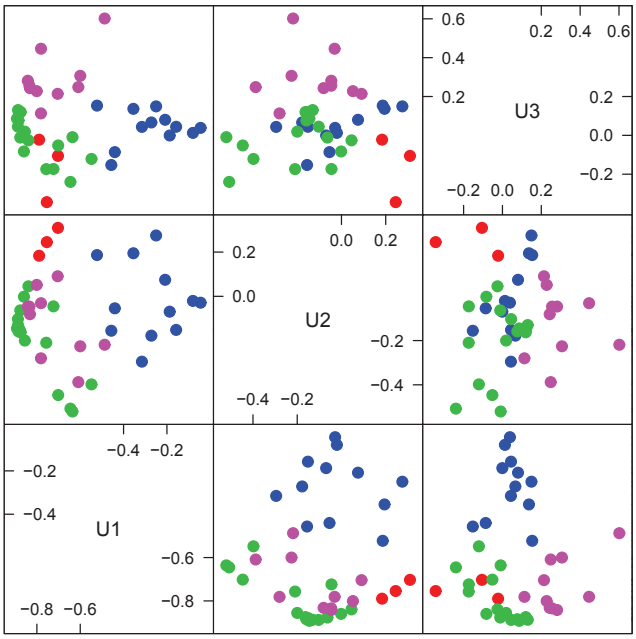

Figure 5.3: Scatterplot matrix of the four classes identified by the dendrogram as viewed within the first three vectors of U^* mapped to one another for the AP Poll data over the 15 weeks for which the data were collected. The dark blue points are the best class (as noted by the lowest ranking for most of the season, there seems to be two middle classes (in red and light blue) as they have many of their rankings in the middle, while the lowest class consists of the green teams as they are ranked above 26 for most weeks and only fall into the ranking for a short time period.

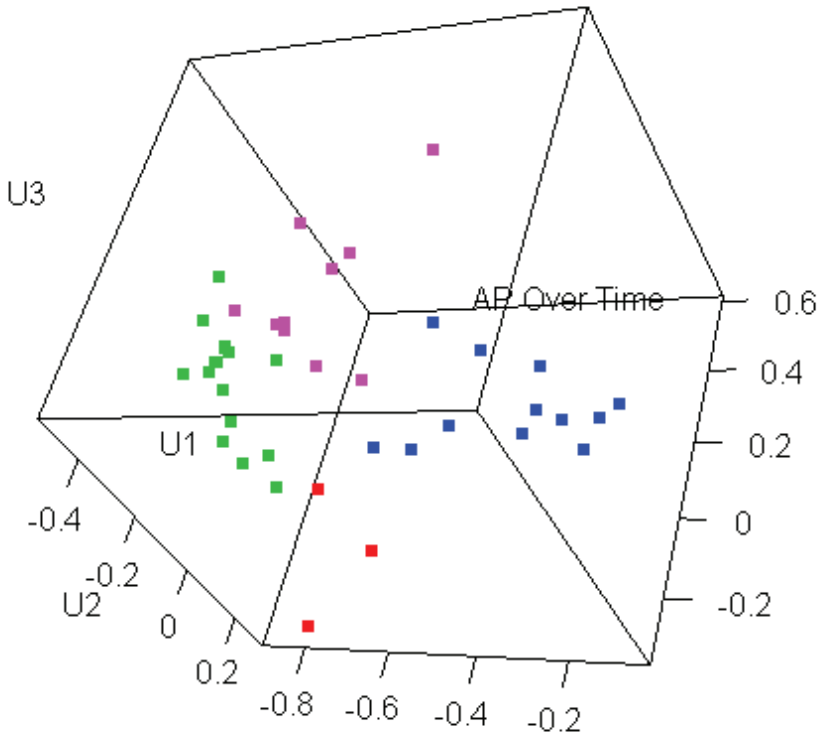

Figure 5.4: A three-dimensional view of the partitioning of the schools as ranked by the AP Poll over the fifteen weeks of the 2011 college football season. Here, the classes are more discernible. The dark blue points are the best class starting the first position on the top left of the graphic. There seems to be two middle classes (in red and light blue) as they have many of their rankings in the middle, while the lowest class consists of the green teams as they are ranked above 26 for most weeks and only fall into the ranking for a short time period.

the three-dimensional plot (Figure 5.4), however, these relationships become evident and the partitions (other than some overlap of the green and light blue points) are easily made. In this graphic, the partitioning supports the dendrogram and provides meaning to the separation of the classes. While it is clear that the dark blue points represent the class containing the best teams, the remaining colors are not as easily classified. Arguably, there seems to be two middle classes (in red and light blue) as they have many of their rankings in the middle, while the lowest class is the green teams as they are ranked above 26 for most weeks and only fall into the ranking for a short time period.

There is an extensive amount of information gained from using the R-A SVD method, which has been discussed throughout this paper. When considering the AP Poll over fifteen weeks, we have identified overall correlation (which is significant), a set of factions, and realistic equivalence classes. While the factions found may indicate that there are two strong groups in agreement with one another and with disagreement between factions (similar to previous scenarios), it is more likely the second vector is indicating a change over time. In order to see what occurs within these factions, we consider what the data looks like over time (Figure 5.5), as this is suggested by the second vector of V.

In order to rank the data, we considered the best team (classified as a rank of one) as the top team during the final week of the season (week 15). This figure starts at the top left box (Alabama) and continuing to the right are the teams ranked 1 through 6 (where a ranking of 1 would put the points in the top of the box of the team being ranked) during the last week of the season by the AP Poll. The row directly below it includes the next 6 best teams as ranked during week 15 of the season. When we look at Florida, Florida State and Texas A & M, there is evidence that these once highly ranked teams dropped from the rankings over time (these lines are in red). Also of note, Clemson has a very varied pattern, which might be why it is almost a class by itself in the dendrogram. Starting with Texas A & M, the

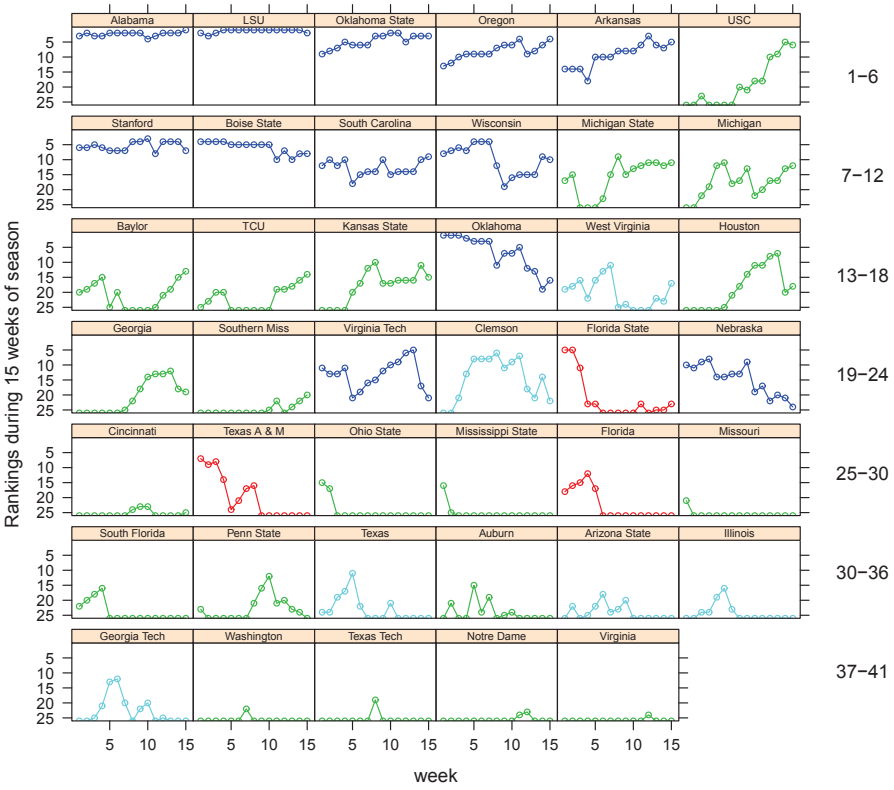

Figure 5.5: A time series plot by team showing the changes over the course of the season as ranked by the AP Poll. The order from top left to top right are the top 6 teams from the final week, the second row are those who were ranked 7 through 12 by the end of the season. The colors of dark blue represent most the highest ranked class (where best ranking starts at the top of each time series). The middle ranked classes are more debatable as there is no clear class second tier in this overall series as it is only looking at the data through the perspective of the rankings during the final week of the season.

remaining points are all teams that were unranked by the end of the season and the ordering was randomly chosen by the program. Although a large portion of these are green, the same is true for many of the middle ranked teams. Most of the teams in the bottom three rows were not originally ranked, made it into the ranking at some point (as evidenced by the small hill), and eventually fell back out of the ranking. Figure 5.5 supplies an interesting view and perspective (especially in light of the pattern found in V_2) suggest that the standard expectation of stationary time series are being observed and clearly identified. For further discussion, more analysis needs to be undertaken in order to include this time series component directly into the R-A SVD method (see Section 6.4).

In order to see any patterns that present themselves from within Figure 5.5, individual plots have been created consisting only of the classes identified from the dendrogram. Figure 5.6 consists of the three teams that were originally ranked as good teams that eventually fell out of the ranking almost entirely midway through the season. (Notice in this graphic the top rankings are a ranking of 26 meaning not considered a contender for the weeks where a flat line is produced above the data). Figure 5.7 identifies the top tiered set of teams for the entire season. These teams were considered to be very strong by the AP Poll throughout the season with little very little fluctuation. Figure 5.8 consists of teams with a decent amount of fluctuation that spent much of the season unranked, although they do fall into the rankings at varying times during the season. Finally, Figure 5.9 represents teams that remained out of the rankings for most of the season but also had fluctuations (some of whom were very variable).

5.1.1 The Remaining Rank Sources Over Time

In this section, we will provide the results found by using R-A SVD to individually analyze the remaining sources ranking the NCAA football data over the time period for which each source provided rankings. Given that

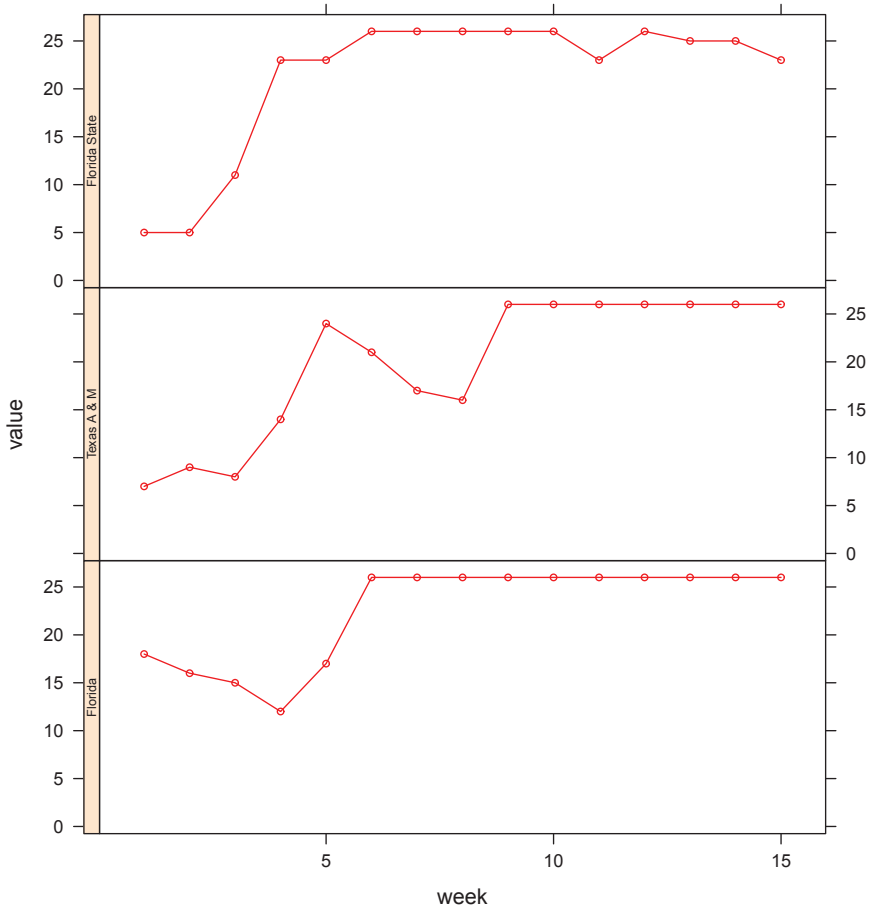

Figure 5.6: The small equivalence class consisting of the three teams (coded in red) to identify if any patterns exist within the class. Notice all three started out very strong and halfway through the season were no longer in the running to be the best team in the NCAA.

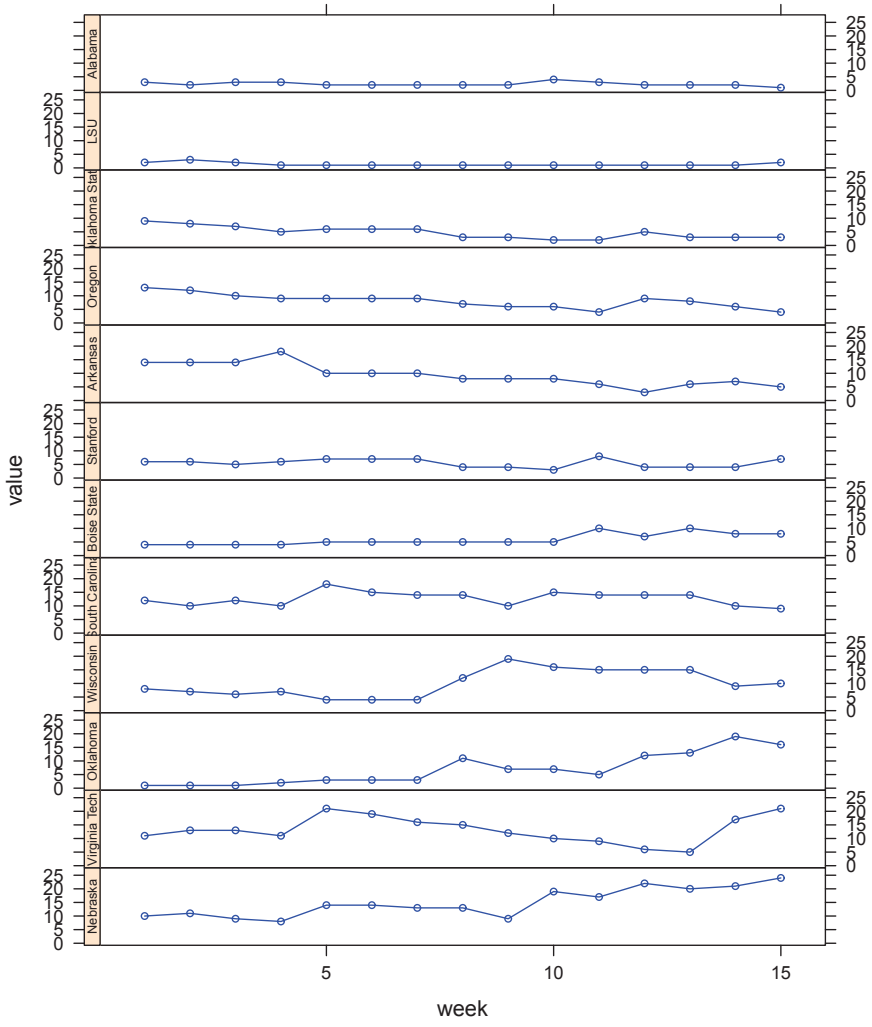

Figure 5.7: The large equivalence class that is identified as the best teams in the NCAA by the end of the season. Each pattern shows relative consistency and low ranking (indicating better performance) throughout the season.

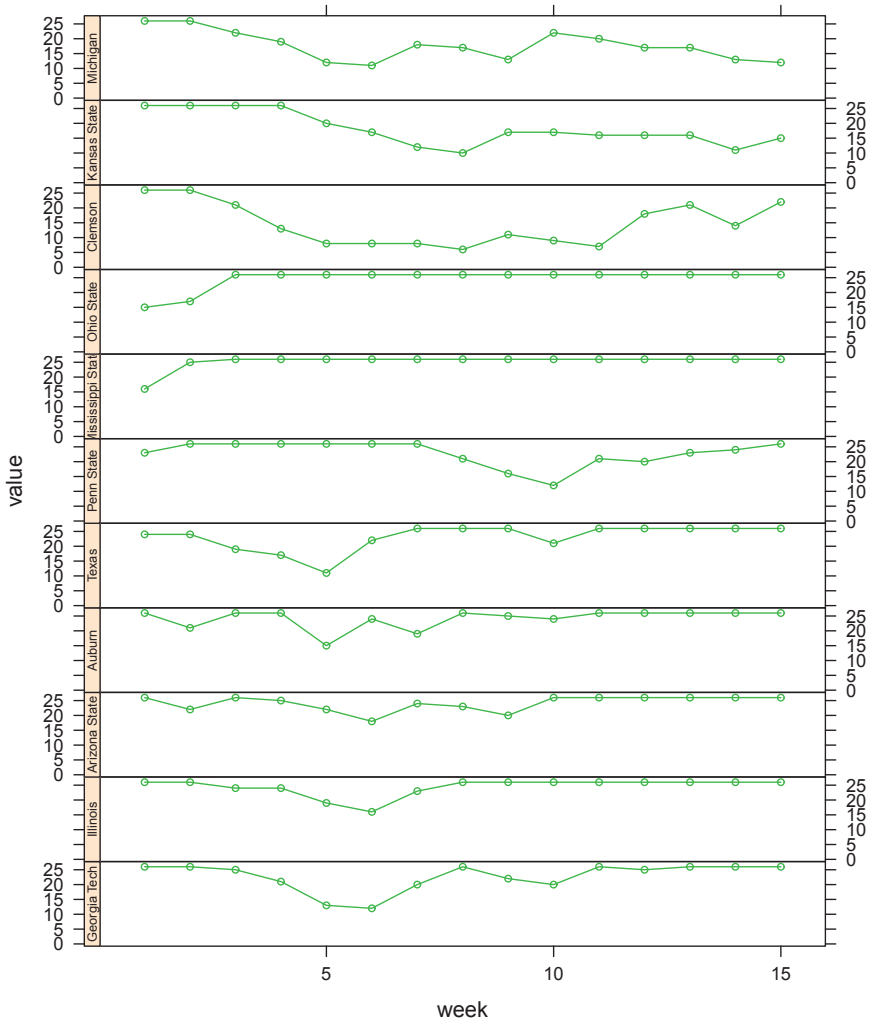

Figure 5.8: Possibly the second lowest ranked class as the green lines all show a similar pattern of remaining out of the rankings for the majority of the season (evidence by the flat line at the top of each box) but being considered as a ranked candidate for some duration of the 2011 season.

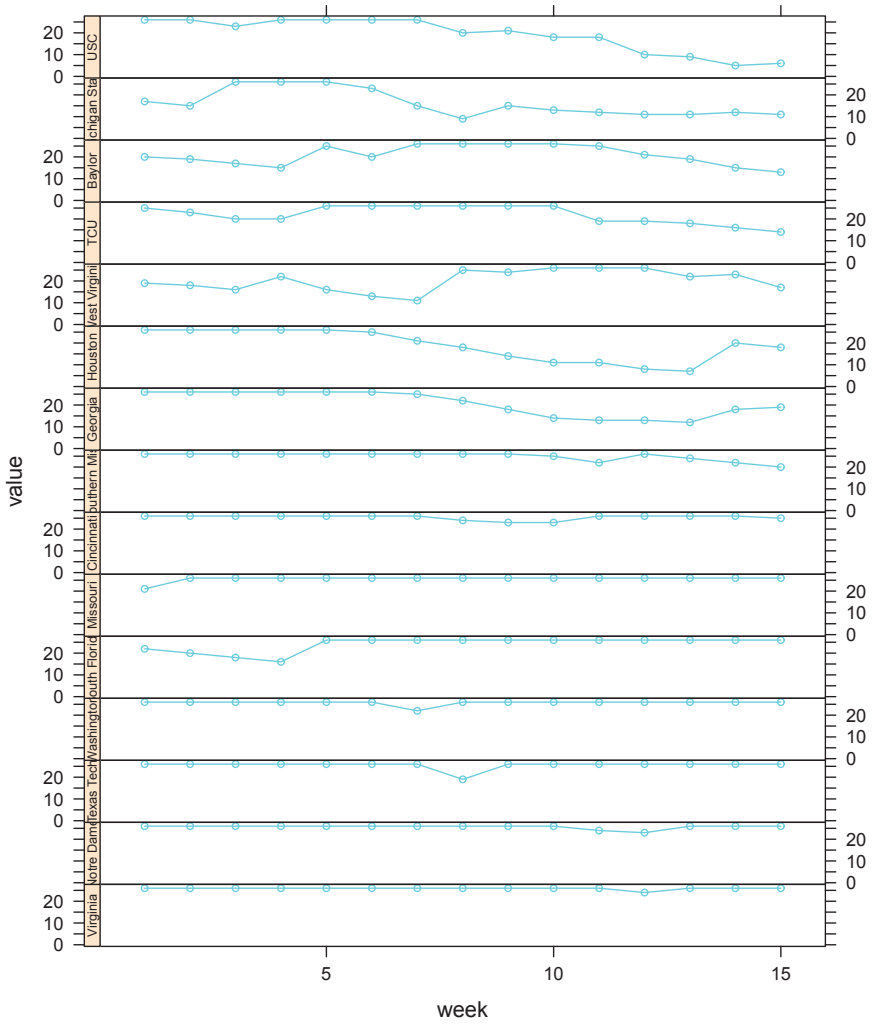

Figure 5.9: Possibly the worst class as the light blue and shows inconsistent patterns though most of the teams are barely ranked during the season.

the three sources (BCS, Coaches, and Harris) did not provide any meaningful insight beyond the patterns already established from the analysis of the AP Poll over the course of fifteen weeks, these discussions will be brief.

Coaches Rank over Time

The output for the Coaches rankings, which also ranked the teams for all fifteen weeks, was identical in nature to that of the AP Poll. The correlation values for $\lambda_{\{R-A\ SVD\}}$ is 0.596 and for W is 0.684. Anderson's test finds this association to be significant at an α level of 0.05 and reveals three significant eigenvalues. The first vectors of V identify an overall association approaching $\sqrt{1/15}$, while the second vectors provide the evidence of the time series trend where the first five weeks form a faction and the last six weeks form a faction. For the third vectors (using 0.3 as a cutoff), the most evident inference is that the first two weeks along with the final week of the season (week 15) are in agreement but have an opposing view to the set of weeks (5, 6, and 7) in the middle of the rank season (similar to a quadratic pattern). The correlations with the principal components confirm the findings of the first three significant vectors of V but to stay concise are not included in this overview of the output.

The final step of R-A SVD considers the first three vectors of U^* for the Coaches data. Starting with the dendrogram (Figure 5.10), there is evidence of four equivalence classes using a $y-$axis between 0.65 and 0.85. As was observed with the AP Poll, Clemson is almost in a class by itself at this cutoff height. The second graphic (Figure 5.11) examines the scatterplot matrix of the first three vectors of U^* mapped to one another. When looking at U_1 mapped to U_2, the elements show clear partitioning. This same set of partitions are mostly visible in U_1 mapped to U_3, with the exception of the red points (Florida, Texas A&M, and Florida State) that also were identified by the AP Poll. U_2 mapped to U_3 does not show any clear partitioning. The three-dimensional view (Figure 5.12) of these vectors can be rotated so that

Table 5.3: The First Three Vectors of the Coaches Poll When Considering the Rankings over the 15 Weeks of the 2011 College Football Season

Week	Vector 1	Vector 2	Vector 3
1	−0.2026	−0.4107	0.2740
2	−0.2122	−0.4487	0.2965
3	−0.2426	−0.3661	0.1726
4	−0.2453	−0.3261	0.0009
5	−0.2553	−0.2135	−0.4013
6	−0.2709	−0.0708	−0.4494
7	−0.2745	−0.0479	−0.3953
8	−0.2751	0.1301	−0.1592
9	−0.2832	0.1639	−0.1464
10	−0.2697	0.2058	−0.0516
11	−0.2655	0.2012	0.1253
12	−0.2671	0.2484	0.1191
13	−0.2615	0.2333	0.2566
14	−0.2771	0.2312	0.2211
15	−0.2557	0.2015	0.3063
Eigenvalues	13568.1007	2817.5923	1378.4974

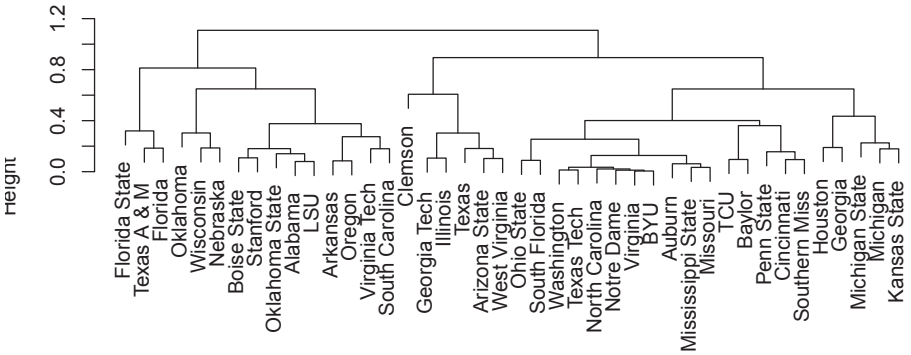

Figure 5.10: Dendrogram of the first three vectors of U^* used to identify equivalence classes for the Coaches data collected over fifteen weeks. A height between 0.65 and 0.85 identifies four classes.

the partitioning of all four classes can be easily seen.

Finally, we consider the data over time. Figure 5.13 is very similar to the time series plot displayed when considering the AP Poll rankings (Figure **??**). Both rank sources once considered the three red teams very highly only to have them disappear from the rankings over time. Also, both graphs show that Clemson is a very volatile team over the course of the season.

5.1.2 BCS and Harris over Shorter Time Periods

The other two sources collected were Harris and BCS. Because the timeframe was shorter and the initial rankings became available later in the season (respectively, 9 and 8 weeks), these two sources ranked fewer teams (36 and 35, respectively). In addition to their shorter ranking timeframes, BCS and Harris produce only two significant eigenvalues when Anderson's test is performed. Because these teams have more data prior to releasing their first rank and a shorter ranking period, they also have slightly stronger correlation values. Harris has a $\lambda_{\{R-A\ SVD\}}$ value of 0.6844 and a W value of 0.817, while BCS has a $\lambda_{\{R-A\ SVD\}}$ value of 0.722 and a W value of 0.854. The vectors for each rank source (Table 5.4) demonstrate the same patterns seen within the first two vectors of both the AP Poll and the Coaches Poll. These patterns indicate an association with the first vector and a time series trend with two identified factions from the second vector. The only difference is that the factions are smaller due to the smaller time span. The correlations with the first two principal components (not shown) concur with the discussion of the vectors. The final step of the R-A SVD process considered the two dendrograms and the two-dimensional xy plot that looks at U_1^* mapped to U_2^* for each source. Again, no new information is garnered from these figures. The dendrograms (not shown) and xy plots (not shown) reveal neither interesting information nor anomalies.

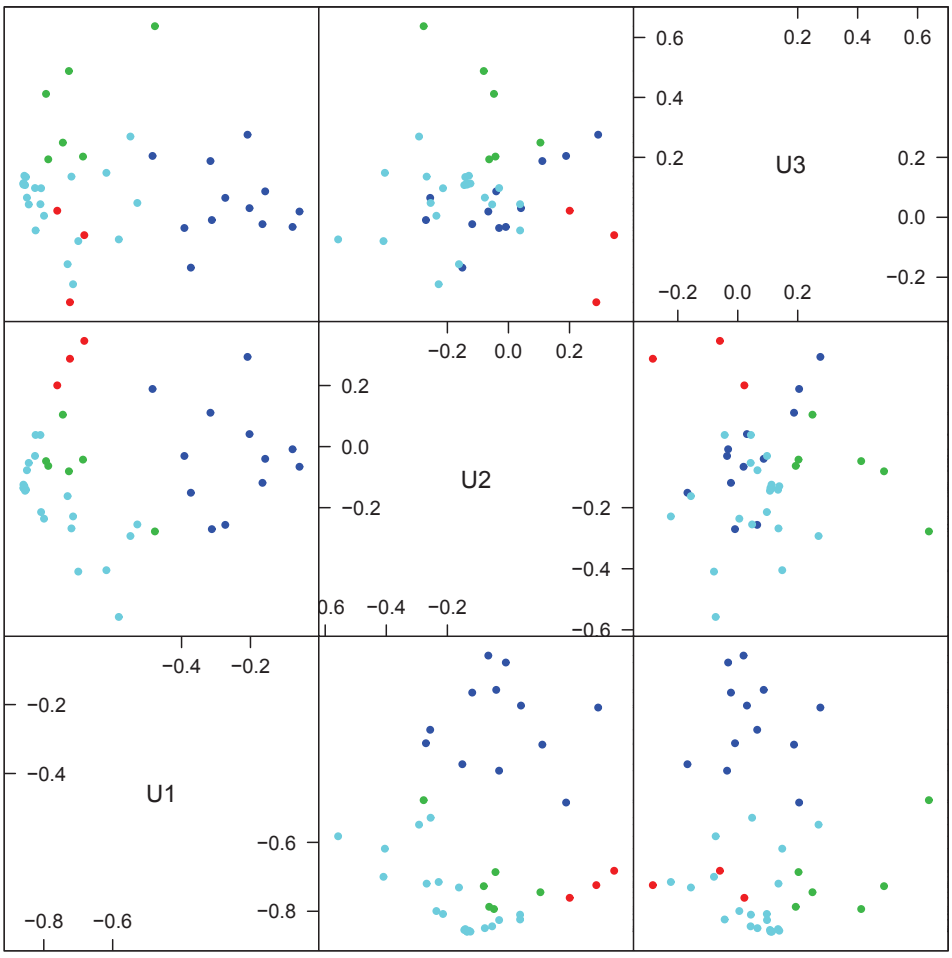

Figure 5.11: Scatterplot matrix of the four classes identified by the dendrogram as viewed within the first three vectors of U^* mapped to one another for the Coaches data over the 15 weeks for which the data were collected. The dark blue points are the best class starting the first position on the top left of the graphic. The remaining classes are not as evident as they hold varying positions during the season.

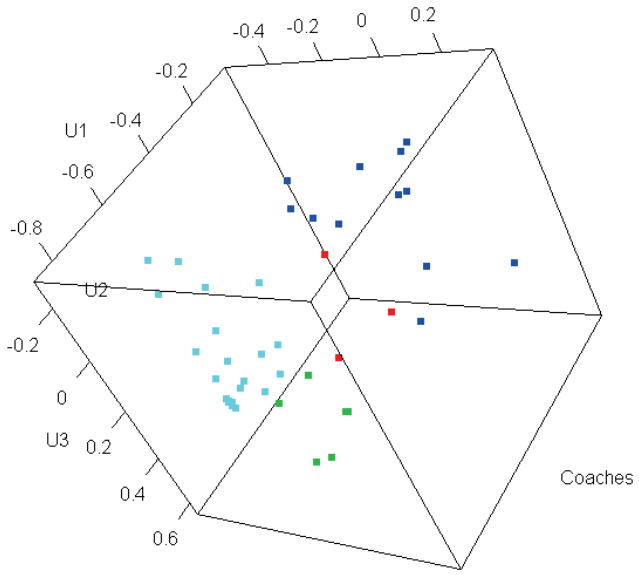

Figure 5.12: Three-dimensional view of the first three vectors of U^* which clearly identify partitioning of the four classes from the Coaches data over the 15 weeks of the 2011 college football season. From the chosen vantage point, the dark blue points are the best class. It appears feasible that the red point may form the second class, although the lowest class is shared between the green and light blue elements.

Table 5.4: The Two Significant Vectors of V from Harris and BCS Rankings over Time

Source	V1 (Harris)	V2 (Harris)	V1 (BCS)	V2 (BCS)
Week 6	−0.2881	0.5436	N/A	N/A
Week 7	−0.3132	0.4973	−0.3215	0.5928
Week 8	−0.3384	0.2297	−0.3449	0.4739
Week 9	−0.3495	0.1516	−0.3644	0.2146
Week 10	−0.3430	−0.1197	−0.3627	−0.0023
Week 11	−0.3538	−0.2304	−0.3675	−0.2132
Week 12	−0.3368	−0.3320	−0.3593	−0.3373
Week 13	−0.3388	−0.3865	−0.3607	−0.3723
Week 14	−0.3334	0.2332	−0.3450	−0.2828
Eigenvalues	6651.17	720.72	7460.06	1124.05

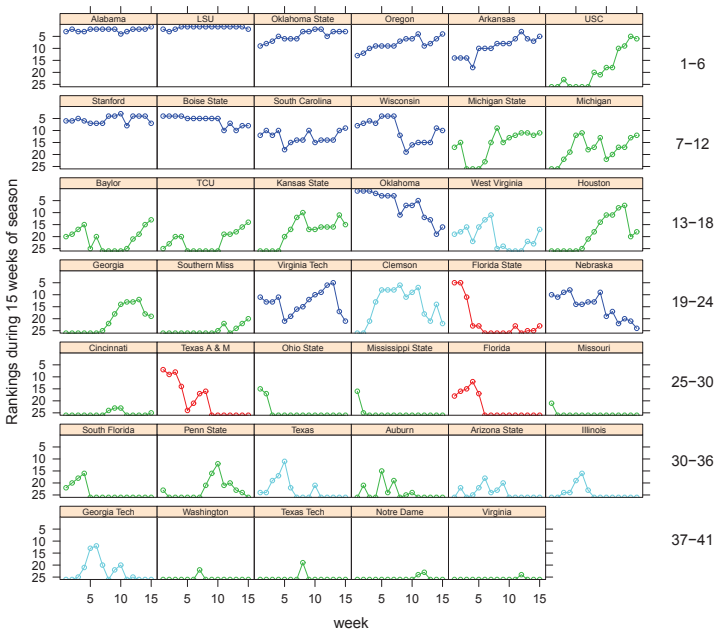

Figure 5.13: A time series plot by team showing the changes over the course of the season as ranked by the Coaches Poll. The order from top left to top right are the top 6 teams from the final week, the second row represents those who ranked 7 through 12 by the end of the season. The dark blue points are the best class starting the first position on the top left of the graphic. The remaining classes are not as evident as they hold varying positions during the season.

5.1.3 Conclusion of Rankings over Time

In addition to comparing ranks, R-A SVD also can be used to identify changes in one source's choice of ranking over time. In all of the cases considered, there was significant correlation and at least two significant eigenvalues found. The second vector in V (for every source) identified a pattern of change over time that was continuous and obvious. Using the significant vectors of U^*, the initial partitioning of the elements into classes can be identified reasonably but not as clearly as in previous cases. There is evidence from the scatterplots matrices of the raw data to suggest a banded correlation method, which is a point of interest for future work. Overall, this set of information exemplifies that R-A SVD is also a good method for identifying changes in rankings over time, something which will be pursued in future work.

5.2 Fifty Greatest Inventions Since the Wheel Data as Discussed by *The Atlantic*

The Atlantic (a literary and cultural magazine founded in 1857) recently assembled a panel composed of twelve scientists, entrepreneurs, engineers, historians of technology, and others to rank their top 25 "innovations that have done the most to shape the nature of modern life" [20]. The compiled findings were released in the November 2013 issue with an accompanying article that discusses the top 50 inventions of the last 6000 years. The individual results were not published but have been granted to us for the purpose of this research. Since much of the information from released ranking data sets typically is undisclosed, this data set provides us with a rare opportunity to apply R-A SVD and then review the underlying data to assess the validity of the method.

In order to complete the task, the panelists were given a list of items

and were instructed that they could add items to the list. Among the twelve panelists, a total of 91 items were given a top 25 ranking. Note that the article claims the top 50 inventions although each source only ranked the top 25. When Anderson's test is performed, all of the eigenvalues are significant. This result most likely is caused by an excessive number of items (39) that were only ranked by a single source. In this case, eleven of the twelve panelists did not consider these items to be significant at all. As a result, there is a large volume of 26 placeholders added into the data. Having observed that a single element being ranked by only one source has an inflating effect on the ranked data in weeks 10, 12, and 14 of the NCAA data set, we recognize that we should consider eliminating these elements so that we may realistically assess the data. In the future, we will need to determine whether this should be considered as an assumption for the initial data analysis.

Since the R-A SVD method was created to compare multiple sources ranking similar elements (here, inventions), these 39 inventions are dropped from the data set. R-A SVD is applied for the remaining 52 inventions. The correlation value of the ranked inventions for $\lambda_{\{R-A\ SVD\}}$ is 0.162 and for W is 0.193. Anderson's test confirm this correlation is significant at an α level of 0.05. Only one significant eigenvalue is found once the data are modified.

Although there were twelve panelists in the initial study, the first vector of V reveals that only nine of them were in any form of agreement as to the ordering of the inventions. Panelists 1, 3 and 8 seem to generate random noise as to the orderings and provide an excessive number of missing points. With nine panelists being considered, the values in the vector should approach $\sqrt{1/9} = 0.3333$. There is evidence that panelists 2, 5, 9, 10, and 11 agree moderately while 4, 6, 7, and 12 agree strongly. This vector indicates that there is one faction but not all of the sources belong within this faction.

To further investigate the agreement, we consider the correlations within the first principal component. Using a lower level cutoff of 0.40, panelists 1, 3, and 8 provide little value to this faction insofar as agreeing on the ranking

Table 5.5: First Vector of V for the Inventions Data.

Source	Vector 1
1	0.0567
2	−0.2996
3	−0.1210
4	−0.3640
5	−0.2601
6	−0.3957
7	−0.3678
8	−0.1748
9	−0.2609
10	−0.2590
11	−0.2695
12	−0.4026
Eigenvalue	7564.2

Table 5.6: The Correlations of Each Panelist with the First Three Principal Components for the Inventions Datal

Source	Principal Component 1
1	0.1044
2	−0.5401
3	−0.3376
4	−0.6025
5	−0.4515
6	−0.6753
7	−0.6004
8	−0.3090
9	−0.4678
10	−0.4584
11	−0.4678
12	−0.6871

of important inventions. Using a range of 0.40 to 0.60 to indicate moderate agreement, panelists 2, 5, 9, 10, and 11 demonstrate moderate agreement within the first principal component. Panelist 4, 6, 7, and 12 agree strongly in both strength and direction (sign) as indicated by these values exceeding 0.60. This correlation also suggests that only nine of the twelve panelists belong within the first faction.

To better understand the results of the correlation, we examine the raw data. The strong agreement group consisted mostly of education and educational journalist, with one highly technological employee. This group focused their rankings on current technological advancements. Of the panelists with moderate correlation, three are historical journalists and four work in the technical industry; two are both. When looking at the weakly correlated panelists, we discover that panelist 3 only chose to rank ten items, which is the likely cause for excluding this individual. Panelist 1 is a historical journalist who provided many entries that predated the advent of the computer. In contrast, panelist 8 works in the computer networking industry and the exclusion of this panelist was somewhat more surprising as the more commonly shared rank elements where technological in nature. It is rare that those disseminating the data would have access to the initial sources to draw these conclusions. These revelations about these sources that belong within the faction are only considered to further discuss their role as the data are readily available.

With the single faction identified, we now consider whether equivalence classes exist within this data set by examining the dendrogram created using the first vector of U^* (Figure 5.14). Choosing a y-axis height below 0.35 prunes the top ranked data into minute classes. Therefore, a value between 0.35 and 0.45 is selected, which indicates that there are three classes (two smaller tighter classes and one larger class). Figure 5.15, an xy plot created by mapping the first vector of U^* to itself, reveals a tight set of inventions in the bottom left corner. As the data move along the diagonal to the upper right corner, the points become more spread. There exists a

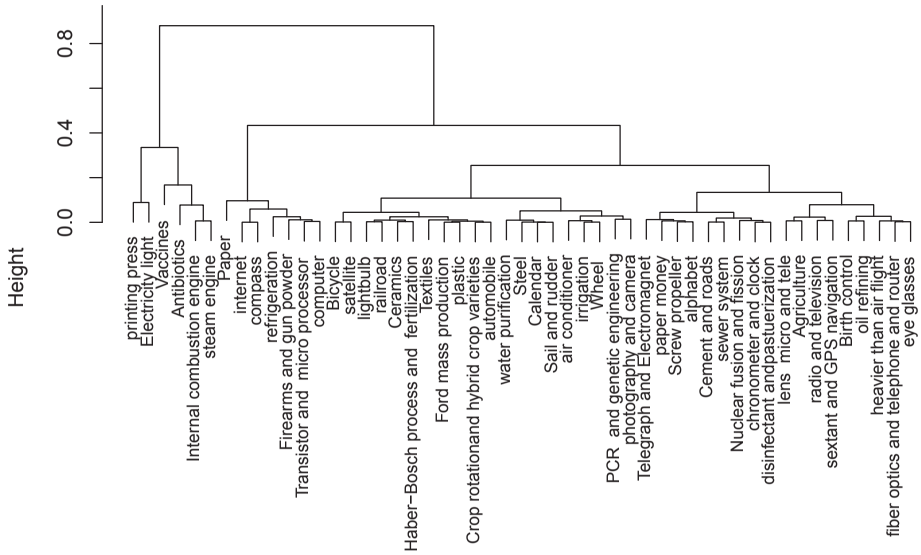

Figure 5.14: A dendrogram of the first vector of U^*, which provides evidence of three feasible equivalence classes using a cutoff height of 0.35 and 0.45.

smaller set of points that could create either one or two equivalence classes, both of which are also identified within the dendrogram. The smallest class identifies six inventions: the printing press, electricity and light, vaccines, antibiotics, combustion engine and steam engine. These six inventions are some of the most common items ranked, and most panelists considered them to be important to some extent. Using the classes identified from the dendrograms (Figure 5.14), the *xy* plot (Figure 5.15) confirms that there is evidence that these partitions are very spread out. In this case, the set of inventions (from paper to computer; second best class) were rated by many of the panelists, whereas the remaining (bottom) class of inventions was ranked by only a handful of panelists. These elements with minimal shared ranking could explain why the bottom (and largest) class forms such a tight cluster separated with so much distance from the next closest class.

Overall, the invention data set provided interesting insights and results. Unlike most ranking data sets released where much of the information is undisclosed, the authors at *The Atlantic* allowed us full access to the files. Once the information was deciphered using R-A SVD, the ability to view the inner working of the relationships that were revealed helped assess the validity of the method. Similar to the the NCAA football rankings, when elements exist that are ranked by a single source, it seems to act as an influential observation that possibly elevates the number of significant eigenvalues. Therefore, the invention data set provides further evidence that an assumption of the R-A SVD method may be that prior to analysis there will need to be elimination of the elements ranked by only a singular source. This assumption will be validated in future work.

5.3 Corporate Sustainability

Studying sustainability, specifically comparative rankings across different venues, led to the initial research for the proposal that evolved into this

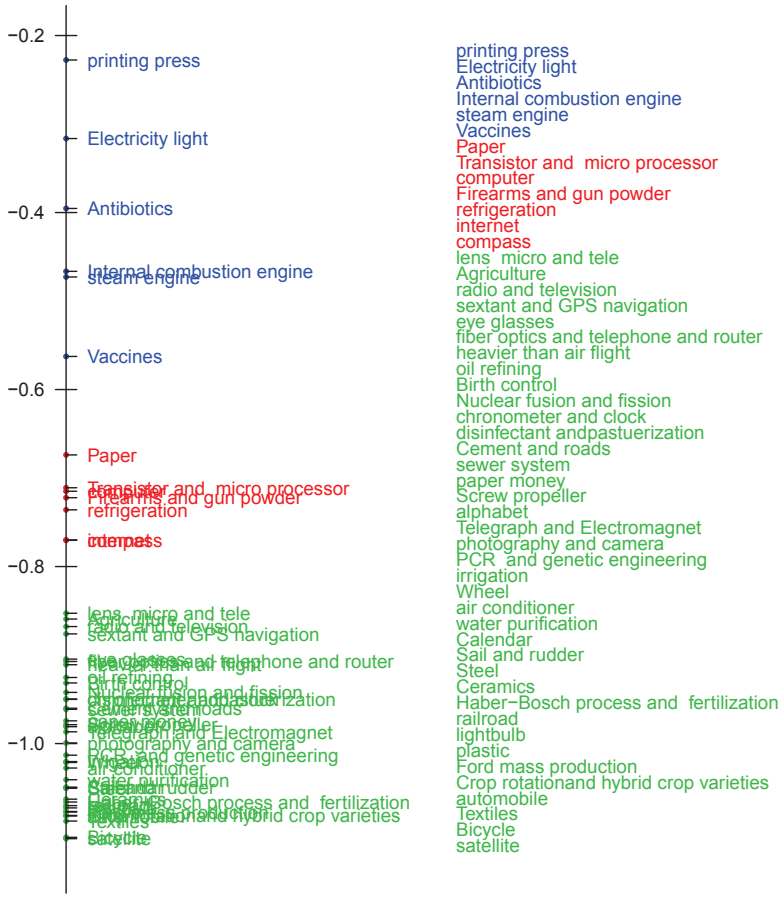

Figure 5.15: An xy plot of the first vector of U^* mapped to itself reveals three feasible equivalence classes in the invention data. The Figure on the left represents the actual positioning of the dots which leads to overprinting, hence we reprint the names in the same order in the second column for clarity of the actual order. Color code key: blue = top class of inventions; red = the second best class green = the bottom class.

Figure 5.16: An xy plot of the first vector of U^* mapped to itself reveals three feasible equivalence classes in the invention data. This is a continuation of the previous plot but the tie-breaker has been broken into two columns (middle and right) for clarity of reading. Color key is the same.

research [12, 13]. Broadly defined, sustainability is the capacity to maintain, support, or endure. Since the late 20*th* century, human sustainability has become congruous to the integration of environmental, economic, and social dimensions towards long-term global stewardship and responsible management of resources. Moreover, those attempting to foster corporate sustainability aim to create long-term consumer and employee value by generating and maintaining a *green* strategy with respect to the natural environment and with consideration to every dimension of how a business operates in the social, cultural, and economic environment. Such a broad and nebulous undertaking leads us to question whether a corporation can really establish and measure this sustainability goal and how well it can be done. Currently, many corporations either claim to be green or imply that they have more sustainable practices than their competition. These claims prompt these questions:

- How well are companies achieving this green initiative?
- Are these companies maintaining the same level of sustainability across the multiple facets of social, cultural, economic, and environmental dimensions?
- How can we tell?

With this in mind, there currently are numerous outside sources gathering information and creating rankings of a top group of sustainable companies. Because this definition of corporate sustainability is so broad, the measuring tools often gather insights about different aspects of the organizations. Here, we must ask whether an overall association of sustainability exists among the sources that are attempting to provide evidence of a truly most sustainable corporation or group of corporations. If this association does exist, our goal is to determine whether classes of good, better, and best overall corporate sustainability can be identified.

The initial study on corporate sustainability looked at four sources, which are named and described as follows:

- Boston College's Corporate Social Responsibility Index (CSRI) considers seven attributes of a company including: performance, products/services, innovation, workplace, governance, citizenship, leadership and performance. Each dimension counts for over twelve percent of the reputation, which is the combined average of the general public's perception along with three key dimensions: citizenship, governance and workplace. This combined score yields greater insight into the influence of stakeholder programs, policies and activities on reputation.

- *Fortune* Magazine's World's Most Admired Companies (herein referred to as *Fortune*) lists the fifty most admired companies overall. The magazine compiles this list from survey data that they send to businesspeople, which asks the recipients to vote for the companies that they admired most from any industry.

- Harris Interactive's Reputation Quotient for the Most Visible Companies (herein called Reputations) asks the general public to measure the reputations of the sixty most visible companies in the United States.

- *Newsweek*'s Green Rankings of America's 500 Largest Corporations (herein referred to as *Newsweek*) attempts to assess the environmental footprint, management of that footprint, and transparency of each of the largest publicly traded companies in the United States and throughout the world.

Previous research identified no real relationship in any pair of rankings considered [12]. Figure 5.17 is a modified xy plot of the raw data that considers all four sources modeled by CSRI. The largest problem with this data set is the excessive number of points gathered along the edges (i.e., a vertical line at $x = 51$ and a horizontal line at $y = 51$), which is indicative of

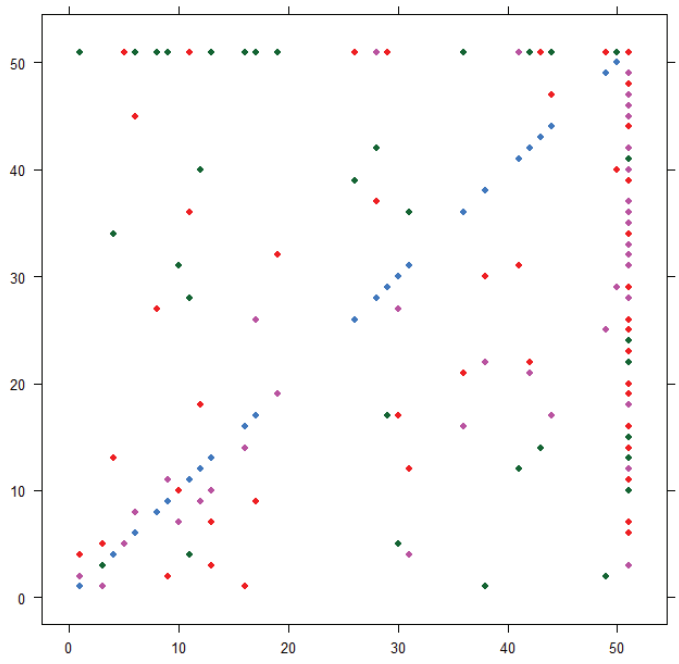

Figure 5.17: A modified *xy* plot for the raw data from the corporate sustainability data set in which all four sources (CSRI, *Newsweek*, Reputations and *Fortune*) are modeled by CSRI. Color code key: pink = Reputations; red = *Fortune*; green = *Newsweek*; blue = CSRI

Table 5.7: Matrix of Correlation by Source for Corporation Data Using Kendall's τ

Source	CSRI	Reputations	*Newsweek*	*Fortune*
CSRI	1.0000	0.2590	−0.0031	0.2022
Reputations	0.2590	1.0000	−0.1451	0.3308
Newsweek	−0.0031	−0.1451	1.0000	−0.0744
Fortune	0.2022	0.3308	−0.0744	1.0000
Eigenvalues	3504.449			

an excessive number of corporations that were not considered by all sources. In these cases, ties were added to the data by inserting a placeholder value 51 (which was not done in the previous study). Using the placeholder value 51 allows the corporation to remain in the data. In contrast, the initial study considered only two sources at a time and Kendall's τ was used only on the ranked data that was shared within each set. Any company that ranked by a singular source was dropped from the data. The final list contained 50 companies (although in the initial study, the data sets for each pair of rank sources were much smaller).

To test these data, the R-A SVD method is applied to the same data set considered by Campbell and Sherman [12] where no individual correlation of data (without ties) exceeded an association value of 0.34 (see Table 5.7). For our current study, ties are added to the data as is noted in the preceding paragraph. The first step of R-A SVD method is to estimate the correlation of the data set, which reveals correlation values of 0.1019 for $\lambda_{[R-A\ SVD]}$ and 0.324 for W. These values vary quite a bit. Next, Anderson's test is considered to confirm that the association is significant and to identify whether factions or equivalence classes exist. As expected with a data set containing such a large amount of ties, Anderson's test fails at an α level of 0.10. In contrast, Kendall's W is significant because it is influenced by the excessive number of ties.

At this point, it would be wise to stop the analysis and concede that there is no association and no faction within the data set, with the most

Table 5.8: The First Vector of V for the Corporate Sustainability Data Set

Source	Vector 1
CSRI	−0.4272
Reputations	−0.6509
Newsweek	0.1764
Fortune	−0.6021

likely reason being that the excessive ties have impacted the data excessively. However, this data set is obtained from various sources that rank the corporations on different scales. Since this case is more similar to a social sciences situation, we decided that allowing for a higher level of α may be appropriate. We determined that the α level for which the data set would have been significant is $\alpha = 0.215$ or greater. In future work, we plan to find appropriate levels of significance for data with an excessive amount of missing elements.

The rest of the steps of R-A SVD are performed to see what information, if any, is provided at this high level of significance. The next step, had we considered this value to be significant, focuses on the first vector of V as only one eigenvalue is significant at $\alpha = 0.25$. On Table 5.8, there is overwhelming evidence that *Newsweek* does not belong with the other sources. It has the opposite sign and approaches zero, indicating that it only adds noise to the analysis. The remaining three sources share agreement on directionality and seem to agree as they approach a value of $\sqrt{1/3} = 0.577$, where 3 represents the number of agreeing sources. Reputations and *Fortune* agree more strongly, although using a cutoff of 0.20 provides evidence that CSRI moderately agrees as well. As a side note, the data were calculated with just these three sources and they were significant at an α level of 0.10. Excluding sources was not a planned part of this study; however, it may be considered to advance the usefulness of R-A SVD for data that have excessive ties.

Using the first vector of V (bearing in mind that it did not pass the initial test) reveals that there is a single feasible faction, and it does not consist of all

Table 5.9: The First Principal Component to Identify the Correlation of Each Source in the Corporate Sustainability Data Set

Source	Component 1
CSRI	−0.5621
Reputations	−0.7966
Newsweek	0.2428
Fortune	−0.7406

four sources but instead only of the three discussed (*Fortune*, Reputations, and CSRI). For consistency, the correlations of each source with the first principal component are also considered (see Table 5.9). Again, there is evidence that *Newsweek* provides no meaningful correlation within the first principal component. CSRI is moderately correlated (between 0.40 and 0.60) with the remaining sources. *Fortune* and Reputation have strong correlation within this component (exceeding the cutoff value of 0.60), and the three main sources also agree on directionality (sign).

Finally, we look at graphics used to identify equivalence classes within the data. A dendrogram (Figure 5.18) created from the first vector of U^* and an xy plot of the first vector of U^* mapped to itself (Figure 5.19) are considered. In Figure 5.18, the best cutoff for partitioning the elements would be at a height on the y-axis between 0.40 and 0.48, which identifies five classes. Note that at this height Google is in a class by itself and no lower height is acceptable for it. The partitioning of these five classes are color coded and presented in the xy plot (Figure 5.19).

In Figure 5.19, there is evidence that Google is the top ranked and most sustainable corporation far above the others corporations. Fed Ex, which is tied with Kraft, has a large separation from General Mills, creating the second partition where the companies labeled in green (Berkshire Hathaway to General Mills) are the next best class. There are some other smaller spaces between WalMart and McDonalds (the fourth class) and between Bank of America and Exxon (the bottom class). Overall, the data represent corporations from around the world and show that there are possible equivalence

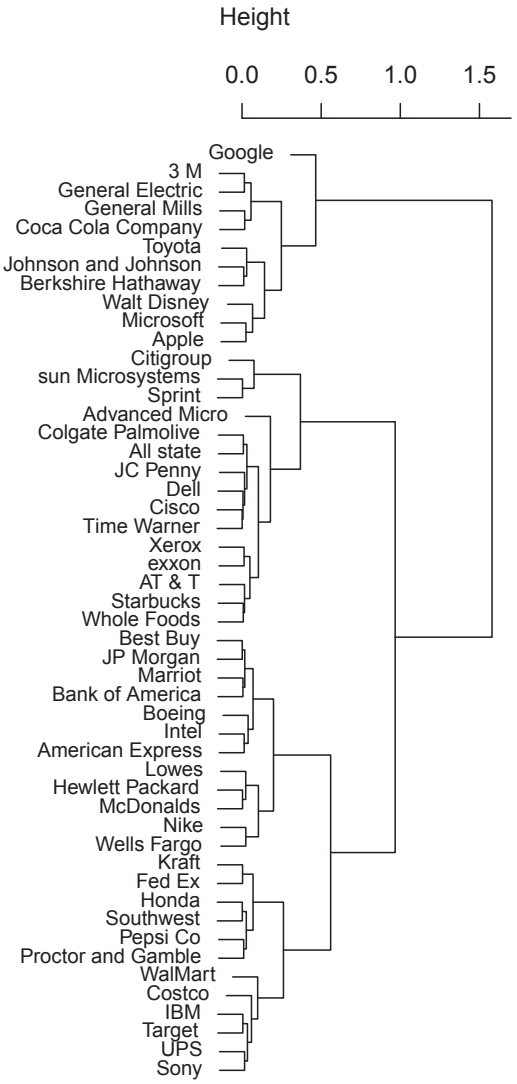

Figure 5.18: Dendrogram of the first vector of U^* demonstrates five feasible equivalence classes for corporate sustainability using a cutoff height between 0.40 and 0.48.

classes when it comes to the most sustainable companies.

Although factions and classes are identified, it is important to note that these data have much weaker association than what is typically seen when the α level is 0.10 significance. Due to the number of missing (non-overlapping) corporations, an excessive number of ties (using the placeholder value of 51) have been added into the data. To be precise, CSRI used the placeholder value a total of 31 times, Reputations used it 14 times, *Fortune* used it 13 times, and *Newsweek* used it 26 times. Note that Reputations and *Fortune* have fewer missing corporations when compared to CSRI and *Newsweek*. However, it is important to recognize the limitations of the test when working with data that comes from such a varied set of sources. In future extensions of the R-A SVD method, establishing baselines for the amount of acceptable missing data will be considered.

5.4 Top 20 M.B.A. Programs in the Northeast Studied from 2006 to 2014

U. S. News & World Reports releases various rankings throughout the year. A common ranking that has a significant effect on academia is the 'best colleges' rankings. Using different classifications ranging from overall schools to business schools, the report has many sets of different itinerary for which they release yearly rankings to help guide consumers in identifying the best college for their needs. This particular data set consists of the top twenty M.B.A. programs in the northeastern region of the United States, and the data were collected from 2006 to 2014.

Using R-A SVD, the correlation values for $\lambda_{\{R-A\ SVD\}}$ value is 0.8103 and for W is 0.893. The scatterplot matrix (Figure 5.21) of the raw data (in which the college rankings are mapped to each other) indicates correlation of each year with one another. However, the strength of the correlation is strongest with the time periods that are closest together and becomes more varied as

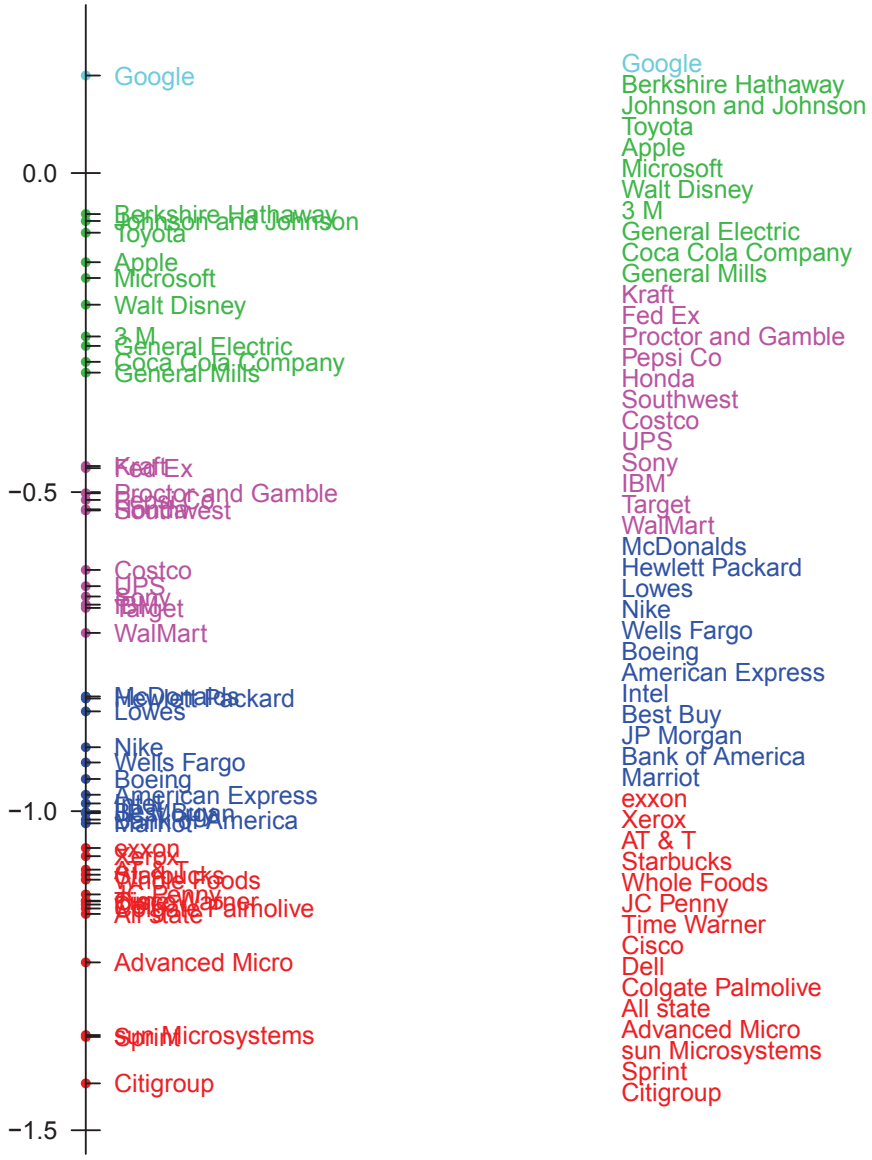

Figure 5.19: The xy plot of the five classes identified by the dendrogram as viewed within the first vector of U^* mapped to itself for the corporate sustainability data set. The Figure on the left represents the actual positioning of the dots which leads to overprinting, hence we reprint the names in the same order in the second column for clarity of the actual order.

Figure 5.20: The same *xy* plot as Figure 5.19 except that the tie-breaker plot on the right is partitioned into two columns (middle and right) for ease of reading.

Table 5.10: First Three Vectors of V for the Best M.B.A Programs in the Northeast Ranking Data

Source	Vector 1	Vector 2	Vector 3
2014	−0.3252	0.3372	0.4294
2013	−0.3288	0.3890	0.4185
2012	−0.3377	0.1797	0.0945
2011	−0.3448	0.1780	0.2367
2010	−0.3268	0.1096	−0.5525
2009	−0.3467	−0.0089	−0.2950
2008	−0.3483	−0.0786	−0.2308
2007	−0.3244	−0.5781	0.0752
2006	−0.3159	−0.5637	0.3525
Eigenvalues	4218.10706	302.02985	196.00423

the amount of time between released rankings increases. When Anderson's test is run, three significant eigenvalues are evident in the data.

Next, the first three vectors of V are considered, and they indicate that association exists as well as identify an overall time series trend similar to those found within the AP Poll NCAA football rankings over 15 weeks. In column 1 of Table 5.10, V_1 identifies that each year approaches $\sqrt{1/9}$ where nine is the number of years under consideration. V_2 identifies only two factions consisting of the first two years (2006 and 2007) and the most recent two years (2013 and 2014). This same vector indicates a linear trend as the values start strong and positive for the top two values and then approach zero by 2010 and then get smaller towards −0.5 by the last row (2006). This data give evidence of a stationary time series trend as would be expected in a time series analysis. Finally, V_3 (using a cutoff of 0.30) provides evidence that years 2014, 2013, and 2006 form one faction that is in opposition (in both sign and strength) to the group 2011, 2010, 2009. This result may indicate a feasible quadratic trend

When considering the number of classes and the elements which belong within each class, the dendrogram in Figure 5.22 is considered. Four classes are identified using a cutoff height of 0.60 on the y-axis.

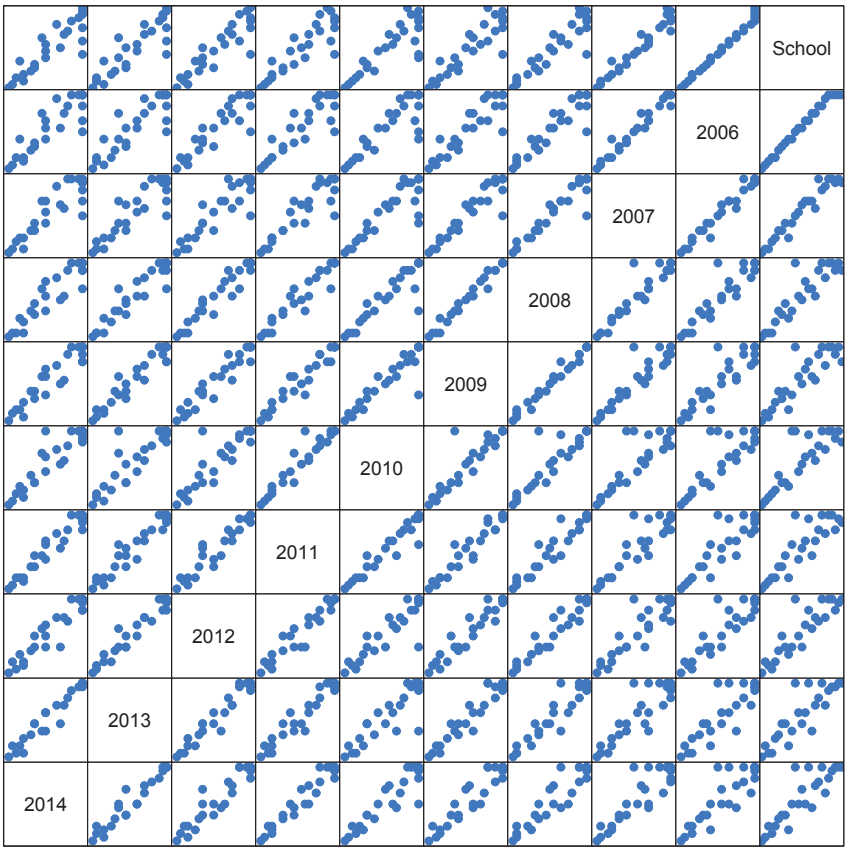

Figure 5.21: Scatterplot matrix of the raw data from the top 20 M.B.A. programs in the northeast region reveals the relationship of the college rankings as mapped to one another starting with the most recent release (2014) and continuing backward to 2006. The boxes closest to the diagonal (bottom left to top right) reveal very strong linear increasing slopes. As the plots move farther from the diagonal, increased variability is evident.

Table 5.11: The Correlation of Each Source within the First Three Principal Components of the Top Twenty M.B.A. Programs in the Northeast from 2006 to 2014

Source	Component 1	Component 2	Component 3
2014	−0.8924	0.2476	0.2540
2013	−0.9007	0.2852	0.2471
2012	−0.9218	0.1312	0.0556
2011	−0.9362	0.1307	−0.1385
2010	−0.8907	0.0799	−0.3246
2009	−0.9430	−0.0065	−0.1730
2008	−0.9542	−0.0576	−0.1363
2007	−0.8810	−0.4200	0.0440
2006	−0.8607	−0.4111	0.2071

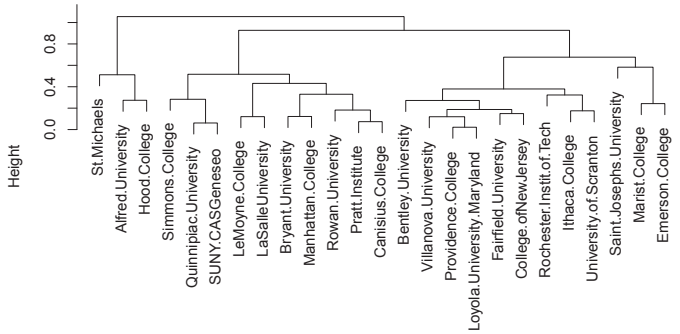

Figure 5.22: Dendrogram of the top twenty M.B.A. programs in the Northeast as ranked by *U. S. News & World Reports*. The height of 0.60 is used to identify the four equivalence classes.

Using the partitioning identified from the dendrogram (Figure 5.22), the classes present themselves in the scatterplot matrix (Figure 5.23) where the top ranked schools (dark blue) clearly separate from the other classes whenever any vector is mapped to U_1. The second best class (light blue) are clearly evident when any other vector is mapped to U_2. Finally, the bottom two classes are most clearly partitioned whenever any vector is mapped to U_3. Although there is some overlap in each picture, initial separations are obvious, and they lead to a very clear three-dimensional view in Figure 5.24. When looking at the three-dimensional view of the data, the classes are easily partitioned with no overlap (Figure 5.24). When this plot is viewed dynamically, the classes also can be shown in order (from best to worst) although some overlap remains in certain instances. However, this figure provides further evidence that the correct items are placed within each class.

R-A SVD was established as a method to help identify association among multiple sources. When at least one eigenvalue is significant, this method further identifies feasible factions and equivalence classes (when they exist). The case of the top twenty M.B.A. programs in the Northeast over a nine-year time frame allows the use of the method from a different perspective. A look at each college ranking over time (Figure 5.25) are considered where the top row has the top five colleges as ranked by *U. S. News & World Reports* in 2014. Starting from the top left, Villanova is ranked 1^{st} in 2014. The next school to the right (Providence College) is ranked 2^{nd} in 2014, with the rank number increasing when looking across the row from left to right. The second row shows the teams ranked 6 through 10 in the 2014 ranking. Each remaining row shows the next five schools as they were ordered in 2014.The best-ranked schools (labeled in dark blue) have lines that travel across the graphic along the top of the y-axis (representing top ranked although the ranking value is actually 1^{st}); these lines show little fluctuation. The second best class (light blue points) has points that remain in the middle for most of the ranking, with some small spikes (upwards) over time. The third best

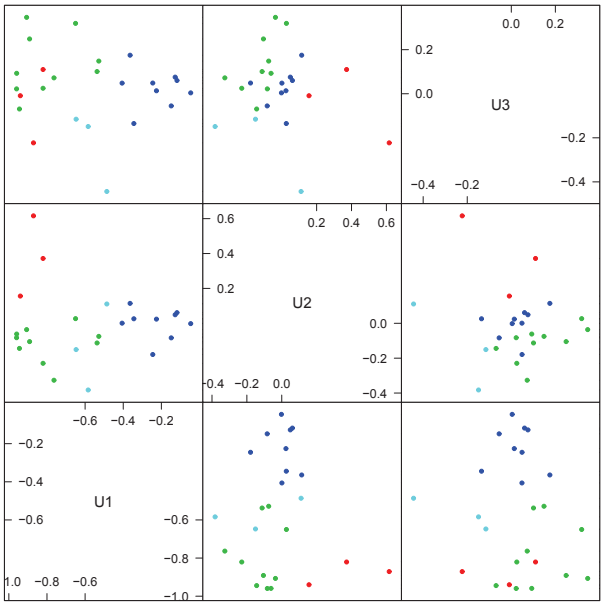

Figure 5.23: Scatterplot matrix of the first three vectors of U^* mapped to one another for the twenty top M.B.A. programs in the Northeast data set. This graphic attempts to show the partitioning of the four classes provided by the dendrogram. The dark blue points represent the top class, the light blue points are the second best, green and red are classes but their order is not as easily classified.

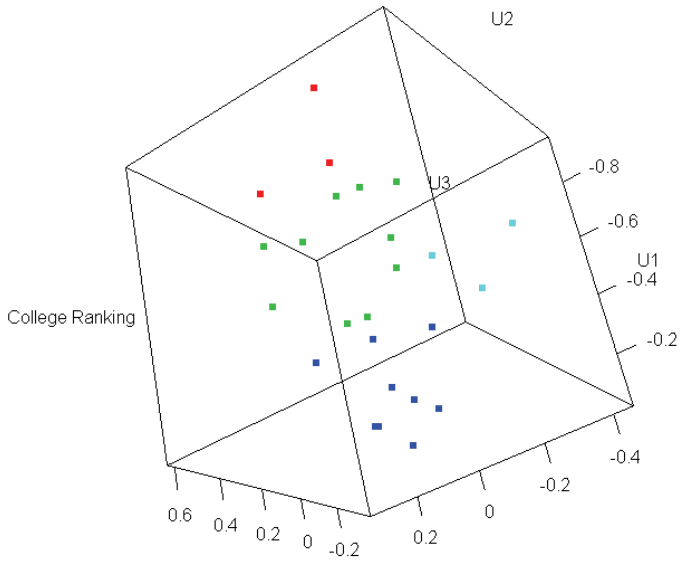

Figure 5.24: A three-dimensional view of the classes identified by the dendrogram for the twenty best M.B.A programs in the Northeast data set. The dark blue points represent the top class, the light blue points are the second best, green and red represent the set of next best classes but the ordering is not as easily determined.

classes (in green and red) present lines that are in higher ranked y-value (between 12 and 20) on the timeline and also has a slightly spiky nature (downwards), although there is slight evidence that the red lines represent the schools that are rarely ranked and found a short-time frame that they were included in the ranking (as is evident by the flat line below the y-value of 20 at the placeholder value 21). Within each box, along the x-axis, the rankings at time 1 represent the rankings for 2014 and descend by one year for each dot to the right until the final dot which represents 2006.

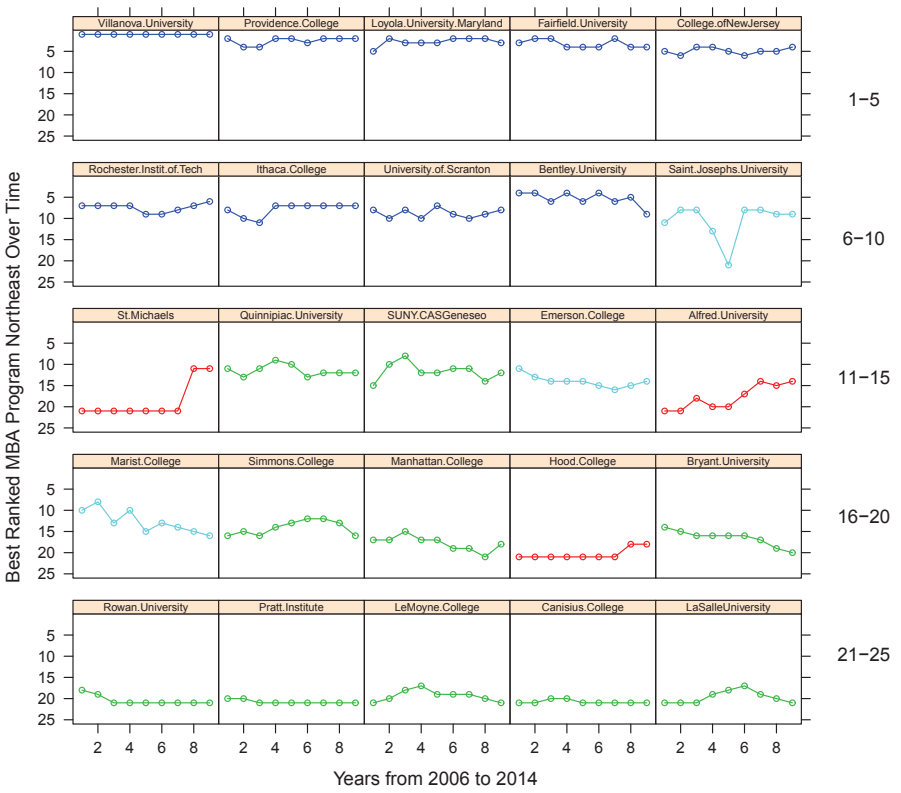

Figure 5.25: Time series graphic representing the top 20 MBA. programs in the Northeast from 2014 continuing backwards to 2006. Villavova, the top ranked school, can be found in the top left box, the second through fifth ranked school follow along the top row (from left to right). Dark blue = top class, light blue = second tiered class, red and green are the remaining classes though which order is not as clear.

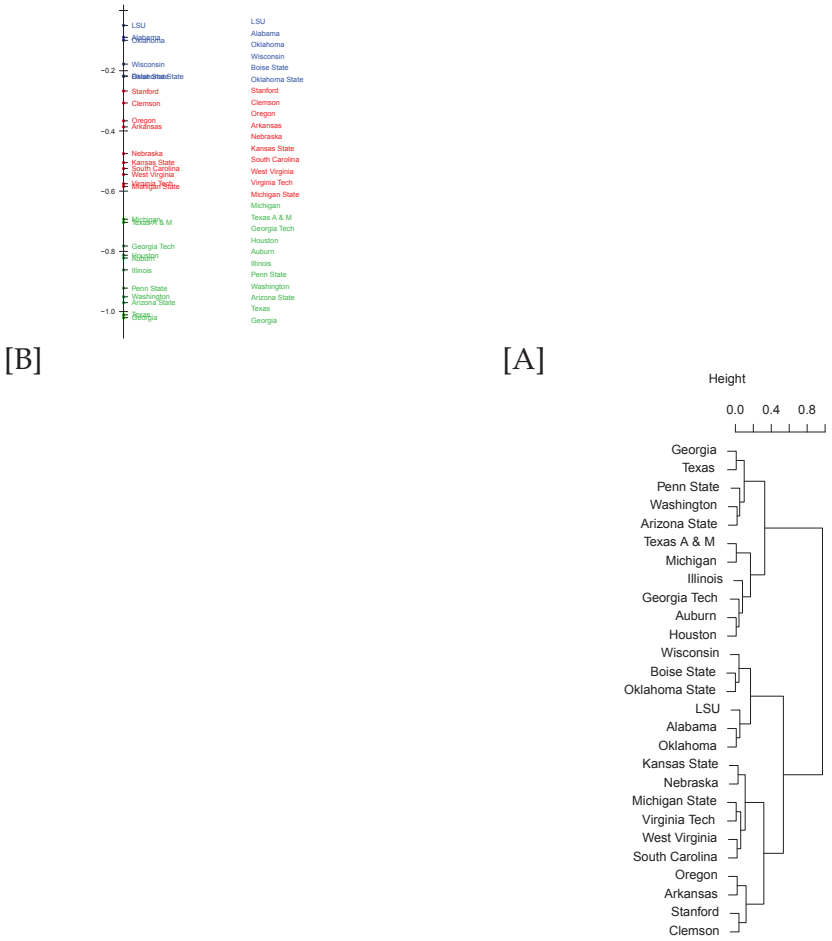

Figure 5.26: [A] Dendrogram for Week 7 rotated to see names clearly. [B] An *xy*plot of week 7 creating by mapping 0f U_1^* to itself. Top class of 6 teams (blue), the next best class of 10 teams (red), while the bottom class of 11 teams (green). Due to overprinting (left) the team names are reprinted in the column (right) using the same order for ease of reading.

CHAPTER 6

CONCLUSION AND FUTURE WORK

In the preceding chapters, we established R-A SVD, a new method for determining overall association. R-A SVD also helps identify the structure of the data by separating agreeing sources into factions and partitioning like elements into equivalence classes. Simulations provided evidence of the method's efficacy in identifying the data structure. Applying the method to real collected data demonstrated the utility of this new tool for ranked data. It quickly became evident how well the method works for data that has identifiable structure, but some anomalies presented themselves and invite further investigation. In this chapter, we address what has been established about the method, limitations that may exist within the assumptions of the test, and what the future direction of our research will be.

6.1 Results of the R-A SVD Method

Many methods have been used within the steps of R-A SVD including: Kendall's τ (see Section 2.1.1); Spearman's footrule, which helped establish rank data being near metric (see Section 2.10); Kaiser's method and the adaptation to it used within this research (see Section 2.3.2); SVD and PCA (see

Section 2.3 and 3.1); and Anderson's test (see Section 3.1.2). Applying this sequence of methods as described in Chapters 2 and 3 provides interesting results and insight into the data structure that have not been previously established for multiple ranking data across multiple sources for small sample sets. This method provides a more conservative replacement to Kendall's W and gives the user a more versatile tool that can assess the data beyond simply determining the correlation measure and whether it is significant.

6.2 Underlying Assumptions Needed for R-A SVD

There are many assumptions that apply to each of the steps and tests used throughout the application of the R-A SVD method. For our purposes, most of these have been met entirely. In Section 2.1.1, the asymptotic normality of τ was shown for $n \geq 10$. Using Fagin et al's [19] distance equivalence class (see Section 2.2.1) allows for Kendall's τ to be considered near-metric, which is a required assumption for the remaining steps in the method. Kaiser's original method (see Section 2.3.1), which uses the largest eigenvalue to determine the correlation, was established for population data. The adaptation to this method (see Section 2.3.2) reduces the bias of the original method and allows for R-A SVD to consider all levels of correlation (even minute) for small samples, such as the majority of the studies discussed in Chapter 4.

The assumptions necessary for the initial application of SVD are that we are attempting to factorize a real or complex matrix. Having established the data as near-metric, we are assuming that the structure of the data allows for applications of this method. From this point, we recognize that eigenvalue decomposition only can be applied to certain classes of square data. We now consider PCA because this approach can be applied to both correlation and covariance matrices as the eigenvectors correspond to principal components and the eigenvalues to the variance explained by the principal components.

After performing these steps to generate the data from the R-A SVD

but prior to their interpretation, it is necessary to gauge the number of significant principal components. When considering correlation matrices in standard PCA, the method for determining the number of significant eigenvalues is directly correlated to the number of eigenvalues that exceed the number one; however, there is no test for significance and applications vary. We wanted a more thorough test that would allow us to separate the eigenvalues into groups where the smallest eigenvalues (starting with the smallest) are equivalent to one another and are considered to be minute enough to be inconsequential. The remaining eigenvalues provide evidence that they are significant, which standard PCA lacks. As Girshick [24, 25] and Anderson [9, 38] demonstrated, the results hold as n becomes large, although how large is not clearly stated. It is important to consider that these are large sample distribution properties (being applied in the test) for the case of m distinct population roots Morrison [35]. As these tests precede the modern explosion of "big data", we allowed for the assumption that, although not huge, our samples of sizes of 20×9 and 25×4 and above were acceptable for the initial applications of the test.

Overall, our method and the data loosely met all of the requirements of the assumptions made within the steps of R-A SVD. As the number of rankings produced and the various data topics being ranked continue to increase and permeate society, establishing a metric application to parse these data and their underlying structure is pertinent for data sets of all sizes. Although we focus on a method that works with smaller data, it could easily be applied to larger data structures as well.

6.3 Future Work to Improve R-A SVD

One major adaptation to the initial data structure is the addition of a value of $k + 1$ where any element is unranked by a source (see Section 2.2.2). The use of a placeholder value prevents multiple source rankings from reducing

to very small sample sizes. Recall that previous tests would eliminate any element that was not considered by all of the sources. While analyzing these data, it appeared that this adaptation had two effects on the data that need to be considered in future assumptions. The first of these effects is observed when only one source highly ranks an element and no other sources rank this element. In this scenario, it is possible that this element could increase the number of significant eigenvalues identified using R-A SVD. Future research will attempt to identify the effects that points like this may have on the data. Second, it is highly likely that the presence of too many ties in the data can adversely effect the testing procedure to the point where no significance can be found at reasonable levels of α. Unlike W, which increases with the amount of ties, this method is less likely to identify significance if an excessive amount of ties exist. Future research will look at various sample sizes and the how the proportion of placeholders effect the eigenvalues and the significance level needed. This investigation will be essential for applying R-A SVD to social science data sets or other data sets where the sources are ranking similar objects but from varying viewpoints, as was the case with the corporate sustainability data set in Chapter 4.

6.4 Future Work Expanding R-A SVD for Other Applications

When reviewing the different rank scenarios, we examined both the NCAA football ranking data over the weeks for which each source ranked the data and the *U. S. News & World Report's*Report's top twenty M.B.A. programs in the Northeast ranking data from 2006 through 2014. R-A SVD was used to ascertain whether the association over time was correlated and to determine what other information could be gained from the vectors and values of the method. The method worked well, and multiple significant eigenvalues were identified for each analysis. The vector V identified the

overall correlation in V_1 as well as identifying a linear trend in V_2, which identifies a correlation pattern of change over time. V_3 also provides a quadratic trend, which is not surprising considerin that V_2 is identifying this change over time. Although we assumed that these trends existed since the data provided a realistic example of stationary data, we did not expect the method to identify it so clearly. In future studies, we hope to apply the method to longer time periods and to less organized data to see what other information can be gained and identified using R-A SVD.

6.5 Conclusion

Non-parametric data analysis has a broad range of applications in statistics. Rank order data (specifically for smaller sample sizes) has had little new metrics or modern application since these methods were established by Kendall in 1939 [**?**]. The original methods are valid for the data they were intended to test. However, changes in the types of data to which they can be applied and the advances of technology allow for advances in the original techniques. Many of the initial ideas explored in this research were topics posed by Kendall. In our attempt to address Kendall's questions and concerns, we also have identified a method that answers the question of how to identify the underlying correlation and its significance when factions exist with opposing views as well as allows one to understand the structure of the original ranked order data and feasible equivalence classes.

REFERENCES

[1] Bloomberg l. p., 2014. https://www.bloomberg.com/.

[2] Consumer reports, 2014. https://www.consumerreports.org/.

[3] Edmund's, 2014. http://www.edmunds.com/.

[4] J. d. power & associates, 2014. http://www.jdpower.com/.

[5] Kelly's blue book, 2014. http://www.kbb.com/.

[6] Newsweek, 2014. http://www.newsweek.org/.

[7] Statistical dictionary, 2014. http://dictionary.reference.com/browse/correlation.

[8] Michael L. Anderson. The benefits of college athletic success: An application of the propensity score design with instrumental variables. *National Bureau of Economic Research Working Paper No. 18196*, 2012.

[9] T. W. Anderson. Asymptotic theory for pinciple component analysis. *Annals of Mathematical Statistics*, 34:122–148, 1963.

[10] J. Bar-Ilan. Comparing rankings of search results on the web. *Information Processing and Management*, 41:1511–1519, 2005.

[11] M. S. Bartlett. A note on multiplying factors for various chi-squared approximations. *Journal of the Royal Statistical Socitey*, 16:296–298, 1954.

[12] Kathleen T. Campbell and W. Richard Sherman. Lists & more lists: Making sense of corporate reputations. *Journal Of Business & Economics Research*, 8, No. 7, 2010.

[13] Kathleen T. Campbell and W. Richard Sherman. The most sustainable car. *Academy of Business Research Journal*, 1, 2012.

[14] Ron Coddington and Alex Richards. 30 ways to rate a college. *The Chronicle of Higher Education*, 2010.

[15] Tim Crosby. How inventory management systems work, July 2007. `http://money.howstuffworks.com/how-inventory-management-systems-work.htm`.

[16] Sanjoy Dasgupta and Philip M. Lon. Performance guarantess for hierarchical clustering. *Elsevier Science*, 2010.

[17] W. Dunlap, C. Silver, and G. Phelps. A monte carlo study of using the first eigenvalue for averaging intercorrelations. *Educational and Psychological Measurement*, 47:917–923, 1987.

[18] B Everitt and T Hothorn. *An Introduction to Applied Multivariate Analysis with R (Use R!)*. Springer, 2011.

[19] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top *k* lists. *Society for Industrial and Applied Mathematics*, 17:134–160, 2003.

[20] James Fallows. The 50 greatest breakthroughs since the wheel. *The Atlantic*, 17:134–160, 2003.

[21] Forbes. "special report the forbes top 100 best buy colleges 2012", 2010. `http://www.esf.edu/welcome/Forbes_Best_Buys_2012.pdf`.

[22] S. A. Freidman and H. Weisberg. Interpretting the first eigenvalue of a correlation matrix. *Educational and Psychological Measurement*, 41:11–21, 1981.

[23] D. L. Fulks. Revenues and expenses 2004-2010: Ncaa division i intercollegiate athletics programs report. *Technical Report, National Collegiate Athletic Association, Indianapolis, IN*, 2011.

[24] M. A. Girshick. Principal components. *Journal of the American Statistical Association*, 31:519–528, 1936.

[25] M. A. Girshick. On the sampling theory of the roots of determinantal equations. *Annals of Mathematical Statistics*, 10:203–224, 1939.

[26] Google. Google information for webmasters, 2004. Available from `http://www.google.com/webmasters/4.html`.

[27] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:498–519, 1933.

[28] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, fifth edition, 2002.

[29] H. Kaiser. A measure of the average intercorrelation. *Educational and Psychological Measurement*, 1968.

[30] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30:81, 1938.

[31] Maurice G. Kendall. *Rank Correlation Methods*. Charles Griffin and Company Limited, 1948.

[32] D. N. Lawley. A modified method of estimation in factor analysis and some large sample results. *Uppsala Symposium of Psychological FActor Analysis*, pages 35–42, 1956.

[33] D. N. Lawley. On testing a set of correlation coefficients for equality. *Annals of Mathematical Statistics*, 34:149–151, 1963.

[34] James Monks and Ronald G. Ehrenberg. U.s. news and world report's college rankings: Why do they matter. *the National Bureau of Economic Research*, 31:42–15, 1999.

[35] Donal F. Morrison. *Multivariate Statistical Methods*. McGraw-Hill, 1976.

[36] U.S. News and World Reports. Princeton, williams top u.s. news best colleges rankings, 2014. `http://www.usnews.com/education/best-colleges/articles/2013/09/10/princeton-williams-top-us-news-best-colleges-rankings`.

[37] Pranab K. Quade, Dana Sen and Ibrahim A. Salama. Spearman's footrule: Asymptotics in applications. *Chilean Journal of Statistics*, 2:3–20, 2011.

[38] Second Berkeley Symposium on Mathematical Statistics and Probability. *The asymptotic distribution of certain characteristic roots and vectors.*, 1951.

[39] David Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press, 1997.

[40] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15:88, 1904.