

Rank Analysis in the Evaluation of Stock Forecasts:
Hypothesis Testing on Predicted Rankings Assuming
Distance-Based Models

Yu-Jay Huoh

July 22, 2008

Contents

1	Introduction	3
2	Background Information	3
2.1	Distances	7
2.1.1	Spearman's Distances	7
2.1.2	Kendall Distance	8
2.1.3	Hamming Distance	9
2.1.4	Cayley Distance	9
3	Models	9
3.1	Babington Smith Models	10
3.2	Thurstonian Models	11
3.3	Mallows / Distance-Based Models	12
4	Model Fitting	13
4.1	Assessing Goodness of Fit	14
5	Testing	15
6	Application to Stock Forecasts	16
7	Empirical Study	18
8	Future Work	21
9	References	21

1 Introduction

In the half century that has passed since the introduction of the Markowitz Bullet and efficient frontier, the financial world has given rise to many statistical challenges. One challenge that most financial institutions have yet to conquer is developing methods that accurately forecast stock returns. Stock returns, like most forms of financial data, generally contain a high amount of noise and fluctuation, making methods that produce insightful forecasts hard to come by. This noise that makes forecasting difficult also introduces difficulties into assessing the effectiveness of predictions. Traditional methods for measuring closeness and prediction accuracy, such as correlation or mean squared prediction error, are not very useful when evaluating forecasts. Too high of a correlation often results in overfitting. Yet most prediction schemes used in practice produce correlations below .10, which is not very indicative of the true predictive power contained within the forecasts.

In this paper, a method will be developed to determine the effectiveness of forecasting methods based on the predicted rankings of stocks. Instead of coming up with a statistic to summarize how effectively a method can forecast, this method attempts to determine whether the ranks of the forecasted stock returns are significantly close to those of the actual stock returns through hypothesis testing. As with most methods based on ranks, the impact of the noise inherent in the data is mitigated by using the ranks of the forecasted actual returns. One drawback of this method is that it does not yield a numerical value to compare different sets of forecasts. While a numerical value that summarizes performance has its uses, the amount of noise present in all financial data inhibits the ability to make meaningful assessments based on single values.

This method will be applied on monthly forecasts - kindly provided by a large financial services company - for stocks in the Russell 1000 Index, which is an index containing the thousand largest companies being traded in the United States based on the number of shares available for trading. Forecasts and actual returns are available at the end of each month starting from January 31st, 1996 to December 31st, 2006, giving a total of 120 months or data points. The Russell 1000 Index was chosen because the stocks on this index are less volatile due to the high volume of shares being traded. Month-to-month returns are studied because these contain fewer random fluctuations than daily returns. Annual returns are not used because there is no way to get a large collection of data points - few stock forecasts are available before the 1970s.

2 Background Information

The goal of this paper is to develop a method to test for whether a set of forecasted stock returns accurately predicts actual stock returns. Suppose there is a group of 100 stocks that one would like to be able to forecast. Any method of forecasting stock returns is implicitly predicting the rankings of the stock returns at the same time. After all, if the forecasted returns are to be similar to the actual returns, one would expect that the order of the companies based on forecasted returns would be close to the order of the companies based on the actual returns. Naturally, one way of deciding whether or not a set of stock forecasts is effective is to test if the forecasts have

this property whether the forecasted rankings are similar to the actual rankings. The goal of this paper is to develop enough of an understanding about permutations and probability models on permutations in order to conduct hypothesis tests for this property.

While rankings are prevalent in our everyday life, people rarely spend time doing formal mathematical or statistical work with them. Anyone that has read academic literature on rankings or permutations is familiar with the confusion that often arises when working with these objects. Thus, throughout this paper, every attempt will be made to prevent such confusion in the reader.

To begin, consider a situation where two people, Alice and Bob, are trying to rank four fruits in order of their preferences: apples, bananas, oranges, and pears. Alice likes bananas the most, followed by oranges, pears, and apples. Bob likes oranges the most, followed by pears, bananas, apples. One could write the two rankings by listing them in order of preference. Throughout this paper, the most preferred item will always be the left-most or first-listed item. The right-most or last-listed item will always have the lowest rank and least preference. Alice's ranking would be (bananas, oranges, pears, apples) and Bob's ranking would be (oranges, pears, bananas, apples).

However, listing the rankings in this manner is rather cumbersome, especially if the opinions of more than two people are under consideration. One natural simplification would simply be to use the first letter of the names of each fruit. This gives Alice the ranking (b, o, p, a) and Bob the ranking (o, p, b, a). Difficulty arises when there are multiple fruits starting with the same letter, e.g. if Alice and Bob were asked to rank pears, peaches, pineapples, and plums.

An alternative simplification that avoids this problem is to assign each fruit a number. This turns a ranking of the four fruits into a permutation of the numbers $\{1, 2, 3, 4\}$. If the fruits were assigned numbers alphabetically, apples would be 1, bananas would be 2, oranges would be 3, and pears would be 4. Under these labels, Alice's preferences would be written as (2, 3, 4, 1) or simply (2 3 4 1). Similarly, Bob's preferences would be written as (3 4 2 1). It is worth noting that there is no reason to assign the numbers alphabetically after all, pears are not twice as fruity as oranges.

One particularly useful method of labeling would be to assign numbers to the fruits based on Alice's preferences. Bananas, the fruit most preferred by Alice, would be represented by a 1. Oranges, Alice's second most-preferred fruit, would be represented by a 2, and so on. Using this set of labels, Alice's preferences would be the permutation (1 2 3 4), which is known as the identity permutation because it assigns item 1 the first rank, item 2 the second rank, etc. Bob's preferences would result in the permutation (2 3 1 4).

There are countless ways to assign labels, and an entire paper can be written about ways to assign fruits a numerical label. However, it is worth noting that given any collection of preferences on a set of items, there is always a set of numerical labels that will allow at least one of the preferences to be identified with the identity permutation. In fact, if there are m unique sets of preferences in the collection, then there are m different labeling schemes that will result in at least one of the preferences to be identified with the identity permutation.

As stated earlier, a collection of rankings of n items is equivalent to a collection of permutations

of the numbers $\{1, 2, \dots, n\}$. For simplicity, the rest of this paper will deal primarily with permutations. Permutations will always be on the set $\{1, 2, \dots, n\}$, equivalently, this means any set of rankings being examined will be ranking a set of n items. The number m will be used to denote the size of a collection of rankings or the number of permutations in a collection. The set of all permutations of the numbers $\{1, 2, \dots, n\}$ will be denoted S_n and has $n!$ elements. Members of S_n will usually be denoted with either r or s .

Those familiar with group theory in mathematics will recognize S_n as the symmetric group on $\{1, 2, \dots, n\}$ and will also recall that a permutation on $\{1, 2, \dots, n\}$ can be thought of as a bijective function mapping the set $\{1, 2, \dots, n\}$ to itself. Let $r \in S_n$. Then $r(i) = j$ means that r takes the i th element to j . In terms of rankings, this means the ranking associated with r ranks the item i as the j th highest.

Since r is a bijective function from $\{1, 2, \dots, n\}$ to itself, it has an inverse permutation, which will be denoted r^{-1} . Essentially, if r mixes up the numbers $1 \dots n$, then r^{-1} will “unmix” them. The most intuitive way to calculate r^{-1} is to use the fact that $r^{-1}(r(j)) = j$ for any $j \in \{1, 2, \dots, n\}$. So if $r(j) = i$, then $r^{-1}(i) = r^{-1}(r(j)) = j$. As before, $r^{-1}(i) = j$ means that the element that r takes to i is the j th element. In terms of rankings, this means the ranking associated with r assigns j the i th rank.

Since permutations are functions, two or more permutations can be composed together to get another permutation. The result of the composition operation, denoted by \circ , can be calculated directly. If $s(i) = j$ and $r(j) = k$, then:

$$r \circ s(i) = r(s(i)) = r(j) = k$$

For a concrete example, consider two permutations from S_5 , say $(3\ 4\ 1\ 2\ 5)$ and $(1\ 4\ 5\ 2\ 3)$. The composition of the two will be:

$$(3\ 4\ 1\ 2\ 5) \circ (1\ 4\ 5\ 2\ 3) = (5\ 2\ 1\ 4\ 3)$$

The identity permutation on the set $\{1, 2, \dots, n\}$ is always $(1\ 2\ 3\ 4 \dots n)$, which is the permutation that takes every element to itself. It will always be denoted e_n or simply e if the value of n is irrelevant or clear from context. The identity permutation has the important property that when composed with another permutation, the result is the original permutation. More explicitly, for any r in S_n ,

$$r \circ e = r = e \circ r$$

Additionally, if r and s are inverses, then $r \circ s = s \circ r = e$.

Another class of permutations worth mentioning separately are transpositions. Transpositions are identical to the identity permutation at all but two positions, call these i and j , which are in reverse order. For example, on S_5 , $(1\ 2\ 5\ 4\ 3)$ is a transposition with the elements 3 and 5 exchanged. Throughout this paper, transpositions will be denoted with the symbol τ_{ij} with i and j indicating which two elements are transposed. So our example transposition, $(1\ 2\ 5\ 4\ 3)$ would

be denoted τ_{35} . In general, the total number of elements in the permutation, n , should be clear.

Transpositions can be thought of as the most basic or “fundamental” permutations. That is, any permutation on S_n can be decomposed into the composition of the identity with a sequence of transpositions. Explicitly, this means that for any permutation $r \in S_n$, there exists a sequence of transpositions such that:

$$r = \tau_{ij} \circ \tau_{jk} \circ \tau_{kl} \circ \tau_{lm} \circ \dots \circ e_n$$

Additionally, calculating the result of composing a transposition with another permutation is straightforward. Take $r \in S_n$. The composition of r with τ_{ij} , call it s , will be:

$$s(k) = \begin{cases} r(j) & \text{if } k = i \\ r(i) & \text{if } k = j \\ r(k) & \text{otherwise} \end{cases}$$

Essentially, if r is written out as $(r^{-1}(1) \ r^{-1}(2) \ \dots \ r^{-1}(n))$, then $r \circ \tau_{ij}$ just exchanges $r^{-1}(i)$ and $r^{-1}(j)$ with each other.

A majority of this paper will be dealing with random permutations and probability distributions on permutations, so some notation for random permutations will be introduced here. Let X be a random permutation coming from a distribution, P , on the set of permutations of n items, S_n . The “chance that X takes on the permutation r ” or the “probability of permutation r occurring” is denoted:

$$P[X = r] = \pi, \quad \text{where } \pi \in (0, 1) \text{ and } r \in S_n$$

Typically, the distribution P will be a member of a family of distributions that can be parameterized by a vector, θ . In this case, the chance that X takes on the permutation r will be denoted:

$$P[X = r; \theta] = \pi, \quad \text{where } \pi \in (0, 1) \text{ and } r \in S_n$$

Every attempt will be made to indicate what family P comes from and, when appropriate, the values of θ .

One useful permutation for probability distributions on permutations is the mode. The mode is the permutation(s), typically denoted in this paper as r_0 , that has the greatest chance of occurring. Formally, if X is a random permutation from a distribution with mode r_0 ,

$$P[X = r_0] \geq P[X = r] \quad \text{for any } r \in S_n$$

Lastly, it is worth pointing out that the methods developed in this paper are to be used primarily on a collection of complete rankings. Although it is not unusual to encounter incomplete financial data, most of the studies on incomplete rankings focus on partial rankings, which is when a judge is asked to list his k most liked items out of a total of $n > k$ items. A typical example of such a ranking is when a movie critic is asked to pick his ten favorite films out of the hundreds released that year for a “Year’s Best” list. These kinds of rankings are not typically encountered in financial data. In financial data, incomplete rankings usually result from missing data points, which behave very different from partial rankings.

2.1 Distances

In order to evaluate how similar or close two permutations are to one another, a formalized sense of "closeness" or distance is needed. There have been many metrics developed for permutations, some more popular than others. This section of the paper attempts to provide enough understanding of these distances in order to use them later on. For a more thorough treatment of these distances as well as their statistical applications, a good reference would be [1] or [2].

Every distance discussed in this section is, indeed, a metric. That is, it satisfies the four properties of metrics:

1. (Non-Negativity) For any permutations r and s , $D(r, s) \geq 0$
2. For any permutations r and s , $D(r, s) = 0$ if and only if $r = s$
3. (Symmetry) For any permutations r and s , $D(r, s) = D(s, r)$
4. (Triangle Inequality) For any permutations r , s , and t , $D(r, t) \leq D(r, s) + D(s, t)$

When studying a collection of permutations or rankings, the labels assigned to each item should not have an impact on any calculations. After all, labels are arbitrarily assigned to the items to begin with, so it is desirable for distances on rankings to be unaffected by relabeling. More precisely, the distance between two permutations should be independent of the numbers assigned to each item being ranked. In the notation of permutations developed earlier, this means:

$$D(v \circ r, v \circ s) = D(r, s) = D(r \circ v, s \circ v) \quad \text{for all } v \in S_n$$

Distances that have this property are said to be bi-invariant and each distance discussed in this section has this property.

One very useful consequence of a distance being bi-invariant is that in order to calculate the distance between two permutations r and s , it is enough to know how to calculate the distance between a permutation and the identity permutation:

$$D(r, s) = D(s^{-1}r, s^{-1}s) = D(s^{-1}r, e)$$

This result will be used later when describing efficient methods for calculating the distance between two permutations.

2.1.1 Spearman's Distances

Of all the distances covered in this section, the family containing Spearman's ρ and Spearman's Footrule most closely resembles distances in the traditional or spatial sense. A distance in this family is defined by:

$$D_p(r, s) = \begin{cases} \sum_{i=1}^n |r(i) - s(i)|^p & \text{if } p \in (0, 1] \\ (\sum_{i=1}^n |r(i) - s(i)|^p)^{\frac{1}{p}} & \text{if } p \geq 1 \end{cases}$$

When $p = 2$, this is the well-known Euclidean distance between the two permutations and is known in nonparametric statistics as Spearman's ρ or the Spearman rank correlation coefficient. When $p = 1$, this is known as Spearman's Footrule.

In practice, these two distances are almost always normalized to lie on the interval $(-1,1)$. The formulas used are:

$$D_{Spearman}(r,s) = 1 - 6 \frac{\sum_{i=1}^n (r(i) - s(i))^2}{n(n-1)}$$

and

$$D_{Footrule}(r,s) = 1 - 3 \frac{\sum_{i=1}^n |r(i) - s(i)|}{n-1}$$

The result is a value that behaves similarly to the correlation coefficient: a value of 0 indicates no relation between r and s ; a value of 1 means r and s are identical, and a value of -1 means that r and s are complete opposites (e.g. $r = (1 \ 2 \ \dots \ n)$ and $s = (n \ \dots \ 2 \ 1)$).

The rest of the distances covered in this section are based on the size of a certain set. The notation $|S|$ will be used to denote the size or number of elements in the set S .

2.1.2 Kendall Distance

One of the most widely used distances in the field of nonparametric statistics is Kendall's distance. The Kendall distance between two permutations is usually defined as the number of discordant pairs. More formally,

$$D_{Kendall}(r,s) = |\{(i,j) \in \{1,\dots,n\} \times \{1,\dots,n\} : r(i) < r(j) \text{ and } s(i) > s(j)\}|$$

In group theory terms, the Kendall distance is equivalent to the minimum number of pairwise adjacent transpositions of items needed to transform one permutation into another. This definition of Kendall's distance in terms of transpositions provides a convenient way of calculating the Kendall distance between two permutations. $D_{Kendall}(r,s)$ is equal to the number of swaps made when sorting $s^{-1} \circ r$ with a bubble sort algorithm. Unfortunately, the bubble sort is one of the slowest sorting algorithms, so calculating the Kendall distance in this fashion is not the best approach. For a more efficient way of calculating this distance, see [3].

In practice, the Kendall distance is often normalized for easier interpretation. One way to do so is to divide by $\binom{k}{2}$ so that the resulting value lies on $(0, 1)$. A value of 0 indicates perfect agreement (i.e. they r and s are the same permutation) and a value of 1 indicates complete discordance. Another way to normalize the Kendall distance is to use the quantity

$$\frac{1 - 2D_{Kendall}(r,s)}{\binom{k}{2}}$$

which gives a value on the interval $(-1, 1)$.

2.1.3 Hamming Distance

Hamming's distance is defined as the number of elements that rankings r and s disagree on. Mathematically,

$$D_{Hamming}(r,s) = |\{(i,j) \in \{1,\dots,n\} \times \{1,\dots,n\} : r(i) \neq r(j)\}|$$

The Hamming distance has been well-studied in information theory, and there are many efficient ways to calculate this distance when r and s are just strings of bits. When working with permutations, calculating this distance becomes a bit slower, but it should still be reasonably efficient.

2.1.4 Cayley Distance

The Cayley distance between two permutations is the minimum number of transpositions that needs to be made in order to change one permutation into another. This is different from Kendall's distance, which is the number of adjacent transpositions required to change one permutation into another. In mathematical terms, Cayley's distance is defined as:

$$D_{Cayley}(r,s) = \min \text{ over all collections of transpositions } |\{\tau_1, \tau_2, \dots : s = \tau_1 \circ \tau_2 \circ \dots \circ r\}|$$

This distance has the least easily interpretable formula because Cayley's distance originates from the Cayley graph. The above definition of the Cayley distance is equivalent to the number of edges (or 'distance') between two vertices on the Cayley graph for S_n generated by transpositions. Its foundations in graph theory provides many efficient algorithms for calculating the Cayley distance between two permutations any algorithm that finds the shortest path between two points on a set of edges and vertices should be sufficient.

For those looking for a less graph-theoretic yet reasonably efficient method of calculating the Cayley distance, the answer lies in group theory. The Cayley distance between two permutations is equivalent to n minus the number of cycles in $s^{-1}(r)$. The proof of the equivalence of the two definitions is omitted. A straightforward derivation of this fact can be found in [4].

3 Models

Since there are a total of $n!$ possible permutations, any probability distribution on rankings can be completely determined by a list of $n!$ probabilities. More explicitly, any probability distribution on rankings can be parameterized by a vector with $n!$ elements. When n is small, say 2 or 3, $n!$ probabilities is not that many, and estimating that many parameters is feasible. However, as n grows, $n!$ probabilities becomes too many parameters to estimate. Consequently, for the first half of the twentieth century, statisticians working with permutations or rankings assumed the permutations were coming from the null or uniform distribution. In the null model, each permutation occurs with probability $\frac{1}{n!}$.

In 1950, Maurice Kendall realized the limitations imposed by this assumption, and he challenged the statistical community to come up with a better probability model for rankings or permutations. In the fifty years that have passed since Kendall's challenge, many statisticians have proposed a

variety of models. This part of the paper attempts to provide a cursory overview of models along with intuitive interpretations of these models. The family models that will actually be used and sampled from are the distance-based models, the justification for which will be provided after its introduction.

Most probability models on permutations attempt to either reduce the number of parameters from $n!$ down to a much smaller number or to have estimators for parameters that are easily computed and have known or easily calculable distributions. However, due to advances in computing, the latter quality is not nearly as crucial in today's world as formerly.

There are many properties of probability models on ranking worth discussing, and this paper will not attempt to discuss all of them. Critchlow, Fligner, and Verducci [5] have already done a superb job in their overview of probability models on rankings. The one property that will be mentioned in this paper is strong unimodality, which will be used to guarantee optimality of a local estimator in a later section of the paper.

A probability model is said to be strongly unimodal if (1) it has a mode, r_0 , and (2) the probability of a permutation, r , is nonincreasing as the agreement between r and r_0 decreases. Here, agreement is defined in the following sense: Take a pair of numbers, i and j . Take any permutation that has i and j adjacent to each other (i.e. $r(i) = r(j) - 1$ or $r(i) = r(j) + 1$).

1. If $r_0(i) < r_0(j)$ and $r(i) < r(j)$, then $r \circ \tau_{ij}$ is in less agreement with r_0
2. If $r_0(i) < r_0(j)$ and $r(i) > r(j)$, then $r \circ \tau_{ij}$ is in more agreement with r_0
3. If $r_0(i) > r_0(j)$ and $r(i) > r(j)$, then $r \circ \tau_{ij}$ is in more agreement with r_0
4. If $r_0(i) > r_0(j)$ and $r(i) < r(j)$, then $r \circ \tau_{ij}$ is in less agreement with r_0

Essentially, agreement, as defined above, is based on the relative orderings of a pair of elements. For any two elements i and j , permutations that have the same relative ordering are in more agreement with r_0 than permutations that have an opposing relative ordering. Formalized in the mathematical sense, a probability model is strongly unimodal if

1. There exists a r_0 such that $P(r) \leq P(r_0)$ for all $r \in S_n$
2. For any pair (i, j) in $\{1 \dots n\} \times \{1 \dots n\}$ such that $r_0(i) = r_0(j) - 1, P(r_0) > P(r_0 \circ \tau_{ij})$
3. For any pair (i, j) in $\{1 \dots n\} \times \{1 \dots n\}$ such that $r_0(i) < r_0(j)$ and any permutation r such that $r(i) = r(j) - 1, P(r) \geq P(r \circ \tau_{ij})$

3.1 Babington Smith Models

One of the first probability models for permutations to be developed was the family of Babington Smith models, which originates from the idea of comparing pairs of items. Given a collection of n items, suppose instead of placing all the n items in order of preference, a judge takes all of the $\binom{n}{2}$ pairs of items and lists which one he prefers. As an example, consider Alice trying to rank

the fruits {apples, bananas, pears, oranges}. Suppose Alice prefers bananas to apples, bananas to oranges, bananas to pears, oranges to pears, oranges to apples, and pears to apples. In this case, because there are no “circular rankings” (e.g. bananas to apples, apples to oranges, and oranges to bananas), this collection of paired comparisons constitutes a complete ranking of the fruits: (bananas, oranges, pears, apples). Similarly, a complete ranking can be decomposed into a collection of paired comparisons. If Bob’s preferences for fruit are (oranges, pears, bananas, apples), then Bob would prefer oranges to pears, oranges to bananas, oranges to apples, pears to bananas, pears to apples, and bananas to apples.

Random permutations can also be constructed in this manner. Suppose a judge independently compares each pair of items, with probability p_{ij} of preferring item i to j . If the results of the comparisons have circular rankings, then the items are compared again. Eventually, a set of non-circular, or *consistent*, preferences will result, which can then be composed into a permutation. If X is a random permutation generated in that manner, the probability that X takes on permutation $r \in S_n$ is:

$$P[X = r] = c \prod_{r(i) > r(j)} p_{ij}$$

Where c is a normalizing constant, and is calculated as:

$$c = \frac{1}{\text{chance of a consistent ranking}} = \frac{1}{\sum_{x \text{ is consistent}} P[X = x]}$$

Models from this family of probability distributions on permutations are known as Babington Smith models. They are parameterized by just $\binom{n}{2}$ parameters since $p_{ij} = 1 - p_{ji}$. While $\binom{n}{2}$ does not grow as fast as $n!$, it still gets intractable quite quickly. However, there are many simplifications to the model that reduce the number parameters dramatically.

One popular simplification of this family of models is the Mallows-Bradley-Terry model. In these types of models, instead of using $\binom{n}{2}$ probabilities, it associates to each of the n items a number, call it v_i , $i = 1, 2, \dots, n$. p_{ij} is then calculated as:

$$p_{ij} = \frac{v_i}{v_i + v_j}$$

3.2 Thurstonian Models

Models on permutations in the Thurstonian family are based on the order statistics of random variables. This family of models extends on Thurstone’s studies [6] on how people determine how loud different sounds are - hence the name . However, it is Daniels [7] who applied Thurstone’s work to create this family of permutations .

The model is based on the idea that given a set of n items, how much each judge “is stimulated by” or “prefers” item i is determined by a random variable, call it Z_i . Presumably Z_1, Z_2, \dots, Z_n come from the same family of distributions, but with different location parameters. If all the Z_i ’s are independent and have the same distribution, then this is just the null or uniform distribution. The actual ordering of the Z_i ’s is the observed permutation. For example, if a judge is trying to

rank his or her preference for {cats, dogs, fish}, and $Z_{cats} = 1.5$, $Z_{dogs} = 0.5$, and $Z_{fish} = 2.5$, then the permutation from that judge will be (dogs cats fish).

The exact number of parameters in a specific model depends on the number of parameters for each of the Z_i 's as well as the number of assumptions one is willing to make. For example, Thurstone originally used the normal distribution for the Z_i 's. Without any other assumptions, this comes out to n means, n variances, and $\binom{n}{2}$ covariances, giving a total of $2n + \frac{n(n-1)}{2}$ parameters. While this is quite a few parameters to estimate, it is much smaller than $n!$ when n is greater than 7 or 8. If one is willing to assume that the Z_i 's are independent and have the same variance, then there are only n means and one variance to estimate, giving $n + 1$ parameters.

After some thought, one quickly realizes that this family of models is quite large. The Z_i 's can come from almost any family of distributions. Work has been done that uses both the Gumbel and Gamma distributions (see [8, 9, 10]). Additionally, the Z_i 's do not necessarily have to be independent as was originally assumed by Thurstone. For example, suppose one is trying to conduct a study on rankings of different movies. In this case, if a judge has a liking for the action film *Gladiator*, then it is likely that the judge will have an increased preference for other action films, such as *Braveheart* or *The Terminator*. Relaxing the assumption of independence among the Z_i 's, while complicating the models significantly, makes this family of models much more inclusive.

3.3 Mallows / Distance-Based Models

The family of distance-based models, sometimes known as metric-based models, is characterized by two parameters, a mode and a spread parameter. The mode, denoted r_0 , is the ranking which the distribution is "centered" around. The spread parameter, θ , can be thought of as how much to "penalize" a permutation for each unit of distance it is away from the mode. For a random permutation X coming from a distance-based model with mode r_0 and spread θ ,

$$P[X = r; \theta, r_0] = \frac{\exp(\theta d(r, r_0) - \psi(\theta))}{n!}$$

Where $d(\cdot, \cdot)$ can be any of the distances mentioned earlier or any metric for permutations, and $\psi(\theta)$ is a constant that normalizes the distribution to sum to one. Two famous models of this type are Mallows' [11] ϕ model, which uses the Kendall distance. Mallows' θ model is the distance-based model with Spearman's ρ . It is worth noting that while $\psi(\theta)$ does depend on θ as the notation would suggest, most formulas for $\psi(\theta)$ are also dependent on the distance being used.

If θ is negative, this corresponds to the previously mentioned idea of a distribution centered around r_0 that tapers as the permutations get further and further from r_0 . However, in some cases, there may be models where θ is positive. In these cases, the distribution r_0 is the least probable permutation, and as permutations get further from r_0 , their chances of occurring increases. When θ is zero, this corresponds to the uniform or null distribution.

The analysis in this paper will primarily use distance-based models. Of the models discussed in this paper, they have the fewest parameters, which is desirable given the limited amount of data available. Specifically, the two models used will be the distance-based model with Kendall's and Cayley's distances. These two distances are used because there is a closed form expression for the normalizing constant in their distribution for each of the distances. Being able to calculate this constant allows the estimation of parameters with a maximum likelihood method and also simplifies the simulation process.

When the Kendall distance is used, the normalizing constant is:

$$\psi_{Kendall}(\theta) = \sum_{i=1}^m \log \left(\frac{1 - \exp(\theta i)}{i(1 - \exp(\theta))} \right)$$

For Cayley's distance, the normalizing constant is:

$$\psi_{Cayley}(\theta) = \sum_{i=1}^m \log((i-1)\exp(\theta) + 1) - \log(m!)$$

There is also a closed form expression for the constant with the Hamming distance:

$$\psi_{Hamming}(\theta) = m\theta + \log \left(\sum_{k=0}^m \frac{1}{k!} (\theta - 1)^k \right)$$

Time limitations did not allow the exploration of fitting the Hamming distance-based model to the data. If one is interested in the derivations of the formulas for the constants, they can be found in [4].

4 Model Fitting

To fit a distance-based model to a set of data, the parameters can be estimated by taking a maximum likelihood approach or through Bayesian estimation. In this paper, the maximum likelihood method suggested in [4] will be used. The section in Marden's comprehensive book details the derivation of the estimation method well, but contains little useful information on actually performing the necessary steps to actually estimate parameters from a data set. Consequently, this section of the paper will focus more on the methodology rather than the derivations.

So let $M := \{x_1, x_2, \dots, x_m\}$ be a collection of m rankings on S_n . For a distance-based model on, the log likelihood is given by:

$$l(\theta, w; M) = \theta S(w) - m\psi(\theta) - \log(n!)$$

Where $S(w)$ is defined as:

$$S(w) = \sum_{i=1}^m d(w, x_i)$$

The θ and w that maximizes $l(\theta, w; M)$ will also maximize

$$l'(\theta, w; M) = \theta S(w) - m\psi(\theta)$$

In order to maximize l' , one first finds the permutation w that maximizes and minimizes (for the case of $\theta < 0$) $S(w)$. To solve for the permutation that maximizes and minimizes $S(w)$, Fligner and Verducci have suggested the following local search method [12]. The search starts from the “mean rank,” which is the rank of the average ranking in each position. For example, if the average ranking in each position for a set of data is (2.4 1.3 3.5 3.2), the “mean rank” would be (2 1 4 3). From the mean rank, a local search for the minimizing and maximizing permutation is performed among the the $n - 1$ permutations that are within one Kendall distance of the mean rank. This local search will converge satisfactorily to the global maximum and minimum permutations if the model is strongly unimodal [12]. The distance-based model under the Kendall distance is strongly unimodal, but with the Cayley distance it is not [5].

After solving for the two permutations, call them \hat{w}_{min} and \hat{w}_{max} , find the value of θ that maximizes the log-likelihood for each permutation. This can be done using any kind of nonlinear maximization algorithm such as a descent method. This produces a pair of estimates: $(\hat{w}_{min}, \hat{\theta}_{min})$ and $(\hat{w}_{max}, \hat{\theta}_{max})$. The pair of parameters with the greatest log-likelihood will be the maximum likelihood estimate.

4.1 Assessing Goodness of Fit

There are many ways to assess how well the model fits the actual data, and some of the more useful methods will be discussed here. While the previous section on model fitting applies only to distance-based models, almost all the ideas contained in this section can be applied to any kind of models.

The simplest and quickest way to assess how well a model fits the data would be to compare the expected value and the observed value of certain statistics. Possible choices for comparison include the expected distance from the mode or the expected rank in each position. With enough time, the sampling distribution for these statistics can be calculated through simulation and a formal assessment can be made.

Cohen and Mallows [13] suggest a clever method to assess how well a model fits to a set of data. Let $\{r_1, r_2, \dots, r_m\}$ be a collection of random permutations from S_n that are independent and identically-distributed with distribution F . For any fixed permutation, s , the distance between r_i and s , $d_i = d(r_i, s)$, is like drawing a number from $\{1, \dots, k\}$ with probabilities $\{p_1, p_2, \dots, p_k\}$, where k is the maximum value that $d(\cdot, \cdot)$ can take on S_n , and p_1, p_2, \dots, p_k are probabilities that depend on both F and d . Since each r_i is independent, the distances d_i are also independent, which means the random vector

$$\mathbf{d} = \left(\sum_{i=1}^m \mathbf{1}(d_i = 1), \sum_{i=1}^m \mathbf{1}(d_i = 2), \dots, \sum_{i=1}^m \mathbf{1}(d_i = k) \right)$$

is distributed as a multinomial with n trials and probabilities p_1, p_2, \dots, p_k . The most logical choice of s would probably be the estimated mode, \hat{r}_0 or the identity, e_n . A comparison of the

expected value of each entry of \mathbf{d} and the observed values in $\hat{\mathbf{d}}$ provides a simple way to assess how well a model fits the data. With appropriately chosen bins, a χ^2 test can even be performed on the data to formally test for goodness of fit.

If the model was fit using a maximum likelihood method, another way to assess the goodness of a model is to find a larger model that contains the one being fit, and compare likelihoods of the two models. If the model is a good fit, then the larger model will not improve the fit of the model to the data and so the two likelihoods should be comparable. If the model fits the data poorly, then using a larger model will have a much better fit and consequently a much greater likelihood.

When dealing with the distance-based models, the best model to compare likelihoods with is the generalized exponential model using the same distance. The distance-based model is a special case of the generalized exponential model, so comparing the two is the most natural. The generalized exponential model requires a lot of background in permutations to develop, and will not be introduced in this paper. Additionally, the parameters of a generalized exponential model are difficult to estimate. When just using Kendall's distance, [14] describes a very clever method of estimation based on the branch and bound algorithm. For non distance-based models, a larger model that can be used in a likelihood ratio comparison can be found in [5].

5 Testing

After a model is fit and parameters are estimated, hypothesis testing on the rankings can be performed. As mentioned earlier, the hypothesis tests will attempt to determine if a method of forecasting stock returns predicts rankings that are similar to the actual rankings. Luckily, testing for this property is equivalent to testing if a set of permutations come from a distribution that has the identity as the mode. The equivalence of these seemingly different testing problems will be discussed in detail later. For now, the focus will be on development of the actual testing procedure.

Let $\{x_1, x_2, \dots, x_m\}$ be a set of permutations on S_n coming from a distribution F . Then the hypotheses to be tested are:

H_0 : The mode of F is the identity permutation or $r_0 = (1 \ 2 \ \dots \ n)$

H_A : The mode of F is different from the identity permutation or $r_0 \neq (1 \ 2 \ \dots \ n)$

In the empirical study in this paper, F was assumed to be the distribution of a distance-based model. Under the null hypothesis, the mode is assumed to be the identity permutation, as is stated in the hypotheses above. The spread parameter is not specified, but can either be assumed, if one has the luxury of such information a priori, or can be estimated using the maximum likelihood described in the previous section, which is the method that will be used in this paper.

There are many possible test statistics to use, the most basic ones being the average distance from the mode. Any of the distances mentioned previously can be used, regardless of which

distance, if any is used in the model. There are more advanced possibilities, but the average of the distances mentioned earlier should be more than enough for most situations since there are distances based on permutations, actual number of different elements, special separation, etc. The test statistic that results in the test with the most power should probably be used in most cases. Unfortunately, no considerations on power were made due to time limitations.

In general, the null distribution of the test statistic is too difficult to calculate analytically. Marden [4] summarizes a few asymptotic results for certain distance-based models, such as approximations to the likelihood or approximate distributions of the distances. These results may be useful in certain cases, but rarely does one have a large enough collection of permutations to justify the use of asymptotic results.

Instead, the null distribution of the test statistic can be estimated using simulations. Most of the families of models mentioned in this paper either (1) were given interpretation that allows for a straightforward way to simulate data (such as the Babington Smith or Thurstonian family of models) or (2) have an explicit formula for the probability of a permutation (such as distance-based models) which means an acceptance-rejection method can be used to simulate permutations. It should be noted that although these methods are easy to implement, there are no guarantees made regarding their speed or efficiency.

With a simulated sampling distribution, the p -value can be estimated and the null hypothesis can be accepted or rejected. Acceptance the null hypothesis does not necessarily mean that the mode of the distribution F is the identity permutation. Rather, this result implies that there is no reason to believe otherwise. While the differences between the two are subtle, this distinction must be made.

6 Application to Stock Forecasts

This section will detail how to apply the methodology developed up until this point to stock forecasts. For any collection of n stocks, the month to month returns give rise to a collection of rankings, which will be referred to as the actual rankings. Let $\{s_1, s_2, \dots, s_m\}$ be the collection of permutations associated with each of the actual rankings. The labels for each company can be arbitrarily chosen, but for simplicity we will assume that each company is numbered alphabetically. Applying any forecasting methodology applied to the same n companies gives rise to another set of rankings, which will be called the forecasted rankings. Let $\{r_1, r_2, \dots, r_m\}$ be the collection of permutations associated with the forecasted rankings. Once again, it does not matter what labels are used as long as the labels used for r_j are the same as the labels used for s_j for $j = 1, 2, \dots, m$.

Suppose that the forecasting method used has perfect foresight and is able to predict the returns of each company perfectly. In this case, $r_i = s_i$ for $i = 1, 2, \dots, m$. Additionally, by applying s_i^{-1} to each side, this gives $s_i^{-1} \circ r_i = e_n$ for $i = 1, 2, \dots, m$. This is true regardless of the labeling system used to assign numbers to each company. In fact, this is true even if the labels for month i are different from the labels used for month j . If the predicted returns for month j are perfect, then

$$s_j^{-1} \circ r_j = e_n.$$

This restriction of the forecasts having perfect foresight can be loosened because ranks are generally not sensitive to small fluctuations. As long as the forecasted returns are similar enough to preserve the ordering of the stocks, then $s_j^{-1} \circ r_j$ will be equal to the identity. Additionally, if r_i and s_i are close to each other, say they are off by a transposition - $r_i = s_i \circ \tau_{jk}$ for some $j, k \in 1, 2, \dots, m$, then applying s_i^{-1} to both sides gets:

$$s_i^{-1} \circ r_i = s_i^{-1} \circ s_i \circ \tau_{jk} = \tau_{jk}$$

which is close to the identity permutation, e_n . This means that testing if r_i is close to s_i for $i = 1, 2, \dots, m$ is equivalent to testing if $s_i^{-1} \circ r_i$ is close to e_n for $i = 1, 2, \dots, m$.

It is worth pointing out that $s_i^{-1} \circ r_i$ is actually just a relabeled version of the forecasted rankings. As mentioned earlier, for any two rankings, there is always a set of labels that will force one of the permutations associated with the rankings to be the identity permutation. If the labels for each stock are chosen according to the actual rankings (e.g. the company with the highest ranking is assigned 1, the company with the second highest ranking is assigned 2, and so on), then s_i will be the identity permutation, which means s_i^{-1} is the identity permutation as well, and so r_i under this choice of labeling will remain the same under s_i^{-1} , that is $s_i^{-1} \circ r_i = r_i$.

To determine whether or not a forecasting method is effectively forecasting the returns of a collection of n stocks over m periods of time, first calculate the set of actual rankings of the n stocks, $\{s_1, s_2, \dots, s_m\}$, using the actual returns of the stocks in the m periods being studied. It should be noted that the m periods need not be contiguous. For example, if there was a major catastrophe that caused abnormalities in the stock market, that period of time can be excluded.

Next, compute the forecasted returns of the n stocks over the m periods and calculate the forecasted rankings, $\{r_1, r_2, \dots, r_m\}$. With the two sets of permutations, $\{s_1^{-1} \circ r_1, s_2^{-1} \circ r_2, \dots, s_m^{-1} \circ r_m\}$ can be computed, and the previously described model fitting and hypothesis testing can be applied to this collection of permutations.

If the null hypothesis is accepted at significance level α , then for reasonably chosen α , it is reasonable to believe that the forecasting method that is used to produce the rankings r_1, r_2, \dots, r_m is effective at predicting stock returns. If the null hypothesis is rejected at significance level α , then it is likely that the method producing r_1, r_2, \dots, r_m is not accurately predicting stock returns. Unfortunately, without any considerations on the power of this test, there is no way to tell how certain to be about this statement.

As with any hypothesis test, there are certain practices one should be wary of. First, it is not possible to compare two different forecast methods based on p -values. That is, if testing two different methods, say Alice's method and Bob's method, Alice receiving a higher p -value than Bob does not mean that she is able to forecast better than Bob's method.

Additionally, since there are so many possible test statistics to use, one might feel compelled to compute all of them and perform a hypothesis test using each of the statistics. However, as with

any kind of hypothesis testing, these multiple comparisons will cause inference errors, particularly the incorrect rejection of null hypotheses. Consequently, when using multiple test statistics to assess one set of forecasts, if the null hypothesis is ever rejected, one must be aware of the possibility of multiple testing causing those types of results.

7 Empirical Study

A large financial services company has kindly provided stock forecasts for the companies in the Russell 1000 Index for an empirical study. The predictions are monthly forecasts over a period of ten years, giving 120 total data points. There are not nearly enough data points to fit a model on all 1000 companies. Instead, the twenty companies whose predicted returns (not ranks) was most highly correlated with actual returns was chosen. The distance-based model with both the Kendall distance and the Cayley distance were fit to the data using the previously described maximum likelihood method. Under both distances, the permutation w that minimized $S(w)$ had the higher likelihood and was used.

For the distance-based model with the Kendall distance, the estimated mode was:

$$\hat{r}_{0,Kendall} = (3\ 2\ 1\ 5\ 7\ 11\ 4\ 15\ 8\ 13\ 6\ 10\ 12\ 14\ 9\ 16\ 18\ 17\ 19\ 20)$$

For the distance-based model with the Cayley distance, the estimated mode was:

$$\hat{r}_{0,Cayley} = (1\ 3\ 2\ 4\ 6\ 9\ 5\ 16\ 8\ 12\ 7\ 10\ 13\ 14\ 11\ 15\ 17\ 19\ 18\ 20)$$

The estimates for θ and the maximum log likelihood achieved is given in Table 1. Worth noting is that the log likelihood for the model under the Kendall distance is much higher.

Distance	$\hat{\theta}$	Loglikelihood
Kendall	-0.0575	46.130
Cayley	-0.3535	17.223

Table 1: Estimated parameters and loglikelihoods.

Position	1	2	3	4	5	6	7	8	9	10
Observed	8.76	8.91	8.49	9.21	9.58	10.53	9.28	11.53	10.11	10.78
Kendall	9.02	8.83	8.63	9.40	9.81	10.61	9.21	11.38	10.00	10.99
Cayley	10.29	10.36	10.33	10.37	10.41	10.47	10.37	10.62	10.44	10.52

Position	11	12	13	14	15	16	17	18	19	20
Observed	9.95	10.61	11.03	11.43	10.61	11.58	11.60	11.96	12.01	12.05
Kendall	9.60	10.40	10.79	11.20	10.21	11.59	11.98	11.78	12.18	12.36
Cayley	10.42	10.49	10.55	10.57	10.51	10.59	10.65	10.68	10.65	10.71

Table 2: Comparison of observed average ranks and expected average ranks.

To assess the fit of the models, the expected and observed average rank in each position was compared - see Table 2. The expected average ranks were estimated numerically by simulating permutations from both models.

Based on the average ranks, the distance-based model with the Kendall distance fits the observed data rather closely. Most of the observed average ranks are within .2 to .3 of the expected average rank. They even exhibit the same behavior - the average rank for position 11 is lower than the average rank for position 10 for both the observed and expected cases. The distance-based model with the Cayley distance does not fare so well, with the average rank being around 10.5 for all the positions. This is not surprising considering the likelihood for the model with the Cayley distance was much lower than that of the Kendall distance.

Distance	Observed	Expected
≤ 50	1	2.01
51-60	9	7.68
61-70	26	18.78
71-80	27	29.00
81-90	21	29.47
91-100	12	20.10
101-110	20	9.38
111-120	3	2.92
≥ 120	1	0.68

Table 3: Observed and expected counts of distances for the distance-based model using Kendall's distance

Distance	Observed	Expected
≤ 11	1	0.91
12	1	2.71
13	9	7.50
14	12	16.31
15	35	26.25
16	29	30.33
17	18	23.34
18	14	10.56
≥ 19	1	2.08

Table 4: Observed and expected counts of distances for the distance-based model using Cayley's distance.

Additionally, the distance from the mode was used to partition the permutations, as suggested by Cohen and Mallows, and the observed and expected counts in the bins were compared. The results of this comparison are reproduced in Tables 3 and 4. The observed counts for each bin in the data follow the expected counts for the distance-based model using Cayley's distance rather closely. The distance-based model using Kendall's distance does not fit as well as with Cayley's distance, but the observed counts for the Kendall distances still do not deviate much from the

expected counts. Based on these two assessments, both models fit the observed permutations data reasonably well, but the model based on Kendall's distance fits better due to the higher loglikelihood and the greater resolution of Kendall's distance from the wider range.

A hypothesis test as described in section 6 was conducted with the the distance-based model with Kendall's distance. The hypotheses being tested are:

H_0 : The forecasted rankings come from a distribution that is centered about the identity permutation.

H_A : The forecasted rankings come from a distribution that is not centered at the identity permutation

To estimate a p -value, almost 2,000,000 permutations were simulated from the selected type of model with the identity permutation as the mode and with the maximum likelihood estimate of $-.0575$ for the spread parameter. These 2,000,000 permutations were randomly organized into 16,000 groups of 120 permutations. The average Kendall distance from the identity for each group was computed to give 16,000 simulated data points, with the simulated sampling distribution drawn in Figure 1.

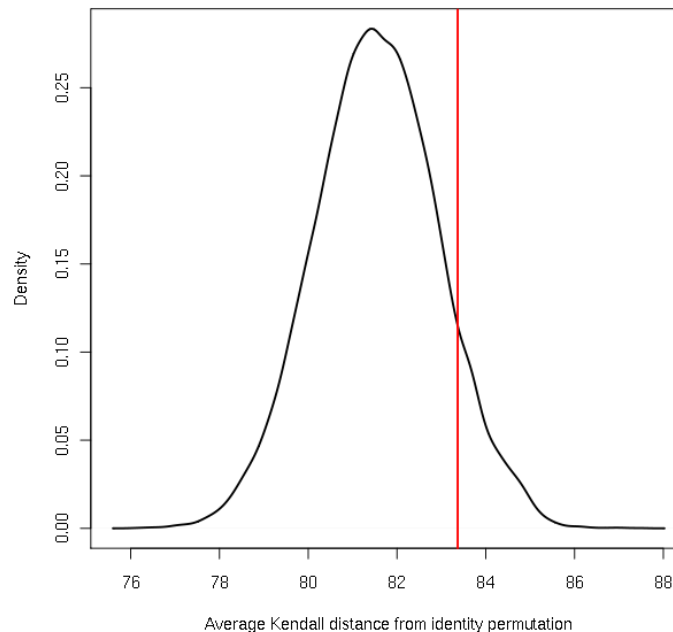


Fig. 1: The sampling distribution from 16,000 samples of the average Kendall distance from the mode under the null hypothesis. The observed average distance from the mean is denoted in red.

The observed average Kendall distance from the identity permutation for the 120 actual forecasted rankings was 83.3667 and is indicated in red. This gives an estimated p -value of .0954, which is not significant at the 5% level and so for the twenty companies being studied, the null hypothesis is reasonable - the forecasted rankings come from a distribution that is centered about the identity permutation or that the forecasts are making good predictions.

8 Future Work

At this point there are three issues that have yet to be fully considered. First and foremost, this method should be applied to larger data sets. A study of just twenty companies, as done in this paper, is not particularly useful when there are tens of thousands of companies to choose from on a typical stock exchange. The feasibility of this method when the number of companies being considered increases is unknown and needs to be determined.

Second, some analysis on the power of the hypothesis test should be done. After all, a hypothesis test is not particularly useful without any sense of the power. The main difficulty in power analysis is trying to come up with a specific alternative hypothesis. One might try to avoid this by analyzing a power function, but that may prove difficult to interpret as well. For a collection of n stocks, the power function is a function from S_n to $(0, 1)$.

Last, a method of generalizing this process in order to be able to compare two or more sets of forecasts would be useful. The biggest drawback of this method is not being able to have a numerical value to compare multiple methods with. Being able to decide between two methods of forecasts would make this type of rank analysis much more applicable in non-academic settings.

9 References

- [1] Persi Diaconis. *Group Representations in Probability and Statistics*. Institute of Mathematical Sciences, Hayward, CA, 1988.
- [2] Douglas E. Critchlow. *Metric Methods for Analyzing Partially Ranked Data*. Springer-Verlag, New York, NY, 1985.
- [3] William R. Knight. A computer method for calculating kendall's tau with ungrouped data. *Journal of the American Statistical Association*, 61:436–439, 1966.
- [4] John I. Marden. *Analyzing and Modeling Rank Data*. Chapman and Hall, London, UK, 1995.
- [5] Michael A. Fligner Douglas E. Critchlow and Joseph S. Verducci. Probability models on rankings. *Journal of Mathematical Psychology*, 35:294–318, 1991.
- [6] L. L. Thurstone. The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology*, 31:384–400, 1927.
- [7] H. E. Daniels. Rank correlation and population models. *Journal of the Royal Statistical Society Series B*, 12:171–191, 1950.
- [8] R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, NY, 1959.
- [9] Robert J. Henery. Permutation probabilities for gamma random variables. *Journal of Applied Probability*, 20:822–834, 1983.

- [10] Hal Stern. Models for distributions on permutations. *Journal of the American Statistical Association*, 85:558–564, 1990.
- [11] C. L. Mallows. Non-null ranking models i. *Biometrika*, 44:114–130, 1957.
- [12] Michael A. Fligner and Joseph S. Verducci. Multistage ranking models. *Journal of American Statistical Association*, 83:892–901, 1988.
- [13] Ayala Cohen and C. L. Mallows. Assessing goodness of fit of ranking models to data. *The Statistician*, 32:361–374, 1983.
- [14] Arthur Patterson Jeff Bilmes, Marina Meila and Kapil Phadnis. Consensus ranking under the exponential model. *UW Statistics Department Technical Report 515*, 2007.
- [15] Michael A. Fligner and Joseph S. Verducci. Distance based ranking models. *Journal of the Royal Statistical Society Series B*, 48:359–369, 1986.
- [16] M. G. Kendall. Discussion on symposium on ranking methods. *Journal of of the Royal Statistical Society Series B*, 12:153–162, 1950.
- [17] B. Babington Smith. Discussion on professor ross’s paper. *Journal of of the Royal Statistical Society Series B*, 12:54, 1950.