Some Topics Related to Diversity Indices and Applications

SOME TOPICS RELATED TO DIVERSITY INDICES AND APPLICATIONS

BY

QIWEI JIANG, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

© Copyright by Qiwei Jiang, December 2018

All Rights Reserved

Master of Applied Science (2018) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE:	Some Topics Related to Diversity Indices and Applications
AUTHOR:	Qiwei Jiang B.Sc., (Statistics) University of Toronto, Toronto, Canada
SUPERVISOR:	Dr. Shui Feng
NUMBER OF PAGES:	viii, 75

This thesis is dedicated to my family.

Abstract

This thesis discusses three diversity measures: Simpson's Index, Shannon's Index and Berger–Parker Dominance as well as their corresponding True diversity ($\mathbf{N}_q$). Evenness measures the balance of a community which carries different information from diversity indices. We give an example on the application of diversity indices by comparing the surname diversity of China and USA. We also use OLS regression to investigate whether diversified investment strategy leads to higher return rate for mutual funds. Our analysis shows that funds that are diversified in investment have higher return rate. Although highly diversified investment does not translate into high return directly, the market where equity enters also plays an important role. Our analysis reveals the potential of investment diversity and provides motivation for diversifying investment strategy. The diversity indices also have the ability of discriminating categories. With linear discriminant analysis (LDA) and classification tree method, we use Simpson's index and year to date return rate to successfully differentiate mutual fund category. In the last part of the thesis, we introduce the Bayesian approach to estimate diversity indices from an observed sample. We propose four estimators of Simpson's Index based on the sampling distribution of relative abundance and investigate their estimating ability.

Acknowledgements

I would first like to thank my supervisor Prof. Feng, Shui of the Department of Mathematics & Statistics at McMaster University. Prof. Feng continually provided me guidance and insightful conversations with his expertise during the development of the ideas in this thesis.

I would also like to express my gratitude to my committee members: Prof. Fred M. Hoppe and Prof. Roman Viveros-Aguilera for their precious time and knowledge. Without their participation, the defence of my thesis could not be successfully conducted.

I thank Dr. Kai Liu for providing me suggestions on the simulation study.

Finally, I am grateful to have unfailing support from my family and friends. Their encouragement means a lot to me.

Contents

List of Figures

Chapter 1

Introduction

For a given community that contains different types, if we want to investigate its diversity, the total number of individuals in the community does not provide adequate information. A community with q large population but only a few types is not considered as abundant as another community with same population size but more types. On the other hand, the number of types (Richness S) provides better interpretation, but still poorly reflects the diversity of the community. Two communities with same richness could be different in diversity as one community have few types that dominate whole population while the other community is equally distributed among all types. Diversity indices possess such ability by quantifying how types are distributed in community. The true diversity (N_q) transforms the indices into the number of equally abundant types needed to match the abundance level of that community where all types may not be equally distributed. Diversity indices has been used in various fields. In ecology, types may refer to species. In economics, we can analyse industry diversity based on industry sector of that area. The indices can also be applied in finance such as evaluating investment diversity based on investment

portfolio.

In Chapter 2, we are going to introduce several diversity measures as well as their properties. Most diversity indices are derived from N_q where q is the order of diversity. S is a single diversity measure that corresponds to N_0. Using S as diversity measure ignores totally the frequency of each type and only reflects the number of types in a community. Berger–Parker Dominance (BP) is another single measure that takes the value of most dominate frequency that corresponds to N_∞. Simpson's index (D) and Shannon's index (H) are two compound diversity measures that have been commonly used in many literatures. Evenness (E) is defined as "the degree to which the abundances are equal among the species present in a sample or community" (Molinari 1989). There have been a lot of debates over which index provides the best interpretation of a community. So far, none has been identified as the most appropriate. The discussion of choosing diversity indices is included in the last section of Chapter 2. Morris et al. (2014) discuss that different diversity indices provide fairly good abilities for different purposes. Character of a community that is driven mainly by rare species is well captured by H while D performs the best in estimating total number of species. Therefore, it is suggested that at least two measures should be used in analysis (Whittaker 1972; Stirling and Wilsey 2001; Heino et al. 2008).

The last part of Chapter 2 provides a comparison of surname diversity between China and USA in order to illustrate an application of diversity indices. We use D and H to access the name diversity of two countries. Our analysis shows that USA is more diversified in surnames than China. Both D and H provide consistent results. We also report E in our analysis. The result shows that both China and USA have

unbalanced name structure, dominant surnames can be found in the two countries. USA has more unbalanced surname name structure as indicated by rank abundance curves and lower value of E.

Diversity indices do not only reflect the structure of a community, outcomes that are driven by diversity can be revealed by regression model. Goetz et al. (2014) used panel data analysis (PDA) to find out the effect of geographic diversity on risk of bank holding companies (BHCs). BHCs that are more diversified in geography exposed to less idiosyncratic local risk, expansion into economically dissimilar areas helps reducing risk more. In Chapter 3, we are interested in whether higher return is associated with diversified investment strategy for mutual funds. Our results show the trend that more diversified investment leads to higher return rate is significant. The market where the equity enters also matters. D is preferred when estimating one year annual return and H is preferred for estimating year to date return.

Morris et al. (2014) use diversity indices to differentiate site with principle component analysis (PCA), D provides the greatest such ability. In Chapter 4, we are also interested in whether such discrimination ability of diversity indices is preserved in mutual fund data so we can use diversity indices and return rate to differentiate category of mutual fund. Our analysis shows that unsupervised clustering methods fail the task, k-means, k-mediods and hierarchical clustering produce high miss-classification rate. On the other hand, with supervised clustering method, linear discriminant analysis (LDA) and classification tree successfully discriminate data. Diversity indices do have profound ability of discrimination, however, with appropriate methods.

Chapter 5 discusses the Bayesian approach to estimate diversity indices. Suppose we are going to estimate the diversity index from a given sample, using the maximum

likelihood estimator of relative abundance might not be appropriate, especially when the sample size is small. Therefore, we assume the relative abundance $(p_1, p_2, \ldots, p_S)$ follows a Dirichlet distribution. After the sample is observed, the posterior marginal density of p_i is Beta distribution. Gill and Joanes (1979) propose two estimators of Shannon's Index h_α and $\hbar_\alpha$ based on the posterior distribution of p_i. Following the same logic, we propose two estimators of Simpson's index, d_α and $\widetilde{d}_\alpha$. The bias corrected versions of these two estimators d_α^{BC} and $\widetilde{d}_\alpha^{BC}$ are also reported. A simulation study and a real case study are carried out in order to assess the estimating ability. Our proposed estimators are competitive to Simpson's estimator λ as they produce similar estimate mean as λ but smaller variance.

Chapter 2

Introduction to Diversity Indices

2.1 True Diversity ($\mathbf{N}_q$)

Diversity indices reflect how abundant a community is. The true diversity (N_q) known as effective number of types quantifies the diversity of a system by presenting the number of equally abundant types/species needed so that their average proportion matches that of the system. We can compare diversity of different systems by looking at their true diversity.

Suppose we observe the dataset with richness S (S different species). Let N be the population in the community, N_i is the number of individuals for the i^{th} type. Therefore, we can calculate the proportion of type i as $p_i = \frac{N_i}{N}$. The equation of N_q is as following:

$$N_q = \frac{1}{M_{q-1}} = \frac{1}{\sqrt[q-1]{\sum_{i=1}^{S} p_i p_i^{q-1}}} = \left(\sum_{i=1}^{S} p_i^q\right)^{\frac{1}{1-q}} \tag{2.1}$$

M_{q-1} is the weighted generalized mean with exponent $q-1$. Note that when $q = 1$, (2.1) is not defined, however, we can obtain the equation of order 1 by taking

the limit of N_q as q approaches 1 , see Hill (1973)[11]:

$$N_1 = \exp\left(-\sum_{i=1}^{S} p_i \ln\left(p_i\right)\right) \tag{2.2}$$

Theorem 2.1.1. $\lim_{q\to 1}\left(\sum_{i=1}^{S} p_i^q\right)^{\frac{1}{1-q}} = \exp\left(-\sum_{i=1}^{S} p_i \ln p_i\right)$

Proof. Two mathematical results are used in this proof:

1. For small value of x, $\exp(x) \approx 1 + x$
2. For small value of x, $\ln(1 + x) \approx x$

Now the proof could be continued with the above results. Let $q = 1 + a$, and take logarithm on both side:

$$\lim_{a\to 0} \frac{1}{a} \ln\left(\sum_{i=1}^{S} p_i p_i^a\right) = \sum_{i=1}^{S} p_i \ln p_i$$

The left hand side:

$$\lim_{a\to 0} \frac{1}{a} \ln\left(\sum_{i=1}^{S} p_i p_i^a\right)$$

$$= \lim_{a\to 0} \frac{1}{a} \ln\left(\sum_{i=1}^{S} p_i \exp\left(a \ln p_i\right)\right)$$

Then by result 1:

$$= \lim_{a\to 0} \frac{1}{a} \ln\left(\sum_{i=1}^{S} p_i (1 + a \ln p_i)\right)$$

$$= \lim_{a\to 0} \frac{1}{a} \ln\left(\sum_{i=1}^{S} \left(p_i + p_i a \ln p_i\right)\right)$$

$$= \lim_{a\to 0} \frac{1}{a} \ln\left(\sum_{i=1}^{S} p_i + a \sum_{i=1}^{S} p_i \ln p_i\right)$$

Note that $\sum_{i=1}^{S} p_i = 1$, and by result 2:

$$= \lim_{a \to 0} \frac{1}{a} a \sum_{i=1}^{S} p_i \ln p_i = \sum_{i=1}^{S} p_i \ln p_i$$

Therefore, we have shown that $\lim_{q \to 1} \left(\sum_{i=1}^{S} p_i^q \right)^{\frac{1}{1-q}} = \exp\left(- \sum_{i=1}^{S} p_i \ln p_i \right)$. □

The most commonly used diversity indices are derived from N_q as they can be expressed as monotonic functions of $\sum_{i=1}^{S} p_i^q$.

Theorem 2.1.2. *Suppose A is a continuous and monotonic diversity measure. The value of N_q depends on q and the frequency of the species p_i, it is independent of A.*

Proof. Assume any index of the dataset $A(\sum_{i=1}^{S} p_i^q)$ has some certain value a. x is the value of true diversity. Therefore, $\sum_{i=1}^{x} (\frac{1}{x})^q = \sum_{i=1}^{S} p_i^q$. Now we have:

$$A\left(\sum_{i=1}^{x} (\frac{1}{x})^q \right) = a$$

$$A\left(x(\frac{1}{x})^q \right) = a$$

$$A\left((\frac{1}{x})^{(q-1)} \right) = a$$

$$x^{(1-q)} = A^{-1}(a)$$

$$x = \left[A^{-1}(a) \right]^{\frac{1}{(1-q)}}$$

Substitute a with $A\left(\sum_{i=1}^{S} p_i^q \right)$ yields:

$$x = \left[A^{-1}\left(A(\sum_{i=1}^{S} p_i^q) \right) \right]^{\frac{1}{1-q}} = \left(\sum_{i=1}^{S} p_i^q \right)^{\frac{1}{(1-q)}}$$

x is independent of the function A. □

q from (2.1) and (2.2) is called the order of diversity. It shows how sensitive the diversity measure is to abundant and rare types. A true diversity with order 0 ignores the frequency of each species which represents the richness of the system. Values of q that are less than 1 favour the rare species while values of q greater than 1 add more weight on abundant species.

Theorem 2.1.3. *The value of N_q gets doubled for any order q except for order 1 if the system is divided equally into two groups and treated as separate "species".*

Proof. Let $\sum_{i=1}^{S} p_i^q = a$. After doubling the community, the frequency of each group gets halved and the size of the group gets doubled. Therefore, the sum of doubled community becomes:

$$\sum_{i=1}^{S} 2\left(\frac{p_i}{2}\right)^q = 2^{(1-q)} \sum_{i=1}^{S} p_i^q = 2^{(1-q)} a$$

. Noting that the original diversity: $N_q = \left(\sum_{i=1}^{S} p_i^q\right)^{\frac{1}{q-1}} = a^{\left(\frac{1}{1-q}\right)}$. The diversity for the doubled community, by (2.1) is:

$$N_q^\star = \left(\sum_{i=1}^{S} 2\left(\frac{p_i}{2}\right)^q\right)^{\left(\frac{1}{1-q}\right)} = \left(2^{1-q}a\right)^{\left(\frac{1}{1-q}\right)} = 2\left(a^{\left(\frac{1}{1-q}\right)}\right) = 2 \times N_q$$

□

This doubling property is proposed by Hill in 1973 [11], it is independent of the frequency of species.

2.2 Simpson's Index (D)

The Simpson's index is proposed by Simpson in 1949. If two substances are randomly selected from the community, the probability that they are of the same type is:

$$D = \sum_{i=1}^{S} p_i^2 \tag{2.3}$$

D in equation (2.3), indeed, is the *Simpson's Index*. It is also known as *Herfindahl-Hirschman index* (HHI) in many economics literature. The reciprocal of Simpson index $\frac{1}{D} = \left(\sum_{i=1}^{S} p_i^q\right)^{-1}$ is a true diversity of order 2 for the system. The complement $1 - D$ captures the uncertainty that two substances are of different types. Therefore, the smaller value of simpson's index indicates a more diverse system while greater value of Simpson's index suggests a more concentrated system.

Equation (2.3) assumes the first substance taken is replaced to the dataset before the second selection. For small data set, if sampling without replacement, the index is calculated as

$$\lambda = \frac{\sum_{i=1}^{S} n_i(n_i - 1)}{N(N-1)} \tag{2.4}$$

2.3 Shannon's Index (H)

Shannon Index is originally proposed to measure the uncertainty of occurrence of letters in strings (Shannon 1948). Given a community with relative abundance p_is, a diversity measurement, or measure of entropy, should possess the following properties[17]:

- H is a continuous function of p_i.

- If $p_i = \frac{1}{S}$ for all i, H is a monotonic increasing function of S.
- If a choice is decomposed into two successive choices, the original H is the weighted sum of individual values of H.

Theorem 2.3.1. *The only H satisfying the three above assumptions is of the form:*

$$H = -K \sum_{i=1}^{S} p_i \ln p_i \tag{2.5}$$

where K is a positive constant.

Proof. Denote $H\left(\frac{1}{S}, \frac{1}{S}, \ldots, \frac{1}{S}\right) = A(S)$. According to the third property, a choice from s^m equally likely possibilities can be decomposed into m choices of s equally likely possibilities and obtain

$$A(s^m) = mA(s).$$

We can construct another series of events with

$$A(t^S) = SA(t).$$

S could be arbitrarily large and we can always find an m such that

$$s^m \leq t^S \leq s^{m+1}.$$

Taking logarithms and divided by $S \log s$ yields

$$\frac{m}{S} \leq \frac{\log t}{\log s} \leq \frac{m}{S} + \frac{1}{S}$$

equivalently

$$\left|\frac{m}{S} - \frac{\log t}{\log s}\right| < \epsilon. \tag{2.6}$$

where ϵ is arbitrarily small. The monotonic increasing property of $A(S)$ yields

$$A(s^m) \leq A(t^S) \leq A(s^{m+1})$$

or equivalently

$$mA(s) \leq SA(t) \leq (m+1)A(s).$$

Divided by $SA(s)$, we obtain

$$\frac{m}{S} \leq \frac{A(t)}{A(S)} \leq \frac{m}{S} + \frac{1}{S}$$

or

$$\left|\frac{m}{S} - \frac{A(t)}{A(s)}\right| < \epsilon. \tag{2.7}$$

Equation (2.6) and (2.7) yields

$$-\epsilon < \frac{\log t}{\log s} - \frac{m}{S} < \epsilon \tag{2.8}$$

and

$$-\epsilon < \frac{m}{S} - \frac{A(t)}{A(s)} < \epsilon \tag{2.9}$$

(2.8) and (2.9) yields

$$-2\epsilon < \frac{\log t}{\log s} - \frac{A(t)}{A(s)} < 2\epsilon$$

$$\left|\frac{A(t)}{A(s)} - \frac{\log t}{\log s}\right| < 2\epsilon.$$

Therefore, $A(t) = K \log t$, where K is positive in order to satisfy the second property. Suppose we have a choice from S probabilities with probability $p_i = \frac{N_i}{\sum N_i}$. A choice from $\sum N_i$ possibilities can be decomposed into a choice from S possibilities $p_1, p_2, \ldots, p_S$. If the i^{th} is chosen, then followed by a choice from N_i with equal probabilities. According to the third property,

$$K \log \sum N_i = H(p_1, p_2, \ldots, p_S) + K \sum p_i \log N_i,$$

therefore,

$$\begin{aligned} H(p_1, p_2, \ldots, p_S) &= K \left[\sum p_i \log \sum N_i - \sum p_i \log N_i \right] \\ &= -K \sum p_i \log \frac{N_i}{\sum N_i} \\ &= -K \sum p_i \log p_i. \end{aligned}$$

□

The choice of coefficient K is a matter of convenience and amounts to the choice of a unit of measure (Shannon, 1948). The form

$$H = -\sum_{i=1}^{S} p_i \ln p_i. \tag{2.10}$$

is mostly recognized as Shannon entropy and it is widely used in ecological literature. The logarithmic base corresponds to the unit of the information measured. Base 2 measures information with binary units. If the units is decimal digits, base 10 is used. The natural logarithm is usually used when the analysis involves integration and differentiation. $\exp H$ produces a true diversity of order 1 (N_1) of a system.

If all $p_i = 0$ except for one substance with probability 1, there is no uncertainty remains in the system, therefore, $H = 0$. H is always positive otherwise. On the other hand, if all substances has equal probability ($p_i = \frac{1}{S}$ for S types in the system), the uncertainty of the outcome reaches the maximum $H_{\max} = -\sum_{i=1}^{S} \frac{1}{S} \ln \frac{1}{S} = \ln S$. Any "averaging" operation on p_i of the form

$$p_i^\star = \sum_j a_{i,j} p_j$$

where $\sum_i a_{i,j} = \sum_j a_{i,j} = 1$ and $a_{i,j} > 0$, will increase the value of H. H will remain the same if the operation is only a permutation of the p_j.

Given two sets of events x and y with m and n possible outcomes respectively. Denote $p_{i,j}$ as the joint probability of substance i in x, j in y. The joint entropy of x and y is

$$H(x, y) = -\sum_{i,j} p_{i,j} \ln p_{i,j}. \tag{2.11}$$

while

$$H(x) = -\sum_i p_{i\cdot} \ln p_{i\cdot}$$

$$H(y) = -\sum_j p_{\cdot j} \ln p_{\cdot j}$$

where

$$p_{i\cdot} = \sum_j p_{i,j}$$

$$p_{\cdot j} = \sum_i p_{i,j}$$

It is evident that

$$H(x, y) \leq H(x) + H(y). \tag{2.12}$$

The equality holds when x and y are independent. The inequality implies that the entropy of joint event is less than or equal to the sum of individual entropy.

The conditional entropy of y ($H_x(y)$) is interpreted as the average uncertainty of y when x is known.

$$H_x(y) = -\sum_{i,j} p_{i,j} \ln p_{j|i}. \tag{2.13}$$

where $p_{j|i} = \frac{p_{i,j}}{\sum_j p_{i,j}}$ is the conditional probability of y when x is given. Substitute $p_{j|i}$ with $\frac{p_{i,j}}{\sum_j p_{i,j}}$ in (2.13) yields

$$H_x(y) = -\sum_{i,j} p_{i,j} \ln p_{i,j} + \sum_{i,j} p_{i,j} \ln p_{i\cdot} = H(x,y) - H(x).$$

Therefore,

$$H(x,y) = H_x(y) + H(x). \tag{2.14}$$

The equation implies that the joint entropy of events x and y is the sum of the entropy of x and entropy of y when x is known.

(2.12) and (2.14) yields

$$H(x) + H(y) \geq H(x,y) = H_x(y) + H(x)$$

$$H(y) \geq H_x(y). \tag{2.15}$$

which implies that event y is less uncertain if the information of x is obtained. The value of uncertainty remains unchanged when x and y are independent. Introducing x will never increase the uncertainty of y.

When the information analysed is continuous variable, the Shannon's entropy is

defined as

$$H = -\int_{-\infty}^{\infty} p(x) \ln p(x) dx \tag{2.16}$$

Where $p(x)$ is the density function.

As for an n multi-dimensional distribution, $p(x_1, \ldots, x_n)$, the entropy is given by

$$H = -\int \cdots \int p(x_1, \ldots, x_n) \ln p(x_1, \ldots, x_n) dx_1, \ldots, dx_n \tag{2.17}$$

The continuous Shannon's entropy preserves most properties from discrete case, however unlike discrete information system, the diversity measure of continuous variable is a relative measure according to the coordinate system. Change of variable in general results in a change of entropy. Suppose we change coordinate from x to y, the new entropy

$$\begin{aligned} H(y) &= \int \cdots \int p(x_1, \ldots, x_n) J\left(\frac{x}{y}\right) \ln p(x_1, \ldots, x_n) J\left(\frac{x}{y}\right) dy_1, \ldots, dy_n \\ &= H(x) - \int \cdots \int p(x_1, \ldots, x_n) J\left(\frac{x}{y}\right) dx_1, \ldots, dx_n \end{aligned}$$

where $J\left(\frac{x}{y}\right)$ is the Jacobian of transformation.

2.4 Berger–Parker Dominance (BP)

Berger–Parker Dominance is a single measure of diversity which takes the value of most dominate frequency (Berger and Parker 1970).

$$\text{BP} = p_{\max} \tag{2.18}$$

BP corresponds to the true diversity of order infinity (N_∞).

2.5 Evenness (E)

The evenness reflects the structure of a system, it indicates the balance of all species, more precisely, "the degree to which the abundances are equal among the species present in a sample or community" (Molinari 1989). The quantity proposed by Pielou (1969)

$$J = \frac{H}{H_{\max}} \tag{2.19}$$

captures such ability. Where H is the Shannon's index and the maximum of H achieves, as mentioned previously, when all substances have equal probability $H_{\max} = \ln S$. Hill (1973) proposed a double continuum of measures of evenness

$$E_{a,b} = \frac{N_a}{N_b} \tag{2.20}$$

Where a and b are all possible value of the order of true diversity N_q. Although $J = \frac{\ln N_1}{\ln N_0}$ does not satisfy (2.20), the alternate $J' = H - H_{\max} = \ln \frac{N_1}{N_0} = \ln E_{1,0}$ meets the requirement. $E_{1,0}$ measures the ratio of abundant species to all species. However, as N_1 are too dependent on sample size, $E_{2,1}$, the ratio of very abundant species to abundant species provides more stable analysis.

Bulla (1994) points that the above measures of evenness are inadequate as they overestimate evenness. Inferences are hard to made because of the non-linear behaviour of these indices. Although later works such as $G_{2,1}$ proposed by Molinari (1989) improves the linearity, they are lack of any clear ecological meaning and the statistical property is hard to capture. Moreover, Molinari's measure ignores the

behaviour of rare species in community which may mislead discrimination between communities whose difference is driven mainly by rare species[3]. Therefore, a new measure of evenness is proposed.

$$E = \frac{O - 1/S}{1 - 1/S} \tag{2.21}$$

where $O = \sum \min(p_i, 1/S)$ The newly defined measure presents good sensitivity to rare species compared with other measures.

The value of evenness varies from 0 to 1, the low value indicates the system is dominated by one or a few species while high value suggest the number of individuals for each species are relatively equal.

2.6 Choice of Indices

There has been a lot of debates over whether compound indices (e.g. Shannon index, Simpson index, etc.) provide better interpretation than single measure of diversity (e.g. richness, Berger–Parker Dominance, etc.) and which is more appropriate in different contexts. Since most of the indices are derived from N_q, there exists strong correlations between different diversity measures (Morris et al. 2014). Although different indices are all representations of system's diversity, the choice of true diversity's order q can alter the result significantly. The information of a system can never be fully captured by only one metric, therefore, "no single measure will always be appropriate" (Purvis and Hector 2000). The choice of q, in a way, indicates the weight put on abundant v.s. rare species. BP only focuses on the most abundant species and Simpson's measure D favours relative abundant species, therefore, the effect that is

driven by abundant species is well captured by these indices. However, the behaviour of rare species may be better detected by those indices who favours rare species such as S. As for estimating the total number of species, the compound indices such as D provides better estimation than single measures (e.g. S) as the latter are too dependent on sample size. Other than those indices mentioned above, evenness E may carry different information as it shows inconsistent correlation between other diversity indices when analysing with different data

The choice of indices can depend on the area of study, the type of data as well as the importantness of abundant and rare species. We cannot use only one index to discover all characters of a community. Therefore, at least two indices are suggested to report (Whittaker 1972; Stirling and Wilsey 2001; Heino et al. 2008).

2.7 An Example of Application of Diversity Indices

According to the latest United Nation estimates, China has a population of 1.4 billion which ranks the first in the list of countries by population. The United States possesses a population of 0.328 billion according to the report by the United States census Bureau. The latest report by the United States census Bureau in 2010 [5] shows that about 6.3 million surnames were reported while only about 600 surnames were found in the sixth national population census of the People's Republic of China in 2010. As a nation of immigrants, it is not surprising that the United States is very diverse in surnames. On the contrary, China has much more conservative immigration policy, foreign surnames are rarely seen in China. In ancient China, only ruling family

and upper class were qualified to have surnames, people from lower class were usually given surnames by the family they serve. Therefore, we can see several surnames dominate the whole population in China. Hence, it is expected that USA has greater surname diversity than China. In this section, we are going to use diversity measures introduced in order to investigate the surname structure of these two countries.

2.7.1 Material and Method

The data of surnames in the United State was obtained from the United States census Bureau (`https://www2.census.gov/topics/genealogy/2010surnames/names.zip`) in 2010. It contains 162,253 surnames ranked by their population which accumulates to 90.06% of the population across the nation. The data of surnames in China was obtained from the sixth national population census in 2010. It contains 400 surnames which accumulates to 94.17% of the whole population (See data in Appendix B, Table B.1 to Table B.5). The most common surname in USA is "Smith" which accounts for 0.828% of the population. "Wang" ranks the first in China and takes up 6.96% of the whole population.

The surname in USA has greater richness than that of China. Figure 2.1 is the rank abundance curve which plot the proportion of the surnames against their rank. A steep gradient appears in both rank abundance curves indicates that high-ranking surnames are much more abundant than low-ranking surnames in both USA and China. The gradient of curve of USA is steeper which means this unbalanced structure is more apparent across USA.

We calculate the complement Simpson's Index($1 - D$), Shannon's Index as well as their corresponding true diversity N_2 and N_1 in order to investigate the name

diversity of two countries. We also include evenness to verify the conclusion drawn from the rank abundance curve. Equation 2.21 is used to calculate evenness. The following table provides calculation results.

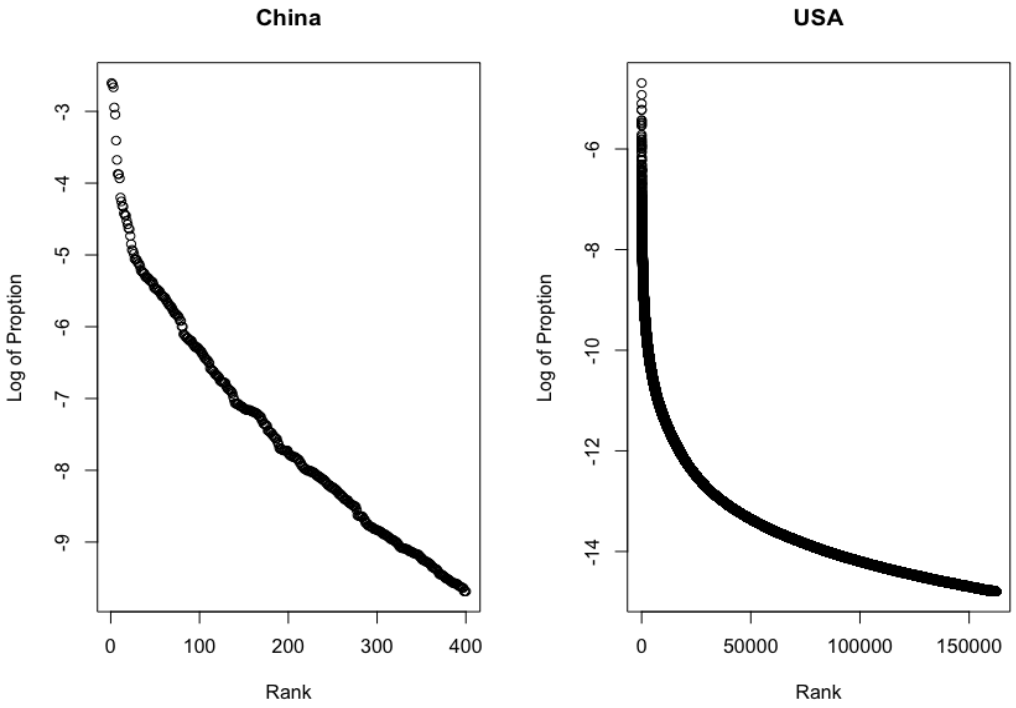

Figure 2.1: Rank Abundance Curves of Surnames of USA and China

Country	S	$1-D$	N_2	H	N_1	E
USA	162253	0.9993054	1439.74440	9.425931	12405.94544	0.2847087
China	400	0.9733317	37.49772	4.500596	90.07077	0.3655664

Table 2.1: Diversity Measures of Surnames across USA and China

From table 2.1 we can see that USA has a Simpson's index of value 0.9993 while China has Simpson's index of value 0.9733, which suggests that USA has higher surname diversity than China. The corresponding true diversity N_2 where USA (1439.74)

has much higher value than China (37.50) makes the conclusion more convincing. It is also evident to conclude that China has less surname diversity than USA using the Shannon's Index and N_1. By looking at evenness, both USA and China appear to have low evenness value while USA has even lower value. This implies that dominant surnames exist in both countries, and USA has more unbalanced surname structure.

It is worth noticing that when change the diversity measure from Simpson's index to Shannon's index, the corresponding true diversity increases from 1439.74 to 12405.95 for USA and 37.50 to 90.07 for China as more weight is put on relative low proportion surnames. The increase appears more significant in USA than China, again, is due to the more unbalanced structure of USA.

Chapter 3

Investment Diversity of Mutual Fund

Diversification is considered as an important factor of investing. Faulkenberry[7] defines it as "a portfolio strategy combining a variety of assets to reduce the overall risk of an investment portfolio." Demsetz and Strahan (1997) point out that more diversified bank holding companies (BHCs) have advantage in operating with lower capital ratios and conducting risky activities. Moreover, geographic diversified BHCs are exposed to less idiosyncratic local risk and expansion into economically dissimilar areas helps reducing risk more (Goetz et al. 2014).

The investment portfolio of different mutual fund categories varies. For example, precious metals equity in general have very concentrated investment sector while Canadian equity have much more diversified sector allocation. We can consider the total capital possessed by a mutual fund as "population", sectors that this fund invests are "species", therefore, the "relative abundance" of each "species" are the proportion of capital that enters in each sector. Then the investment diversity of

thus fund could be calculated based on the above settings. We are interested in whether higher investment return rate is associated with more diversified investing strategy.

3.1 Material and Method

The data is obtained from FUNDATA (`http://www.fundata.com/`) which provides the snapshots of mutual funds over Canada. We randomly sampled 60 funds from each of the following categories: U.S. Equity, Precious Metals Equity, Natural Resources Equity, Canadian Equity and Canadian Fixed Income. For each fund, we record its category, sector allocation and return rate. We pick year to date (YTD) return and one year annual return as two different response. The complement of Simpson's Index $(1 - D)$ and Shannon' Index as well as their corresponding true diversities are calculated according to each fund's sector allocation. We also include evenness in the analysis in order to see which index possesses better predicting ability. We plot return rate against each indices to visually investigate if any trend exist between investment diversity and return.

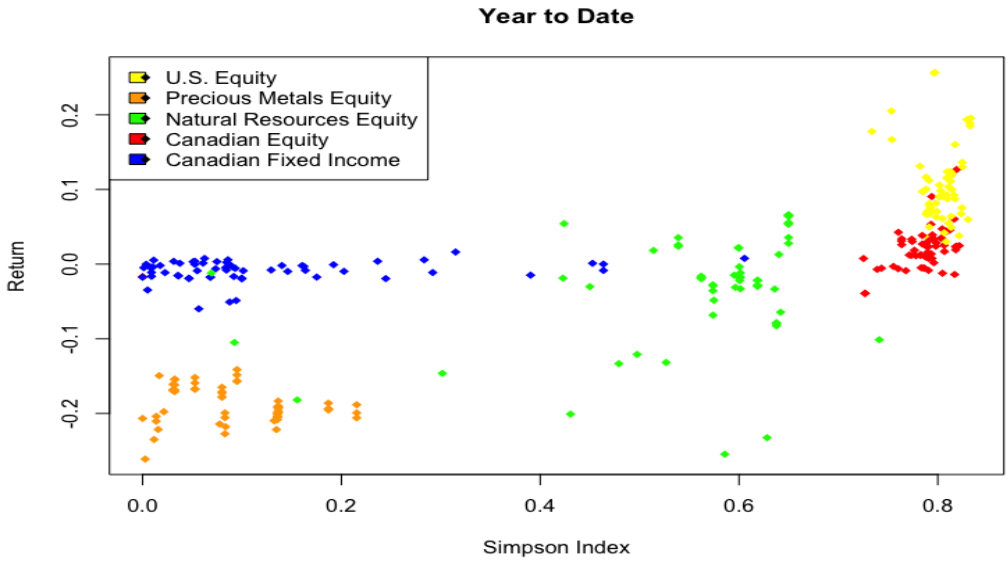

Figure 3.1: Year to Date Return

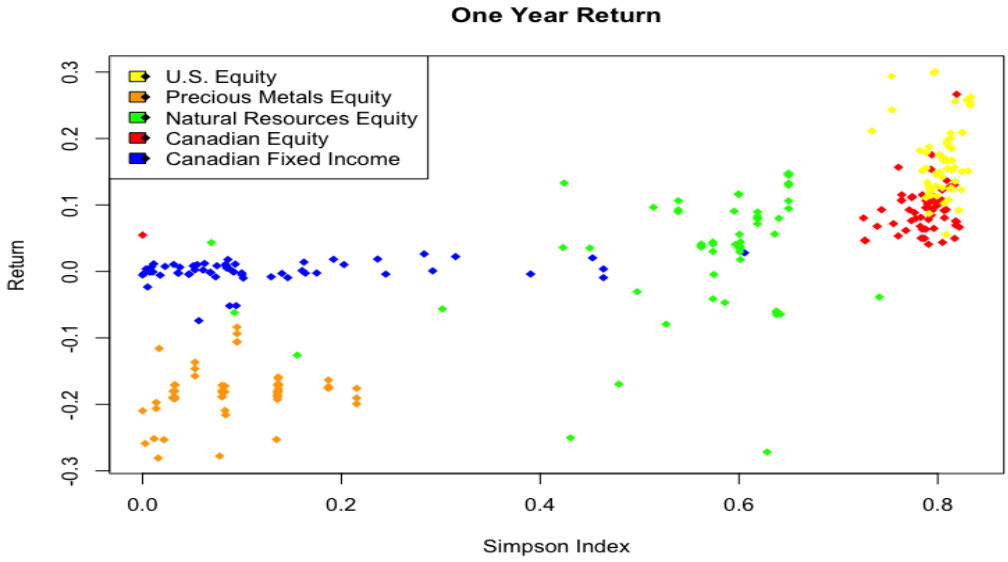

Figure 3.2: One Year Annual Return

The plots of other indices demonstrate the similar trend as Simpson's Index (see Appendix A). From Figures 3.1 and 3.2, we can see that the investment diversity

of different kind of funds varies from low to high. Precious metals equity overall has very concentrated investment sector and the return rate is the lowest among five categories. On the other hand, U.S equity has most diversified investment strategy and possesses the highest return rate. Although the data is not perfectly aligned, we can still see the trend that higher diversity leads to higher return rate. It is worth noticing that Canadian fixed income and precious metals equity have similar diversity index value however, the return rate of Canadian fixed income are in general, higher than that of precious metals equity. Moreover, both Canadian equity and U.S. equity have high diversity index value, the return rate of U.S. equity is higher on average. Therefore, we detect the individual effect from different categories. Based on that, we propose the model:

$$Y_i = \beta_0 + \beta_1 D + \beta_2 \delta_i + \epsilon \tag{3.1}$$

where Y_i represents the return rate, D is the diversity measure and δ_i is the categorical variable that represents the category of the mutual fund. We fit the model with five different diversity measures, the result is as following, $\star$ indicates how significant each variable is:

Model	Coefficient	t-value	p-value	AIC	R^2	δ_i (β, p-value)			
						Canadian Fixed Income	Natural Resources Equity	Precious Metals Equity	U.S. Equity
$1-D$	0.079383	3.146	0.001824$\star\star$	-1033.172	0.8384	0.022885, 0.208393	-0.035047, 0.000246$\star\star\star$	-0.155087, $5.31\times10^{-15}\star\star\star$	0.082486, $<2\times10^{-16}\star\star\star$
H	0.050132	4.035	$6.97\times10^{-5}\star\star\star$	-1039.407	0.8417	0.040892, 0.0312$\star$	-0.022642, 0.0330$\star$	-0.138579, $3.23\times10^{-12}\star\star\star$	0.081849, $<2\times10^{-16}\star\star\star$
N_1	0.024464	4.697	$4.07\times10^{-6}\star\star\star$	-1044.941	0.8446	0.066955, 0.002279$\star\star$	0.003825, 0.786324	-0.112007, $6.36\times10^{-7}\star\star\star$	0.078816, $<2\times10^{-16}\star\star\star$
N_2	0.020141	3.628	0.000336$\star\star\star$	-1036.379	0.8401	0.041606, 0.047100$\star$	-0.006367, 0.665892	-0.137202, $4.23\times10^{-10}\star\star\star$	0.075680, $<2\times10^{-16}\star\star\star$
E	0.060182	2.283	0.0232$\star$	-1028.507	0.8358	0.009126, 0.6191	-0.030161, 0.014$\star$	-0.168122, $4.34\times10^{-16}\star\star\star$	0.081789, $<2\times10^{-16}\star\star\star$

Table 3.1: Year to Date Return

Model	Coefficient	t-value	p-value	AIC	R^2	δ_i (β, p-value)			
						Canadian Fixed Income	Natural Resources Equity	Precious Metals Equity	U.S. Equity
$1-D$	0.115041	3.026	0.0027$\star\star$	-787.1372	0.7516	-0.017166, 0.5308	-0.036195, 0.0115$\star$	-0.177532, $1.30\times10^{-9}\star\star\star$	0.070917, $5.43\times10^{-9}\star\star\star$
H	0.066353	3.526	0.000489$\star\star$	-790.3655	0.7542	0.000187, 0.994789	-0.021884, 0.172733	-0.162443, $4.24\times10^{-8}\star\star\star$	0.070373, $5.86\times10^{-9}\star\star\star$
N_1	0.029406	3.701	0.000256$\star\star$	-791.5995	0.7552	0.023059, 0.487606	0.006384, 0.766790	-0.139048, $4.47\times10^{-5}\star\star\star$	0.067095, $3.29\times10^{-8}\star\star\star$
N_2	0.024265	2.881	0.00425$\star\star$	-786.292	0.7509	-0.007218, 0.81981	-0.005744, 0.79735	-0.169137, $2.88\times10^{-7}\star\star\star$	0.063301, $5.61\times10^{-7}\star\star\star$
E	0.10127	2.558	0.011$\star$	-784.54	0.7494	-0.02827, 0.305	-0.02406, 0.194	-0.18690, $6.94\times10^{-10}\star\star\star$	0.06919, $1.88\times10^{-8}\star\star\star$

Table 3.2: One Year Annual Return

As shown in Tables 3.1 and 3.2, all diversity measures are significant. The coefficients of five diversity indices are positive which indicates that more diversified investment strategy leads to higher return rate (for both YTD and one year annual return). The effect of individual category cannot be ignored. The model sets Canadian equity as default group, the effect of precious metals equity and U.S. equity are consistently significant through all diversity measures. Our analysis suggests that precious metals equity has lower return rate than Canadian equity and U.S. equity on the other hand, produces higher return than Canadian equity. However, other mutual fund categories do not perform such consistent behaviour.

The effect of Canadian fixed income on YTD is significant when using H, N_1 and N_2. It shows that YTD increases when switch fund category from Canadian equity to Canadian fixed income. Similarly, the effect of natural resources equity on YTD is significant when D, H and E are applied as diversity measure. YTD decreases when switch fund category from Canadian equity to natural resources equity. As for one year annual return, the negative effect of natural resources equity appears to be significant only when applying D as diversity measure. Otherwise, Canadian fixed income and natural resources equity appear to have same behaviour as Canadian equity.

Five diversity measures appear to be significant, however, E is the least significant measure. AIC and multiple r-squared of different models are also very close. We suggest to use Shannon's Index as the predictor of YTD and Simpson's Index as the predictor of one year annual return because these two models present more information on individual category effect. The residual plots (see Appendix A) also shows that the models perform good fit.

The sector allocation of each fund contains a category "Other". After removing "Other" and normalizing the proportion of other sectors, we obtain the same analysis results as previous.

3.2 Conclusion

We have successfully find relationship between investment diversity and return rate. The small p-value and positive coefficient suggest that more diversified investment strategy is preferred when expecting higher return. Shannon's Index behaves the best for YTD and Simpson's Index is preferred for year to date return. However, the coefficients of diversity measures in general are very small, an increase in diversity only accounts for a very small increase in return. Moreover, Canadian equity and U.S. equity over all have similar diversity pattern, our analysis shows that the return rate of U.S equity is significantly higher than that of Canadian equity. Therefore, the effect of individual category cannot be ignored. Highly diversified investment does not translate into high return directly, the market where equity enters also plays an important role when analysing return. Our analysis reveals the potential of investment diversity and provides motivation for diversifying investment strategy.

Chapter 4

Discrimination Using Diversity Indices

Diversity indices reflect certain characters of a community. Their values vary among different communities. Many ecology literatures have used diversity indices to differentiate groups. Morris et al. (2014) point out that the Simpson's Index and its corresponding true diversity N_2 perform the best ability in differentiating grassland plots using principle component analysis (PCA).

As shown in Figure 3.1, mutual funds that are from same category appear to be "clustered" together, they share the similar index values and return rates. Therefore, diversity measures along with return rate may provide good ability in discriminating mutual fund categories.

4.1 Material and Method

We choose Simpon's Index and YTD in our analysis. From Figure 3.1, we do not find spherical-shaped data distribution. Canadian fixed income appears to have a line-shaped distribution and natural resources equity is scattered. Therefore, we suspect that k-means and k-medoids might not be sufficient in our analysis as they are sensitive to outliers, especially, k-means.

4.1.1 K-means Clustering

K-means clustering is an un-supervised clustering algorithm that defines k centroids among the data set then associates points that are close to the same centroid into one group. The algorithm is sensitive to the initial centroids selected and the distribution of the data set. Although the data distribution is not suitable for k-means cluster analysis, we want to know how bad the miss-classification would happen.

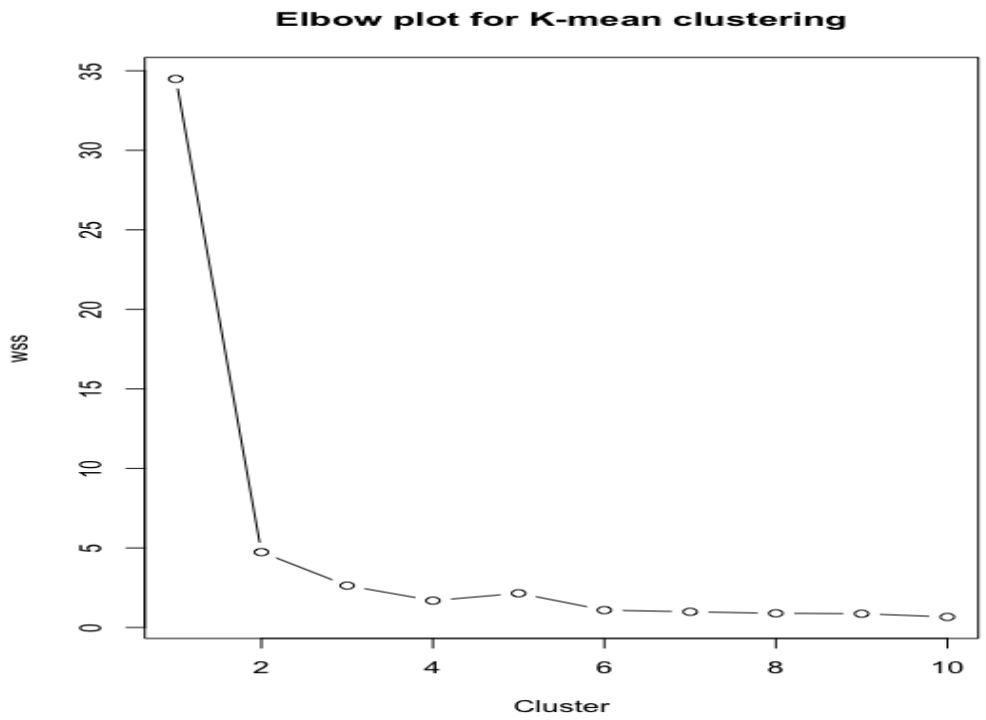

Figure 4.1: K-means Elbow Plot

The elbow plot suggests two clusters. As we already know there are five different categories, k-means method fails to detect all clusters. Noting that the within-cluster sum of squares has a fluctuation at 5 clusters. We set the number of clusters to 5, the result is as following:

	Cluster 1	Cluster 5	Cluster 4	Cluster 2	Cluster 3
Canadian Equity	34	14	0	12	0
Canadian Fixed Income	3	47	9	1	0
Natural Resources Equity	0	2	46	11	1
Precious Metals Equity	4	0	0	50	6
U.S. Equity	4	0	5	1	50

Table 4.1: Cluster Table K-means

The miss-classification is 24.33% which is not really bad considering the unsuitable data distribution. The algorithm provides passable classification for most mutual fund, the miss-classification mainly comes from Canadian equity. 14 funds from Canadian equity are miss-classified into Canadian fixed income and 12 funds are miss-classified into precious metals equity.

4.1.2 K-medoids Clustering

Instead of taking the mean as centroid, k-medoids clustering picks points from the given data set as reference points. K-medoids method is less sensitive to outliers compared to k-means method, both two methods have similar algorithm. Figure 4.2 suggests 2 clusters for k-medoids method. Again, k-medoids fails to detect actual number of categories of mutual fund. Setting the cluster number to 5 produces very bad classification.

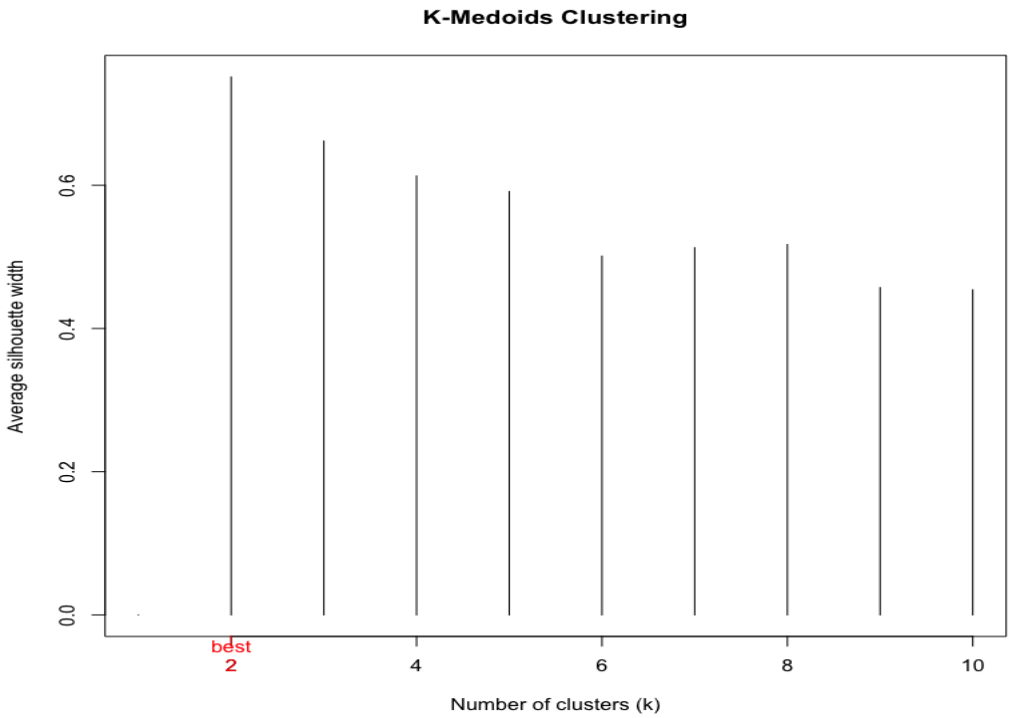

Figure 4.2: K-medoids Average Silhouette Plot

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Canadian Equity	59	0	0	0	1
Canadian Fixed Income	0	0	0	5	55
Natural Resources Equity	1	1	2	55	1
Precious Metals Equity	0	32	28	0	0
U.S. Equity	60	0	0	0	0

Table 4.2: Cluster Table K-medoids

The miss-classification rate is 33%. The algorithm fails to differentiate Canadian equity and U.S. equity also separates precious metals equity into two clusters. A similar result is obtained when we use hierarchical clustering (see result in Appendix A Figure A.11 and Appendix B Table B.6).

So far, all unsupervised clustering methods are inadequate, we propose that supervised clustering method is preferred.

4.1.3 Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) is a supervised clustering algorithm that separates different groups of data through a linear transformation then associates points to the group that maximizes the group density. Unlike PCA, we want to find dimensions that maximize the separation of data from different categories. Therefore, we use LDA to achieve such goal.

We randomly sample 150 mutual funds as training set, the prior probability of each category is set as 0.2. Two discriminants are found when the analysis is performed (see R output in Appendix C). After applying this model on the testing set, the predicting table shows a very good classification.

	Canadian Equity	Canadian Fixed Income	Natural Resources Equity	Precious Metals Equity	U.S. Equity
Canadian Equity	24	0	0	0	0
Canadian Fixed Income	0	30	1	0	0
Natural Resources Equity	3	1	19	1	3
Precious Metals Equity	0	0	0	35	0
U.S. Equity	6	0	0	0	27

Table 4.3: Classification Table LDA

LDA reduces miss-classification rate to 10% which has the best performance so far. The miss-classification is mainly driven by U.S.equity and natural resources equity. 6 mutual funds from U.S.equity are miss-classified into Canadian equity as the data of these two categories are distributed very closely. The scattered distribution of natural resources equity is also the cause of miss-classification. After linear transformation, we

can see the separation between different categories is more apparent in the "rotated" data set as the data from same category is more centralized (see Appendix A Figure A.12).

4.1.4 Classification Tree

Classification tree is another supervised classification strategy that allows us to investigate how each predicting variable contributes to classification. As each fund category contains equal amount of observations, we suggest a stratified sampling method to build decision tree. For each category, we randomly sample 40 mutual funds as training set, we build the model based on this stratified training set. R output can be found in Appendix C. The tree building stops at 4 (nsplit = 4), splitting the next node only decreases overall lack of fit by a factor of 0.01. The relative error for 4 splits is 0.056 indicating a good enough split. Under such splitting rule, a decision tree is built. Our result shows a fairly good classification on training set. Only 5% of the data are miss-classified into natural resources equity which originally belong to Canadian fixed income. There are 2% of the data in Canadian fixed income that are actually Canadian equity and 2% are natural resources equity. Precious metal equity are all classified correctly without any miss-classification. As for Canadian equity, 92% of the data are in the right spot while 2% of them are natural resources equity and 5% are from U.S. equity. 95% of the data in U.S equity are correctly classified while 5% originally come from Canadian equity. Therefore, Simpson's Index and YTD are considered as good classification predictors. We now apply the decision rule defined by classification tree on our testing set, a predicting table is shown below.

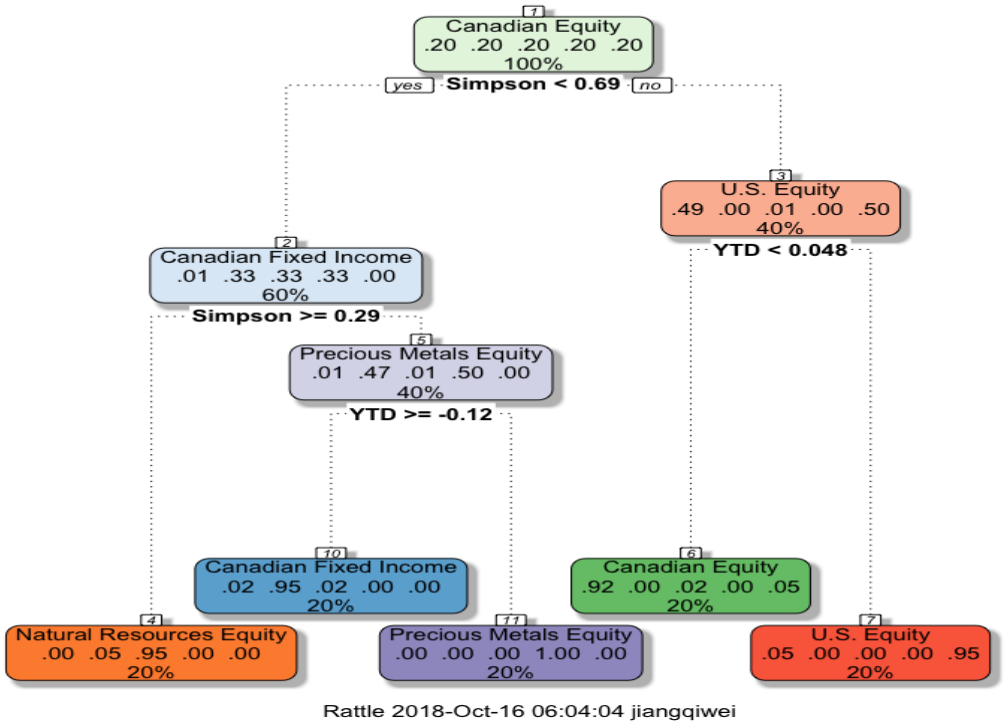

Figure 4.3: Classification Tree

	Canadian Equity	Canadian Fixed Income	Natural Resources Equity	Precious Metals Equity	U.S. Equity
Canadian Equity	18	0	0	0	2
Canadian Fixed Income	0	16	4	0	0
Natural Resources Equity	0	1	18	1	0
Precious Metals Equity	0	0	0	20	0
U.S. Equity	2	0	0	0	18

Table 4.4: Classification Tree Predicting Table

Table 4.4 shows a good predicting ability with only 10% miss-classification rate. Classification tree performs as good result as LDA. Therefore, we conclude that Simpson's Index and YTD have good ability in discriminating mutual fund categories when using supervised clustering analysis.

4.2 Conclusion

We used Simpson's Index and YTD to differentiate category of mutual fund successfully. With unsupervised clustering analysis, the predicting ability is not satisfying. K-means, k-mediods and hierarchical clustering produce miss-classification rate of 24%, 33% and 44% respectively. Because of linear trend distribution of Canadian fixed income and scattered distributed natural resources equity, the distance of a point to its associated centroid could be large enough for k-means and k-mediods algorithms to exclude this point from its actual category. On the other hand, Canadian equity and U.S.equity are too close so that hierarchical clustering would miss-classify them as the same category. LDA transforms data along the dimensions that maximize the septation between different category. As shown in LDA, the mean of Simpson's Index and YTD are significantly different among five categories. After linear transformation, the character driven by Simpson's Index and YTD becomes more apparent, therefore, the predicting ability is very strong. Classification tree provides as good predicting ability as LDA. Our algorithm uses four splits to clearly discriminate all five categories. U.S equity and Canadian Equity have higher diversity than other mutual fund category, according to the classification rule, funds with Simpson's index greater than 0.69 are classified into these two categories. The algorithm differentiate between U.S equity and Canadian equity by evaluating the fund's YTD. Funds with YTD that less than 0.048 are recognized as Canadian equity. Natural resources equity has the highest diversity among the rest three categories. Funds with Simpson's index value between 0.29 and 0.69 are classified into natural resources equity without evaluating its YTD. Precious material equity shares similar diversity pattern with

Canadian fixed income equity, funds with YTD lower than -0.12 are classified into Canadian fixed income. With appropriate strategy, diversity indices do have profound ability of discrimination.

Chapter 5

Bayesian Approach to Shannon's Index and Simpson's Index

Suppose we obtain a sample that contains n observations from a population that falls into S categories (S is known). We denote n_i as the observed number of individuals in each category i from sample. In order to estimate the diversity indices of population, the unknown frequency p_i needs to be estimated. Rather than the maximum likelihood estimator $\hat{p}_i = \frac{n_i}{n}$, we suggest a Bayesian approach to estimate p_i. The p_i's from previous chapters are considered fixed, in this chapter, p_i's are considered as random variables.

Dirichlet distribution is a well known conjugate prior in Bayesian statistics. Therefore, we assume the prior density of $p_1, p_2, \ldots, p_S$ is

$$f(p_1, p_2, \ldots, p_S|\alpha) = \frac{\Gamma(\alpha S)}{[\Gamma(\alpha)]^S} \prod_{i=1}^{S} p_i^{\alpha-1} \tag{5.1}$$

We take the special case of Dirichlet distribution where the parameter α_i for each p_i

has the same positive value because we assume no prior knowledge favouring any p_i over another. The likelihood:

$$L(p_1, p_2, \ldots, p_S | n_1, n_2, \ldots, n_S) = \binom{n}{n_1 n_2 \ldots n_S} \prod_{i=1}^{S} p_i^{n_i} \tag{5.2}$$

Therefore, the posterior joint density of $p_1, p_2, \ldots, p_S$ is

$$f(p_1, p_2, \ldots, p_S | \alpha; n_1, n_2, \ldots, n_S)$$

$$= \frac{\frac{\Gamma(\alpha S)}{[\Gamma(\alpha)]^S} \prod_{i=1}^{S} p_i^{\alpha-1} \binom{n}{n_1 n_2 \ldots n_S} \prod_{i=1}^{S} p_i^{n_i}}{\int_0^1 \int_0^1 \cdots \int_0^1 \frac{\Gamma(\alpha S)}{[\Gamma(\alpha)]^S} \prod_{i=1}^{S} p_i^{\alpha-1} \binom{n}{n_1 n_2 \ldots n_S} \prod_{i=1}^{S} p_i^{n_i} dp_1 dp_2 \ldots dp_S}$$

$$= \frac{\prod_{i=1}^{S} p_i^{n_i+\alpha-1}}{\int_0^1 \int_0^1 \cdots \int_0^1 \prod_{i=1}^{S} p_i^{n_i+\alpha-1} dp_1 dp_2 \ldots dp_S}$$

Let $\boldsymbol{\alpha}^\star = (\alpha_1^\star, \alpha_2^\star, \ldots, \alpha_S^\star)$, where $\alpha_i^\star = n_i + \alpha$.

$$f(p_1, p_2, \ldots, p_S | \boldsymbol{\alpha}^\star; n_1, n_2, \ldots, n_S) = \frac{\prod_{i=1}^{S} p_i^{\alpha_i^\star - 1}}{B(\boldsymbol{\alpha}^\star) \int_0^1 \cdots \int_0^1 \frac{1}{B(\boldsymbol{\alpha}^\star)} \prod_{i=1}^{S} p_i^{\alpha_i^\star} dp_1 dp_2 \ldots dp_S}$$

$B(\boldsymbol{\alpha}^\star) = \frac{\Gamma\left(\sum_{i=1}^{S} \alpha_i^\star\right)}{\prod_{i=1}^{S} \Gamma\left(\alpha_i^\star\right)}$ is the beta function. The integral part in the denominator yields 1, therefore, the posterior joint density of $p_1, p_2, \ldots, p_S$:

$$f(p_1, p_2, \ldots, p_S | \boldsymbol{\alpha}^\star; n_1, n_2, \ldots, n_S) = \frac{1}{B\left(\boldsymbol{\alpha}^\star\right)} \prod_{i=1}^{S} p_i^{\alpha_i^\star - 1} \tag{5.3}$$

which is still Dirichlet distribution with updated parameters $\alpha_i^\star$'s.

We now can derive the marginal density of each p_i from equation (5.3).

$$
\begin{aligned}
&f(p_i|\boldsymbol{\alpha}^\star; n_1, n_2, \ldots, n_S) \\
&= \frac{1}{B(\boldsymbol{\alpha}^\star)} \int_0^{1-p_i} \int_0^{1-p_i-p_1} \cdots \int_0^{1-\sum_{j=1}^{S-2} p_j} \prod_{j=1}^{S-1} p_j^{\alpha_j^\star - 1} \left(1 - \sum_{j=1}^{S-1} p_j\right)^{\alpha_S^\star - 1} dp_{S-1} dp_{S-2} \ldots dp_1
\end{aligned}
$$

We can marginalize out p_{S-1} from the innermost integral:

$$
\int_0^{1-\sum_{j=1}^{S-2} p_j} p_{S-1}^{\alpha_{S-1}^\star - 1} \left(1 - \sum_{j=1}^{S-1} p_j\right)^{\alpha_S^\star - 1} dp_{S-1} \tag{5.4}
$$

Let $z\left(1 - \sum_{j=1}^{S-2} p_j\right) = p_{S-1}$, then

$$
dp_{S-1} = \left(1 - \sum_{j=1}^{S-2} p_j\right) dz \tag{5.5}
$$

and

$$
\begin{aligned}
1 - \sum_{j=1}^{S-1} p_j &= 1 - \sum_{j=1}^{S-2} p_j - p_{S-1} \\
&= 1 - \sum_{j=1}^{S-2} p_j - z\left(1 - \sum_{j=1}^{S-2} p_j\right) \\
&= (1-z) - (1-z) \sum_{j=1}^{S-2} p_j \\
&= (1-z)\left(1 - \sum_{j=1}^{S-2} p_j\right) \qquad (5.6)
\end{aligned}
$$

by (5.5) and (5.6), equation (5.4) becomes

$$\begin{aligned}
&\left(1-\sum_{j=1}^{S-2}p_j\right)^{\alpha^\star_{S-1}-1}\int_0^1 z^{\alpha^\star_{S-1}-1}\left[(1-z)\left(1-\sum_{j=1}^{S-2}p_j\right)\right]^{\alpha^\star_S-1}\left(1-\sum_{j=1}^{S-2}p_j\right)dz\\
&=\left(1-\sum_{j=1}^{S-2}p_j\right)^{\alpha^\star_{S-1}+\alpha^\star_S-1}\int_0^1 z^{\alpha^\star_{S-1}-1}(1-z)^{\alpha^\star_S-1}dz\\
&=\left(1-\sum_{j=1}^{S-2}p_j\right)^{\alpha^\star_{S-1}+\alpha^\star_S-1}B\left(\alpha^\star_{S-1},\alpha^\star_S\right) \qquad (5.7)
\end{aligned}$$

With the result from equation (5.7), the marginal density p_i:

$$\begin{aligned}
f(p_i|\boldsymbol{\alpha}^\star;n_1,n_2,\ldots,n_S)=&\frac{B\left(\alpha^\star_{S-1},\alpha^\star_S\right)}{B(\boldsymbol{\alpha}^\star)}\int_0^{1-p_i}\int_0^{1-p_i-p_1}\cdots\\
&\int_0^{1-\sum_{j=1}^{S-3}p_j}\prod_{j=1}^{S-2}p_j^{\alpha^\star_j-1}\left(1-\sum_{j=1}^{S-2}p_j\right)^{\alpha^\star_{S-1}+\alpha^\star_S-1}dp_{S-2}dp_{S-3}\ldots dp_1
\end{aligned} \qquad (5.8)$$

We can marginalize out p_{S-2} by repeating the same process and get

$$\begin{aligned}
f(p_i|\boldsymbol{\alpha}^\star;n_1,n_2,\ldots,n_S)=&\frac{B\left(\alpha^\star_{S-1},\alpha^\star_S\right)B\left(\alpha^\star_{S-2},\alpha^\star_{S-1}+\alpha^\star_S\right)}{B(\boldsymbol{\alpha}^\star)}\int_0^{1-p_i}\int_0^{1-p_i-p_1}\cdots\\
&\int_0^{1-\sum_{j=1}^{S-4}p_j}\prod_{j=1}^{S-3}p_j^{\alpha^\star_j-1}\left(1-\sum_{j=1}^{S-3}p_j\right)^{\alpha^\star_S+\alpha^\star_{S-1}+\alpha^\star_{S-2}-1}dp_{S-3}dp_{S-4}\ldots dp_1
\end{aligned} \qquad (5.9)$$

Note that $B\left(\alpha^\star_{S-1},\alpha^\star_S\right)B\left(\alpha^\star_{S-2},\alpha^\star_{S-1}+\alpha^\star_S\right)=B\left(\alpha^\star_{S-2},\alpha^\star_{S-1},\alpha^\star_S\right)$. Therefore, by repeating such iteration, we eventually have

$$f(p_i|\boldsymbol{\alpha}^\star;n_1,n_2,\ldots,n_S)=\frac{B\left(\boldsymbol{\alpha}^\star_{-i}\right)}{B\left(\boldsymbol{\alpha}^\star\right)}p_i^{\alpha^\star_i-1}(1-p_i)^{\left(\sum_{j\neq i}^{S}\alpha^\star_j\right)-1} \qquad (5.10)$$

Note that $\frac{B\left(\boldsymbol{\alpha}^{\star}_{-i}\right)}{B(\boldsymbol{\alpha}^{\star})} = \frac{1}{B\left(\alpha_i^{\star}, \sum_{j\neq i}^{S}\alpha_i^{\star}\right)}$. Hence, the marginal posterior density of p_i is a beta distribution, i.e. $p_i \sim Beta(n_i+\alpha, n-n_i+\alpha S-\alpha)$. The property of sampling distribution is obtained.

$$E(p_i) = \frac{n_i+\alpha}{n+\alpha S} \tag{5.11}$$

5.1 Bayesian Estimator of Shannon's Index

The estimator of H: $h_0 = -\sum_{i=1}^{S}\frac{n_i}{n}\ln\left(\frac{n_i}{n}\right)$ takes the MLE $\hat{p}_i$ as the estimator of frequencies. It has been derived by Basharin (1959) using Taylor expansion that

$$E\left(h_0\right) = H - \frac{S-1}{2n} + O\left(n^{-2}\right) \tag{5.12}$$

This suggests h_0 is not an unbiased estimator of H. Adding $\frac{S-1}{2n}$ could remove the bias, however, it is not suitable for small sample size especially when rare species is not observed in sample (Gill and Joanes, 1979). Therefore, Gill and Joanes (1979) propose a Bayesian estimator of H where the relative abundance p_is are estimated by their posterior mean $\frac{n_i+\alpha}{n+\alpha S}$:

$$h_\alpha = -\sum_{i=1}^{S}\frac{n_i+\alpha}{n+\alpha S}\ln\left(\frac{n_i+\alpha}{n+\alpha S}\right) \tag{5.13}$$

Another approach proposed by Gill and Joanes (1979) is by calculating the posterior mean of $p_i\ln(p_i)$

$$E\left(p_i\ln p_i\right) = \int_0^1 p_i\ln p_i f\left(p_i\right)dp_i$$

then

$$\hbar_\alpha = \sum_{i=1}^{S} \frac{n_i + \alpha}{n + \alpha S} \left\{ \psi(n + \alpha S + 1) - \psi\left(n_i + \alpha + 1\right) \right\} \tag{5.14}$$

where $\psi(x)$ is the digamma function.

Therefore, given a set of probability $p_1, p_2, \ldots, p_S$, each n_i is a binomial distribution with probability p_i. The conditional expectation of these two measures can be obtained.

$$\begin{aligned}
&E(h_\alpha | p_1, p_2, \ldots, p_n) \\
&= \sum_{i=1}^{S} E\left[\frac{n_i + \alpha}{n + \alpha S} \ln\left(n + \alpha S\right) - \frac{n_i + \alpha}{n + \alpha S} \ln\left(n_i + \alpha\right) \right] \\
&= \sum_{i=1}^{S} \frac{1}{n + \alpha S} \left\{ \left[E(n_i) + \alpha \right] \ln\left(n + \alpha S\right) - E\left[\left(n_i + \alpha\right) \ln\left(n_i + \alpha\right) \right] \right\} \\
&= \sum_{i=1}^{S} \frac{1}{n + \alpha S} \left\{ \left(n p_i + \alpha\right) \ln\left(n + \alpha S\right) - \sum_{r=0}^{n} \left(r + \alpha\right) \ln\left(r + \alpha\right) \binom{n}{r} p_i^r (1 - p_i)^{n-r} \right\} \\
&= \sum_{i=1}^{S} \frac{n p_i + \alpha}{n + \alpha S} \ln\left(n + \alpha S\right) - \frac{1}{n + \alpha S} \sum_{r=0}^{n} \left(r + \alpha\right) \ln\left(r + \alpha\right) \binom{n}{r} \sum_{i=1}^{S} p_i^r (1 - p_i)^{n-r} \\
&= \ln\left(n + \alpha S\right) - \frac{1}{n + \alpha S} \sum_{r=0}^{n} \left(r + \alpha\right) \ln\left(r + \alpha\right) \binom{n}{r} \sum_{i=1}^{S} p_i^r (1 - p_i)^{n-r}
\end{aligned} \tag{5.15}$$

With the same derivation, we can get the conditional expectation of $\hbar_\alpha$:

$$E\left(\hbar_\alpha | p_1, p_2, \ldots, p_n\right) = \psi(n + \alpha S + 1) - \frac{1}{n + \alpha S} \sum_{r=0}^{n} \binom{n}{r} (\alpha + r) \psi(\alpha + r + 1) \sum_{i=1}^{S} p_i^r \left(1 - p_i\right)^{n-r} \tag{5.16}$$

With Bayesian approach, $p_i \sim Beta(n_i + \alpha, n - n_i + \alpha S - \alpha)$. Therefore, we derive

the expectation of two measures:

$$
\begin{aligned}
&E\left(h_\alpha\right) \\
&= E\left[E(h_\alpha|p_1,p_2,\ldots,p_n)\right] \\
&= \ln\left(n+\alpha S\right) - \frac{1}{n+\alpha S}\sum_{r=0}^{n}\left(r+\alpha\right)\ln\left(r+\alpha\right)\binom{n}{r}\sum_{i=1}^{S}E\left[p_i^r(1-p_i)^{n-r}\right] \\
&= \ln\left(n+\alpha S\right) - \frac{1}{n+\alpha S}\sum_{r=0}^{n}\left(r+\alpha\right)\ln\left(r+\alpha\right)\binom{n}{r}\sum_{i=1}^{S}\int_0^1 p_i^r(1-p_i)^{n-r}\times \\
&\frac{1}{B(r+\alpha,n-r+\alpha S-\alpha)}p_i^{r+\alpha-1}(1-p_i)^{n-r+\alpha S-\alpha-1}dp_i \\
&= \ln\left(n+\alpha S\right) - \frac{1}{n+\alpha S}\sum_{r=0}^{n}\left(r+\alpha\right)\ln\left(r+\alpha\right)\binom{n}{r}\sum_{i=1}^{S}\frac{B(2r+\alpha,2(n-r)+\alpha S-\alpha)}{B(r+\alpha,n-r+\alpha S-\alpha)} \\
&= \ln\left(n+\alpha S\right) - \frac{S}{n+\alpha S}\sum_{r=0}^{n}\left(r+\alpha\right)\ln\left(r+\alpha\right)\binom{n}{r}\frac{B(2r+\alpha,2(n-r)+\alpha S-\alpha)}{B(r+\alpha,n-r+\alpha S-\alpha)}
\end{aligned}
\tag{5.17}
$$

Similarly,

$$
E\left(\hbar_\alpha\right) = \psi(n+\alpha S+1) - \frac{S}{n+\alpha S}\sum_{r=0}^{n}\left(r+\alpha\right)\psi(\alpha+r+1)\binom{n}{r}\frac{B(2r+\alpha,2(n-r)+\alpha S-\alpha)}{B(r+\alpha,n-r+\alpha S-\alpha)}
\tag{5.18}
$$

Gill and Joanes (1979) evaluate the estimating ability of h_0, h_α and $\hbar_\alpha$. It appears that for small sample size, h_α and $\hbar_\alpha$ provide better estimates than h_0. As sample size increases, the standard deviations of h_α and $\hbar_\alpha$ experience a increase at first before decreasing. h_0 has the greatest mean square error when the number of species is large but the sample obtained has small size. They also mention that $\hbar_\alpha$ is slower to converge and in general, produces larger mean square error than h_α.

5.2 Proposed Bayesian Estimator of Simpson's Index

Based on the work established previously, we now propose the Bayesian estimator of Simpson's index:

$$d_\alpha = \sum_{i=1}^{S} \left(\frac{n_i + \alpha}{n + \alpha S} \right)^2 \tag{5.19}$$

Another approach is to calculate the posterior mean of p_i^2.

$$E\left(p_i^2\right) = Var(p_i) + (E(p_i))^2 = \frac{(n_i + \alpha)(n_i + \alpha + 1)}{(n + \alpha S)(n + \alpha S + 1)}$$

Therefore,

$$\widetilde{d}_\alpha = \sum_{i=1}^{S} \frac{(n_i + \alpha)(n_i + \alpha + 1)}{(n + \alpha S)(n + \alpha S + 1)} \tag{5.20}$$

For a given set of probability $p_1, p_2, \ldots, p_S$, the conditional expectation of d_α and $\widetilde{d}_\alpha$ are derived as following:

$$
\begin{aligned}
& E\left(d_\alpha | p_1, p_2, \ldots, p_n\right) \\
&= \sum_{i=1}^{S} E\left[\left(\frac{n_i + \alpha}{n + \alpha S}\right)^2\right] \\
&= \sum_{i=1}^{S} \frac{1}{(n+\alpha S)^2} E\left(n_i^2 + 2n_i\alpha + \alpha^2\right) \\
&= \sum_{i=1}^{S} \frac{1}{(n+\alpha S)^2} \left[np_i^2(n-1) + np_i + 2\alpha np_i + \alpha^2\right] \\
&= \sum_{i=1}^{S} \frac{1}{(n+\alpha S)^2} \left[n(n-1)p_i^2 + (2\alpha+1)np_i + \alpha^2\right] \\
&= \frac{n(n-1)}{(n+\alpha S)^2} \sum_{i=1}^{S} p_i^2 + \frac{n(2\alpha+1) + \alpha^2 S}{(n+\alpha S)^2} \\
&= \frac{n(n-1)}{(n+\alpha S)^2} D + \frac{n(2\alpha+1) + \alpha^2 S}{(n+\alpha S)^2}
\end{aligned} \tag{5.21}
$$

and

$$
\begin{aligned}
& E\left(\widetilde{d}_\alpha | p_1, p_2, \ldots, p_n\right) \\
&= \sum_{i=1}^{S} E\left[\frac{(n_i+\alpha)(n_i+\alpha+1)}{(n+\alpha S)(n+\alpha S+1)}\right] \\
&= \frac{1}{(n+\alpha S)(n+\alpha S+1)} \sum_{i=1}^{S} E\left(n_i^2 + 2n_i\alpha + \alpha^2 + n_i + \alpha\right) \\
&= \frac{1}{(n+\alpha S)(n+\alpha S+1)} \sum_{i=1}^{S} \left\{np_i\left[(n-1)p_i + 1\right] + 2\alpha n p_i + \alpha^2 + np_i + \alpha\right\} \\
&= \frac{1}{(n+\alpha S)(n+\alpha S+1)} \left[\sum_{i=1}^{S} n(n-1)p_i^2 + \sum_{i=1}^{S}(2\alpha+2)np_i + \sum_{i=1}^{S}\alpha(\alpha+1)\right] \\
&= \frac{n(n-1)}{(n+\alpha S)(n+\alpha S+1)} \sum_{i=1}^{S} p_i^2 + \frac{(2n+\alpha S)(\alpha+1)}{(n+\alpha S)(n+\alpha S+1)} \\
&= \frac{n(n-1)}{(n+\alpha S)(n+\alpha S+1)} D + \frac{(2n+\alpha S)(\alpha+1)}{(n+\alpha S)(n+\alpha S+1)} \qquad (5.22)
\end{aligned}
$$

As shown in (5.21), d_α is not an unbiased estimator of D. The bias could be removed by subtracting $\frac{n(2\alpha+1)+\alpha^2 S}{(n+\alpha S)^2}$ then multiplying $\frac{(n+\alpha S)^2}{n(n-1)}$. Therefore, the bias-corrected estimator is

$$
\begin{aligned}
d_\alpha^{BC} &= \left[d_\alpha - \frac{n(2\alpha+1)+\alpha^2 S}{(n+\alpha S)^2}\right] \frac{(n+\alpha S)^2}{n(n-1)} \\
&= \sum_{i=1}^{S} \frac{(n_i+\alpha)^2}{n(n-1)} - \frac{n(2\alpha+1)+\alpha^2 S}{n(n-1)} \qquad (5.23)
\end{aligned}
$$

Similarly, (5.22) suggests $\widetilde{d}_\alpha$ is not an unbiased estimator of D, therefore, the bias-corrected estimator is

$$\begin{aligned}\widetilde{d}_\alpha^{BC} &= \left[\widetilde{d}_\alpha - \frac{(2n+\alpha S)(\alpha+1)}{(n+\alpha S)(n+\alpha S+1)}\right]\frac{(n+\alpha S)(n+\alpha S+1)}{n(n-1)} \\ &= \sum_{i=1}^{S}\frac{(n_i+\alpha)(n_i+\alpha+1)}{n(n-1)} - \frac{(2n+\alpha S)(\alpha+1)}{n(n-1)}\end{aligned} \tag{5.24}$$

5.3 Simulation Study on Proposed Bayesian Estimator of Simpson's Index

We fix the number of species $S = 100$ and generate $(X_1, X_2, \ldots, X_S)$ from a multinomial distribution with corresponding probability $p_1, p_2, \ldots, p_S$ that follow a Dirichlet distribution with parameter $\alpha = 0.5$. The population size is set to 5000. The sample size is set to 60, 100 and 200. For each sample, we calculate d_α, d_α^{BC}, $\widetilde{d}_\alpha$ and $\widetilde{d}_\alpha^{BC}$. We also include Simpson's own estimator $\lambda = \sum_{i=1}^{S}\frac{n_i(n_i-1)}{n(n-1)}$ and the MLE estimator $d_{MLE} = \sum_{i=1}^{S}\hat{p}_i^2$. For each estimator, we report the mean and standard error (SE) of estimates as well as the estimate error (bias and root of mean square error, RMSE). The result can be found in Appendix B, Table B.7.

As shown in Table B.7, $\widetilde{d}_\alpha$ and λ overall produce good estimate result. The bias, SE and RMSE produced by these two estimators are small through all three cases while $\widetilde{d}_\alpha$ always has smaller variance and mean square error than that of λ. For small sample size (n = 60), d_α and $\widetilde{d}_\alpha^{BC}$ show a very poor estimating ability as they produce large bias and RMSE. As sample size increases, all estimators converge to true value as we can see there is a drop in bias and RMSE when n increases. d_α seems constantly underestimate D, the bias is reduced after bias correction, however, still

underestimates D. We can see that d_α^{BC} reduces not only the bias, but also RMSE. The bias-correction method is meaningless for $\widetilde{d}_\alpha$ when sample size is small, the bias is reduced only in the case where sample size is 200. It is worth noticing that the variance increases after applying bias correction on d_α and $\widetilde{d}_\alpha$. d_{MLE} overestimates D and produces the largest bias and RMSE through all three cases.

For large sample size, all estimators present fairly good estimate results except for d_{MLE}. The inadequacy of d_α and $\widetilde{d}_\alpha^{BC}$ appears when the sample size is small. We can see that $\widetilde{d}_\alpha$ and λ constantly present good performance on estimating D. Our proposed estimator $\widetilde{d}_\alpha$ is competitive with the estimator λ proposed by Simpson and slightly outperforms λ because $\widetilde{d}_\alpha$ produce smaller SE and RMSE.

5.4 Estimating Chinese Surname Diversity

We are going to estimate the Chinese surname diversity using estimators we proposed previously. Again, we set the sample size to 60, 100 and 200, the same estimators in simulation study are calculated. Simulation study specify the value of parameter α of Dirichlet distribution. However, in this case study, we are unable to obtain the value of α. The choice of α for estimating has been discussed by Gill and Joanes (1979). They point that large value of α would not be suitable as the estimated probabilities would be very close to $\frac{1}{S}$. Large α smoother the actual distribution. Therefore, they propose that small values of α are more appropriate. Smaller value of α is applied when we expect the relative abundances are widely different while greater value of α is suggested when the relative abundances are fairly equal.

As shown in Figure 2.1, the relative abundance curve suggest a large difference among Chinese surnames. A value of α between 0 to 1 would be suitable for our case.

Therefore, we set α to 0.2, 0.5 and 1. The results are presented in Appendix B, Table B.8 to Table B.10.

When $\alpha = 0.2$, our proposed estimators overall produce good results. However, d_α and $\widetilde{d}_\alpha$ overestimate D. After bias correction, d_α^{BC} and $\widetilde{d}_\alpha^{BC}$ largely reduce bias. The variance and mean square error produced by d_α^{BC} is smaller than that of λ. Moreover, as sample size increases, the bias produced by d_α^{BC} is smaller than that of λ.

When $\alpha = 0.5$, $\widetilde{d}_\alpha$ shows a very competitive estimating ability with λ. They produce similar estimate mean while $\widetilde{d}_\alpha$ has smaller variance and mean square error. $\widetilde{d}_\alpha^{BC}$ does not provide good estimate when sample size is small. d_α^{BC} on the other hand, provides stable estimating results. It produce close estimates as λ, its variance and mean square error are smaller than that of λ.

It is not suitable to set $\alpha = 1$ for our case study. Although when sample size is 200, the bias-corrected estimators d_α^{BC} and $\widetilde{d}_\alpha^{BC}$ provide very close estimates to the true value, the proposed estimators overall produce large variance and mean square error. Moreover, when sample is 60, the negative estimate mean of $\widetilde{d}_\alpha^{BC}$ does not have any statistical meaning.

It is worth noticing that d_{MLE} overestimate D and produces large mean square error. Our analysis is consistent with Simpson's conclusion that it is not suitable to use MLE estimator for D.

5.5 Conclusion and Future Work

We derive the distribution of relative abundance p_i from Bayesian approach. Under Bayesian assumption, the relative abundance $(p_1, p_2, \ldots, p_S)$ follows Dirichlet distribution with parameter α. The marginal posterior density of p_i follows Beta distribution. Therefore, Gill and Joane (1979) proposed two estimator of H, h_α and $\hbar_\alpha$ based on the sampling distribution of p_i. The behaviour of two estimators are well discussed in Gill and Joanes's work (1979)[8]. h_α in general has better performance than $\hbar_\alpha$ as it converges faster and produces less mean square error. Overall, both estimators possess better estimating ability than the MLE estimator h_0. Based on their work, we propose two Bayesian estimators d_α and $\widetilde{d}_\alpha$ as well as their bias-corrected estimators d_α^{BC} and $\widetilde{d}_\alpha^{BC}$. We compare our proposed estimators with Simpson's estimator λ and the MLE estimator d_{MLE}. Our simulation study shows that all estimators converge to true value of D while $\widetilde{d}_\alpha$ and λ present the best performance among all measures. d_α and $\widetilde{d}_\alpha^{BC}$ are not suitable for estimating when the sample size is small. We suggest not using d_{MLE} as estimator because it overestimates D and produces large RMSE. $\widetilde{d}_\alpha$ provides as good estimate results as λ with smaller standard error and RMSE. Our proposed estimators also show a good performance on estimating Chinese surname diversity. With appropriate choice of α, the proposed estimators are competitive with λ.

The work established previously is based on the assumption that S is known. However, in many circumstances, the number of species is unknown. Therefore, S needs to be estimated before applying diversity estimators. The method of estimating S is discussed in many literatures such as the work done by Chao et al. (2000).

The future work could be established based on the Bayesian non-parametric approach. When S is unknown, Dirichlet process can be applied where the number of parameters could be set to infinite.

Appendix A

Figures

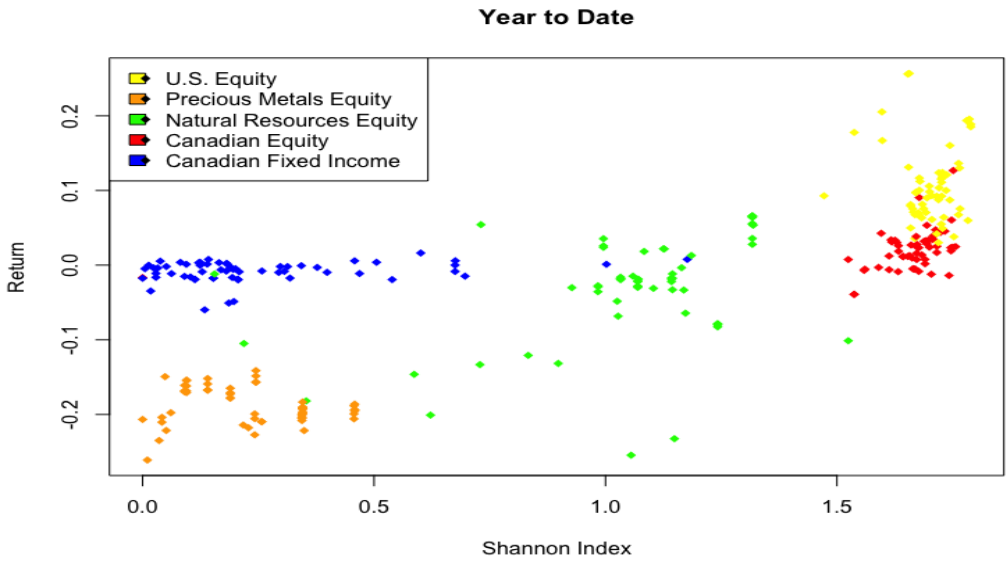

Figure A.1: Year to Date Return (**H**)

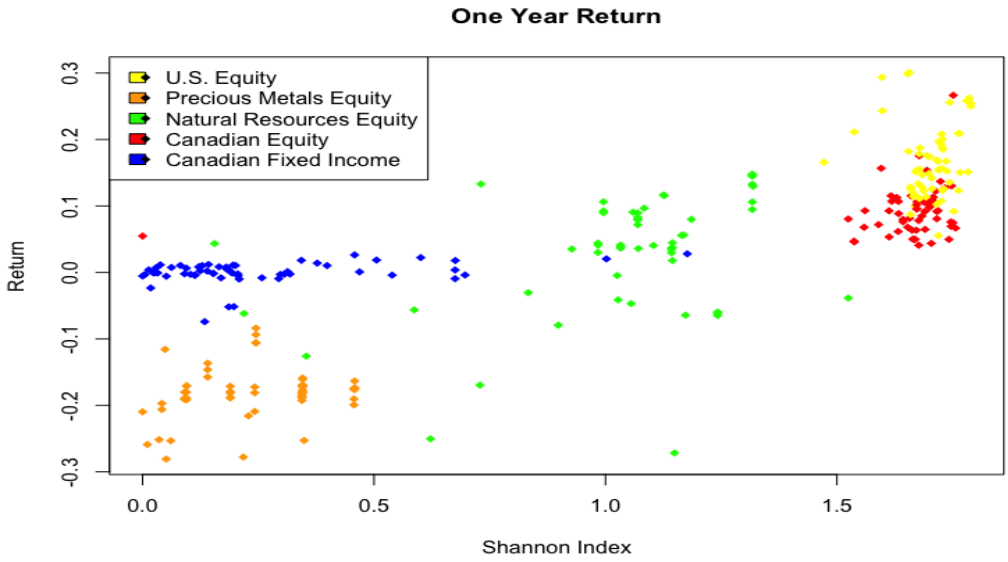

Figure A.2: One Year Annual Return (**H**)

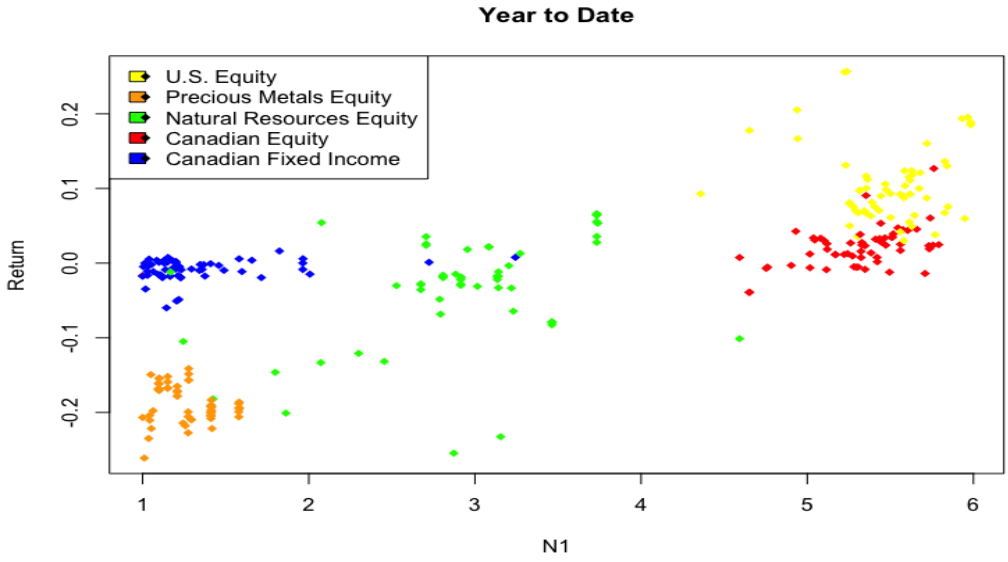

Figure A.3: Year to Date Return (N_1)

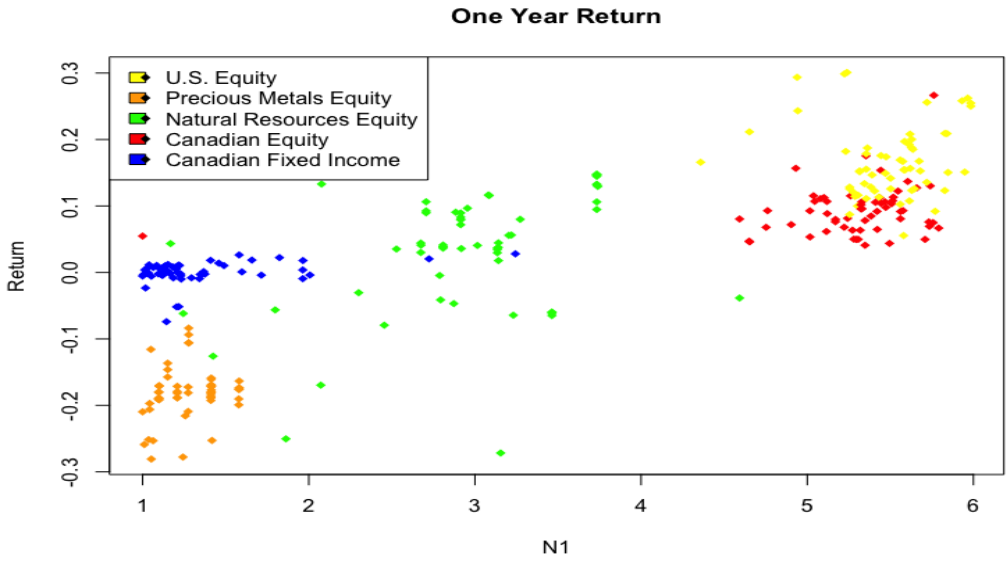

Figure A.4: One Year Annual Return (N_1)

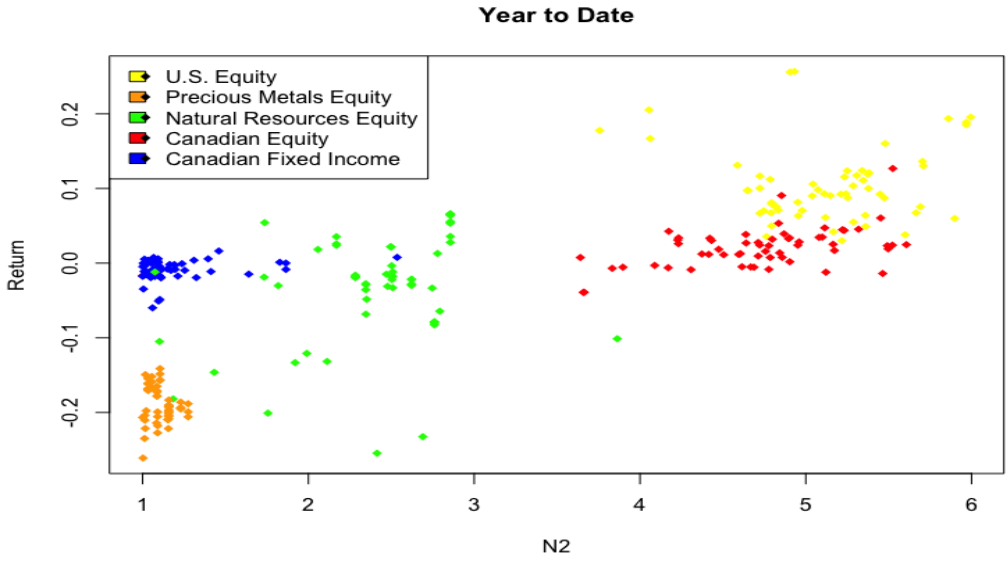

Figure A.5: Year to Date Return (N_2)

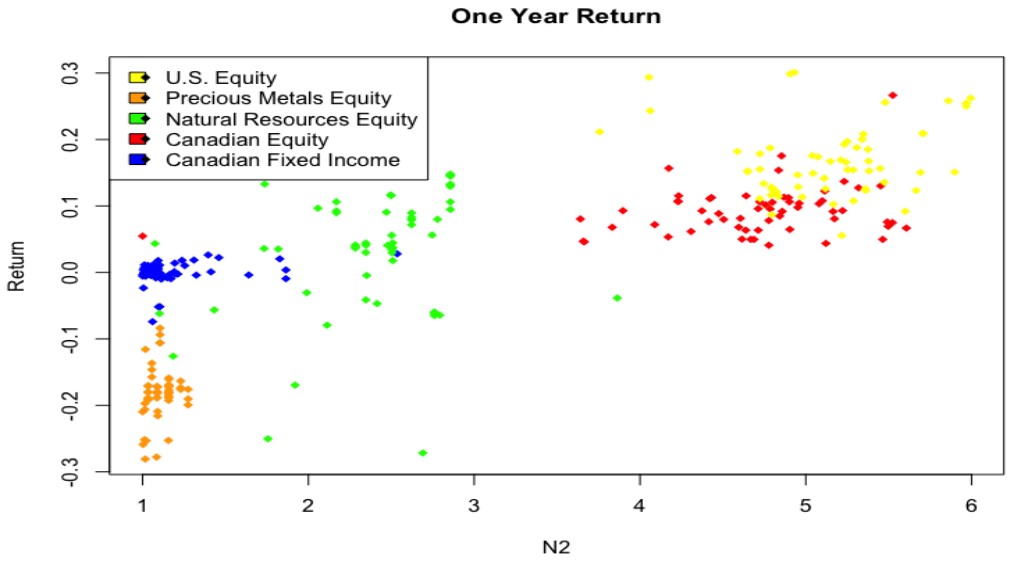

Figure A.6: One Year Annual Return (N_2)

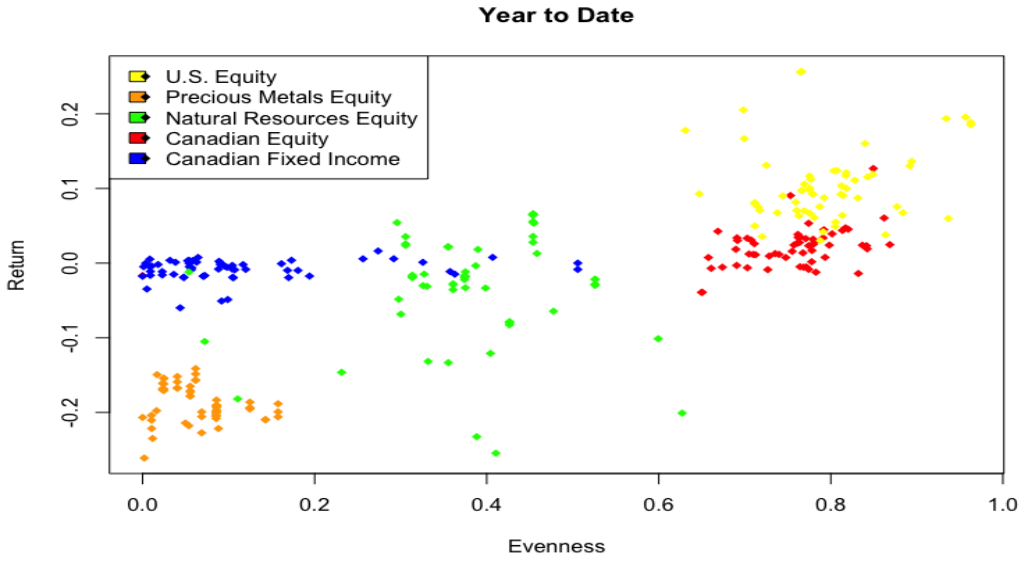

Figure A.7: Year to Date Return (**E**)

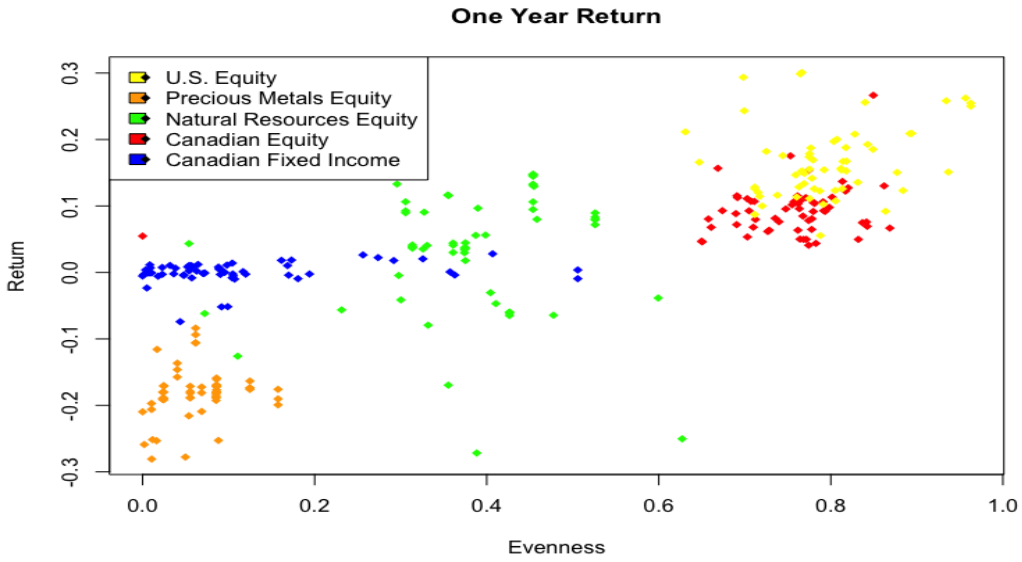

Figure A.8: One Year Annual Return (**E**)

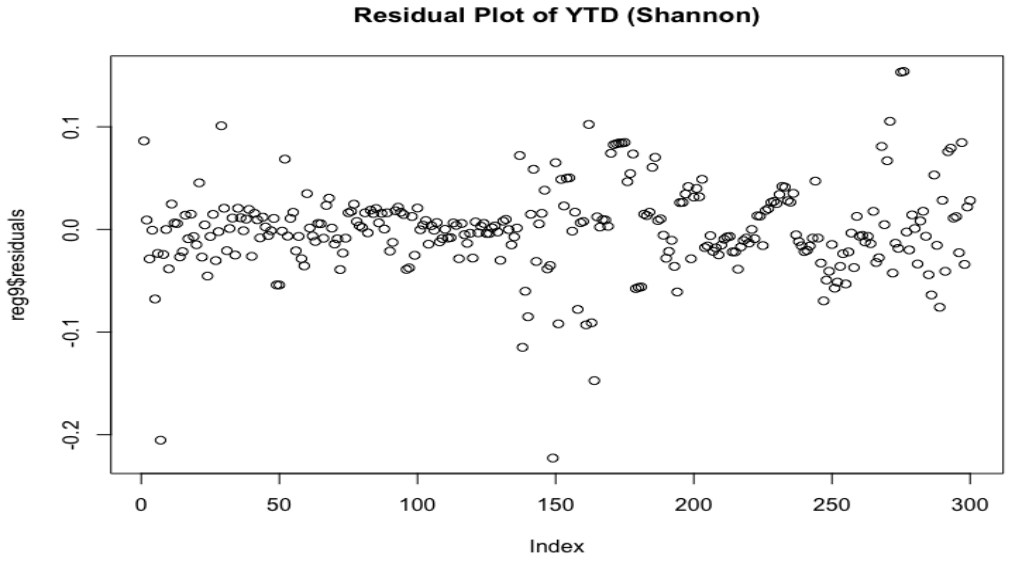

Figure A.9: Residual Plot of YTD

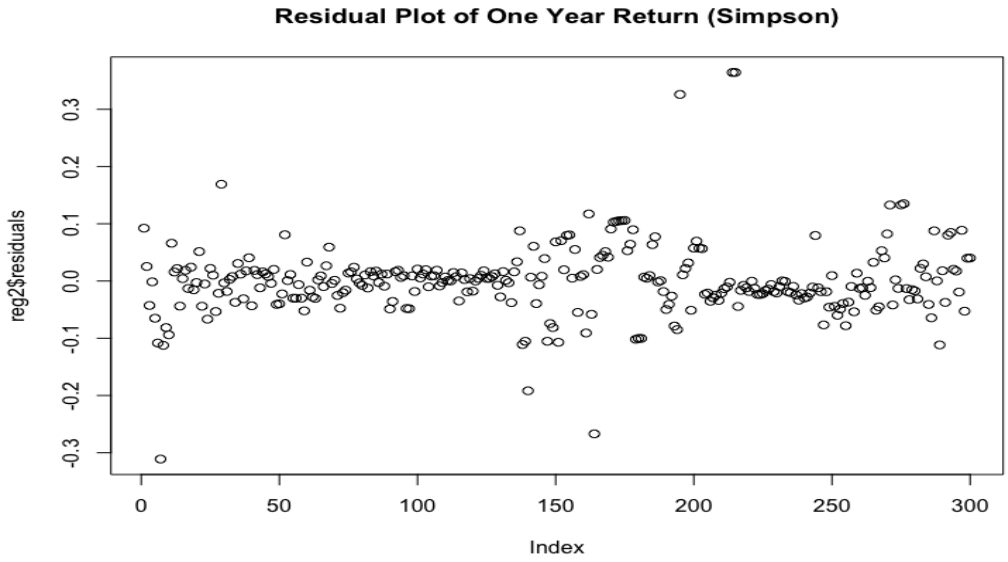

Figure A.10: Residual Plot of One Year Return

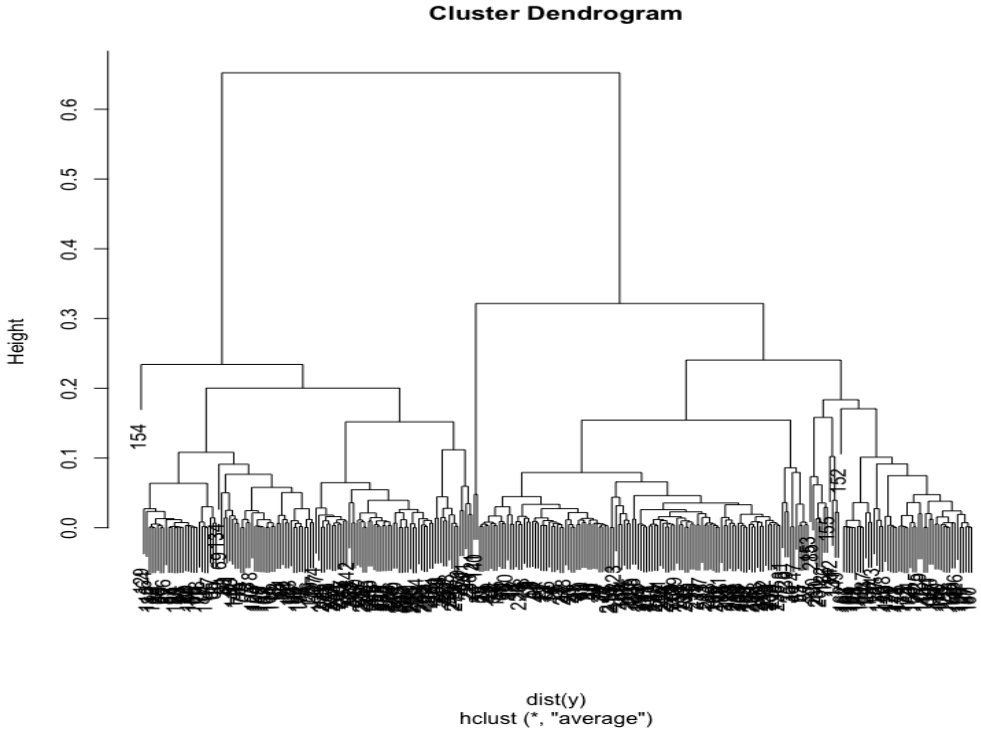

Figure A.11: K-medoids Average Silhouette Plot

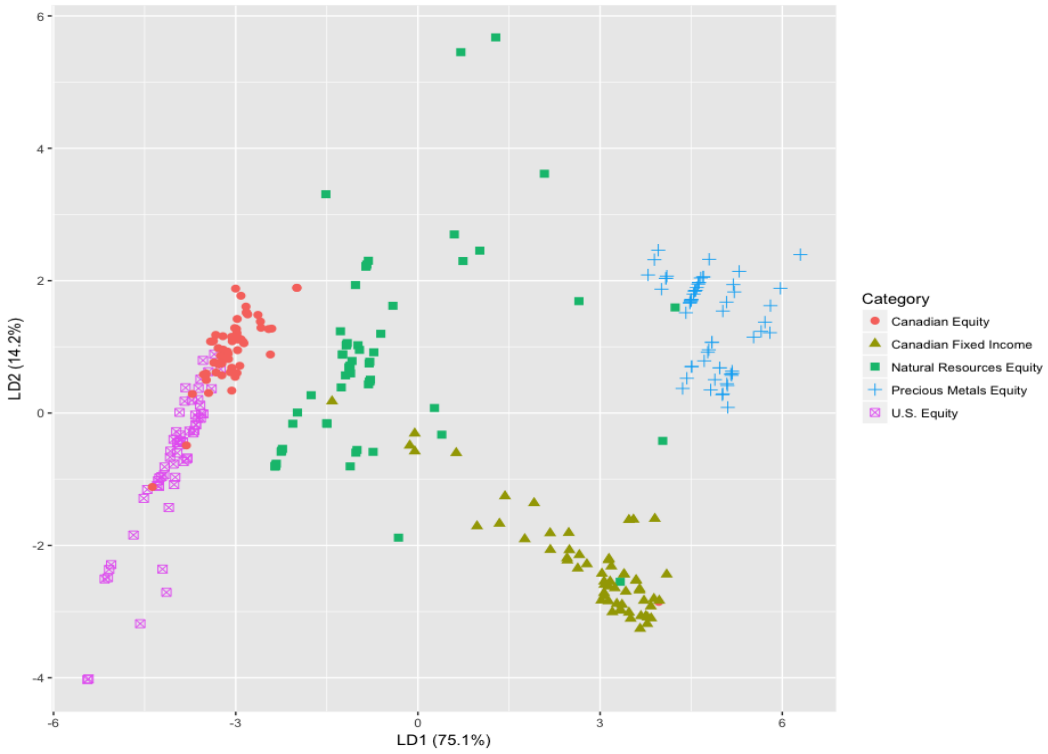

Figure A.12: LDA plot

Appendix B

Tables

Surname	Count 10k	Surname	Count 10k	Surname	Count 10k
王	9520	高	1330	董	677
李	9340	罗	1260	袁	667
张	8960	郑	1240	于	642
刘	6770	梁	1130	余	633
陈	6130	谢	1010	叶	632
杨	4270	宋	932	蒋	632
黄	3260	唐	917	杜	619
吴	2680	许	881	苏	606
赵	2670	邓	821	魏	603
周	2520	冯	818	程	601
徐	1930	韩	815	吕	596
孙	1830	曹	791	丁	576
马	1720	曾	772	沈	550
朱	1700	彭	766	任	547
胡	1550	萧	739	姚	538
林	1510	蔡	701	卢	536
郭	1500	潘	687	傅	536
何	1400	田	685	钟	523

Table B.1: Chinese Surnames

Surname	Count 10k	Surname	Count 10k	Surname	Count 10k
姜	523	龙	281	齐	176
崔	509	陶	274	易	175
谭	499	贺	274	乔	173
廖	487	顾	272	伍	171
范	485	毛	264	庞	167
汪	483	郝	264	颜	164
陆	480	龚	264	倪	163
金	467	邵	262	庄	162
石	455	万	254	聂	159
戴	449	钱	249	章	157
贾	439	严	246	鲁	151
韦	430	赖	240	岳	149
夏	426	覃	240	翟	149
邱	423	洪	240	殷	147
方	413	武	239	詹	147
侯	401	莫	233	申	147
邹	394	孔	231	欧	146
熊	384	汤	227	耿	140
孟	383	向	226	关	137
秦	379	常	218	兰	134
白	374	温	217	焦	133
江	369	康	211	俞	132
阎	360	施	206	左	131
薛	347	文	204	柳	129
尹	346	牛	202	甘	126
段	320	樊	200	祝	120
雷	319	葛	195	包	115
黎	288	邢	192	宁	110
史	285	安	179	尚	109

Table B.2: Chinese Surnames

Surname	Count 10k
符	109
舒	109
阮	109
柯	106
纪	106
梅	105
童	105
凌	103
毕	103
单	101
季	101
裴	100
霍	100
涂	100
成	100
苗	100
谷	99
盛	98
曲	98
翁	97
冉	97
骆	96
蓝	96
路	95
游	94
辛	92
靳	92
欧阳	91
管	87

Surname	Count 10k
柴	86
蒙	83
鲍	82
华	82
喻	81
祁	80
蒲	75
房	75
滕	74
屈	73
饶	73
解	71
牟	70
艾	69
尤	68
阳	67
时	67
穆	64
农	62
司	59
卓	58
古	58
吉	58
缪	57
简	57
车	57
项	57
连	57
芦	57

Surname	Count 10k
麦	55
褚	54
娄	53
窦	53
戚	53
岑	52
景	52
党	52
宫	52
费	51
卜	51
冷	50
晏	50
席	48
卫	48
米	46
柏	46
宗	45
瞿	44
桂	44
全	44
佟	43
应	43
臧	43
闵	43
苟	43
邬	42
边	42
卞	42

Table B.3: Chinese Surnames

Surname	Count 10k
姬	42
师	41
和	41
仇	40
栾	40
隋	40
商	39
刁	39
沙	39
荣	38
巫	38
寇	38
桑	37
郎	37
甄	36
丛	36
仲	35
虞	35
敖	35
巩	34
明	34
佘	34
池	34
查	33
麻	33
苑	33
迟	32
邝	32
官	31

Surname	Count 10k
封	31
谈	31
匡	30
鞠	30
惠	29.8
荆	28.9
乐	28.8
冀	28.5
郁	28.5
胥	28.5
南	27.7
班	27.3
储	27.2
原	27
栗	26.6
燕	26.4
楚	26.3
鄢	26.3
劳	25.9
谌	24.8
奚	23.1
皮	22.9
粟	22.8
冼	22.8
蔺	22.8
楼	22.8
盘	22.5
满	21.9
闻	21.7

Surname	Count 10k
位	20.8
厉	20.6
伊	20.6
仝	20
区	19.9
郜	19.8
海	19.7
阚	19.6
花	19.5
权	19.1
强	19
帅	19
屠	18.9
豆	18.8
朴	18.7
盖	18.6
练	18.5
廉	18.4
禹	18.2
井	17.9
祖	17.7
漆	17.7
巴	17.7
丰	17.6
支	17.3
卿	17.2
国	17.1
狄	16.8
平	16.6

Table B.4: Chinese Surnames

Surname	Count 10k
计	16.5
索	16.5
宣	16.4
晋	16.2
相	16.2
初	15.9
门	15.9
云	15.6
容	15.4
敬	15
来	14.8
扈	14.7
晁	14.6
芮	14.6
都	14.6
普	14.5
阙	14.5
浦	14.4
戈	14.4
伏	14.3
鹿	14
薄	14
邸	13.9
雍	13.9
辜	13.8
养	13.6
阿	13.6
乌	13.5
母	13.5

Surname	Count 10k
裘	13.4
亓	13.4
修	13.3
郃	13
赫	12.8
杭	12.8
况	12.4
那	12.4
宿	12.3
鲜	12.2
印	12.1
逯	12.1
隆	12
茹	11.9
诸	11.8
战	11.7
慕	11.5
危	11.2
玉	11.2
银	11.1
亢	11
嵇	10.9
公	10.9
哈	10.7
湛	10.5
宾	10.2
戎	10.1
勾	10.1
茅	10.1

Surname	Count 10k
利	10
於	9.9
呼	9.8
居	9.6
揭	9.6
干	9.5
但	9.5
尉	9.4
冶	9.3
斯	9.2
元	9.1
束	9
檀	9
衣	9
信	8.9
展	8.9
阴	8.9
咎	8.7
智	8.7
幸	8.6
奉	8.5
植	8.5
衡	8.4
富	8.4
尧	8
闭	8
由	8

Table B.5: Chinese Surnames

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Canadian Equity	59	1	0	0	0
Canadian Fixed Income	0	55	0	5	0
Natural Resources Equity	0	3	2	54	1
Precious Metals Equity	0	60	0	0	0
U.S. Equity	60	0	0	0	0

Table B.6: Cluster Table Hierarchical Clustering

Sample Size	Index	Mean	Bias	SE	RMSE
60	d_α	0.01798876	-0.0069662797	0.001643158	0.007157257
	$\widetilde{d}_\alpha$	0.02414152	-0.0008135167	0.001541163	0.001742016
	d_α^{BC}	0.02052655	-0.0044284863	0.005616444	0.007150133
	$\widetilde{d}_\alpha^{BC}$	0.01123390	-0.0137211417	0.005315708	0.014713879
	λ	0.02517288	0.0002178414	0.005771158	0.005772384
	d_{MLE}	0.04142000	0.0164649600	0.005674972	0.017414592
100	d_α	0.02040338	-0.0045516620	0.001708147	0.004861325
	$\widetilde{d}_\alpha$	0.02566393	0.0007088940	0.001651763	0.001796696
	d_α^{BC}	0.02364404	-0.0013110000	0.003882153	0.004095700
	$\widetilde{d}_\alpha^{BC}$	0.02083717	-0.0041178680	0.003779032	0.005587812
	λ	0.02504747	0.0000924348	0.003935010	0.003934128
	d_{MLE}	0.03479700	0.0098419600	0.003895660	0.010584194
200	d_α	0.02245918	-0.0024958560	0.001557916	0.002941764
	$\widetilde{d}_\alpha$	0.02595039	0.0009953544	0.001497514	0.001797509
	d_α^{BC}	0.02459043	-0.0003646129	0.002446477	0.002472287
	$\widetilde{d}_\alpha^{BC}$	0.02406786	-0.0008871757	0.002426169	0.002582148
	λ	0.02485171	-0.0001033315	0.002456800	0.002457745
	d_{MLE}	0.02972745	0.0047724100	0.002444516	0.005361491
True Value of D: 0.02495504					

Table B.7: Simpson's Measure of Diversity, 5000 Simulation trials, S = 100

Sample Size	Index	Mean	Bias	SE	RMSE
60	d_α	0.02828512	0.0016168336	0.004607358	0.004880642
	$\widetilde{d}_\alpha$	0.03836089	0.0116926038	0.004466372	0.012515810
	d_α^{BC}	0.02627818	-0.0003901043	0.008329686	0.008334655
	$\widetilde{d}_\alpha^{BC}$	0.02276231	-0.0039059800	0.008175731	0.009057175
	λ	0.026981356	0.0003130708	0.008360646	0.008362327
	d_{MLE}	0.04319833	0.0165300482	0.008221301	0.018459813
100	d_α	0.02832784	0.0016595510	0.004145733	0.004463634
	$\widetilde{d}_\alpha$	0.03572555	0.0090572690	0.004070053	0.009928891
	d_α^{BC}	0.02665867	-0.0000096144	0.006030157	0.006027149
	$\widetilde{d}_\alpha^{BC}$	0.02573081	-0.0009374730	0.005969410	0.006039626
	λ	0.02684424	0.0001759573	0.006042373	0.006041914
	d_{MLE}	0.03657580	0.0099075149	0.005981949	0.011571810
200	d_α	0.02791964	0.0012513570	0.003388078	0.003610193
	$\widetilde{d}_\alpha$	0.03225348	0.0055851920	0.003359644	0.006516924
	d_α^{BC}	0.02681685	0.0001485662	0.004120176	0.004120794
	$\widetilde{d}_\alpha^{BC}$	0.02673779	0.0000695059	0.004104168	0.004102704
	λ	0.02683266	0.0001643782	0.004123389	0.004124604
	d_{MLE}	0.03169850	0.0050302150	0.004102772	0.006489913
True Value of D:	0.02666829				

Table B.8: Simpson's Measure of Chinese Surnames, $\alpha = 0.2$

Sample Size	Index	Mean	Bias	SE	RMSE
60	d_α	0.01859143	-0.0080768595	0.002389667	0.008422616
	$\widetilde{d}_\alpha$	0.02488458	-0.0017837028	0.002258513	0.002877043
	d_α^{BC}	0.02258651	-0.0040817738	0.008168071	0.009127516
	$\widetilde{d}_\alpha^{BC}$	0.01379682	-0.0128714631	0.007789957	0.015043182
	λ	0.02698136	0.0003130708	0.008360646	0.008362327
	d_{MLE}	0.04319833	0.0165300482	0.008221301	0.018459813
100	d_α	0.02130114	-0.0053671407	0.002625211	0.005974198
	$\widetilde{d}_\alpha$	0.02676871	0.0001004235	0.002542581	0.002543293
	d_α^{BC}	0.02568442	-0.0009838659	0.005966388	0.006044020
	$\widetilde{d}_\alpha^{BC}$	0.02336477	-0.0033035124	0.005817117	0.006687167
	λ	0.02684424	0.0001759573	0.006042373	0.006041914
	d_{MLE}	0.03657580	0.0099075149	0.005981949	0.011571810
200	d_α	0.02382411	-0.0028441770	0.002613026	0.003861401
	$\widetilde{d}_\alpha$	0.02758789	0.0009196073	0.002577515	0.002735437
	d_α^{BC}	0.02673384	0.0000655528	0.004103370	0.004101841
	$\widetilde{d}_\alpha^{BC}$	0.02653619	-0.0001320979	0.004063795	0.004063911
	λ	0.02683266	0.0001643782	0.004123389	0.004124604
	d_{MLE}	0.03169850	0.0050302150	0.004102772	0.006489913
True Value of D:	0.02666829				

Table B.9: Simpson's Measure of Chinese Surnames, $\alpha = 0.5$

Sample Size	Index	Mean	Bias	SE	RMSE
60	d_α	0.012237617	-0.0144306679	0.001051597	0.014468895
	$\widetilde{d}_\alpha$	0.015956988	-0.0107112975	0.000947473	0.010753079
	d_α^{BC}	0.009401977	-0.0172663077	0.007604768	0.018865313
	$\widetilde{d}_\alpha^{BC}$	-0.008177401	-0.0348456862	0.006894606	0.035520556
	λ	0.026981356	0.0003130708	0.008360646	0.008362327
	d_{MLE}	0.04319833	0.0165300482	0.008221301	0.018459813
100	d_α	0.01549573	-0.0111725601	0.001421616	0.011262552
	$\widetilde{d}_\alpha$	0.01925124	-0.0074170413	0.001344969	0.007537880
	d_α^{BC}	0.02220495	-0.0044633356	0.005743901	0.007271917
	$\widetilde{d}_\alpha^{BC}$	0.01756566	-0.0091026285	0.005461389	0.010613896
	λ	0.02684424	0.0001759573	0.006042373	0.006041914
	d_{MLE}	0.03657580	0.0099075149	0.005981949	0.011571810
200	d_α	0.01946897	-0.0071993184	0.001788455	0.007417921
	$\widetilde{d}_\alpha$	0.02255231	-0.0041159706	0.001748766	0.004471727
	d_α^{BC}	0.02643736	-0.0002309233	0.004044245	0.004048813
	$\widetilde{d}_\alpha^{BC}$	0.02604206	-0.0006262248	0.003967678	0.004014833
	λ	0.02683266	0.0001643782	0.004123389	0.004124604
	d_{MLE}	0.03169850	0.0050302150	0.004102772	0.006489913
True Value of D: 0.02666829					

Table B.10: Simpson's Measure of Chinese Surnames, $\alpha = 1$

Appendix C

R Output

Output for LDA

```
Call:
lda(Category ~ ., data = my_dat, prior = c(1, 1, 1, 1, 1)/5,
    subset = train)

Prior probabilities of groups:
         Canadian Equity    Canadian Fixed Income Natural Resources Equity
                     0.2                      0.2                      0.2
  Precious Metals Equity              U.S. Equity
                     0.2                      0.2

Group means:
                            Simpson          YTD
Canadian Equity          0.76085996  0.021547222
Canadian Fixed Income    0.13511406 -0.008158621
```

```
Natural Resources Equity 0.55126363 -0.046648485
Precious Metals Equity   0.08618181 -0.190464000
U.S. Equity              0.79768729  0.110485185

Coefficients of linear discriminants:
              LD1        LD2
Simpson -6.952245   5.706486
YTD     -9.839496 -19.325782

Proportion of trace:
   LD1    LD2
0.8429 0.1571
```

Output for Classification Tree

```
Classification tree:
rpart(formula = testdat$Category ~ ., data = testdat, subset = train,
    method = "class")

Variables actually used in tree construction:
[1] Simpson YTD

Root node error: 160/200 = 0.8

n= 200
```

```
       CP nsplit rel error  xerror     xstd
1 0.25000      0   1.00000 1.13750 0.025295
2 0.23750      1   0.75000 0.90625 0.039467
3 0.21875      3   0.27500 0.63125 0.044192
4 0.01000      4   0.05625 0.11875 0.025917
```

Bibliography

[1] Basharin, G.P. (1959). *On a statistical estimate for the entropy of a sequence of independent random variable.* Theory of Probability and Its Applications. **4**, 333-336.

[2] Berger, W. H., Parker, F. L. (1970). *Diversity of Planktonic Foraminifera in Deep-Sea Sediments* Science. **168** (3937): 1345–1347.

[3] Bulla, J. (1994). *An Index of Evenness and Its Associated Diversity Measure* Oikos, **70**, 167-171.

[4] Chao, A., Hwang, W.-H., Chen, Y.-C., and Kuo, C.-Y. (2000). *Estimating the number of shared species in two communities.* Statistica Sinica. **10**, 227-246.

[5] Comenetz, J. (2010). *Frequently Occurring Surnames in The 2010 Census* U.S. Census Bureau, 2010 Census.

[6] Demesetz, R. S., Strahan, P. E. (1997). *Diversification, Size and Risk at Bank Holding Companies.* Journal of Money, Credit and Banking. **29**, 300-313.

[7] Faulkenberry, K. *Portfolio Diversification Definition and Purpose* Retrieved from `http://www.arborinvestmentplanner.com/portfolio-diversification-definition-and-purpose/`

[8] Gill, C.A., Joanes, D.N. (1979). *Bayesian estimation of Shannon's index of diversity.* Biometrika. **66**, 1, 81-85.

[9] Goetz, M., Laeven, L. and Levine, R. (2014). *Does the Geographic Expansion of Bank Assets Reduce Risk?* National Bureau of Economic Research, Cambridge.

[10] Heino, J., Mykrä, H., Kotanen, J. (2008). *Weak relationships between landscape characteristics and multiple facets of stream macroinvertebrate biodiversity in a boreal drainage basin.* Landscape Ecol. **23**, 417-426.

[11] Hill, M. O. (1973). *Diversity and Evenness: A Unifying Notation and Its Consequences.* Ecology. **54**, 427–432.

[12] Jost, L. (2006). *Entropy and diversity.* Oikos. **113**, 363–375.

[13] Molinari J. (1989). *A calibrated index for the measurement of evenness.* Oikos, **56**, 319-326.

[14] Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T.S., Meiners, T., Mülla, C., Obermaier, E., Prati, D., Socher, S.A., Sonnemann, I., Wäschke, N., Wubet, T., Wurst, S. and Rillig, M. C. (2014). *Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories* Ecology and Evolution, **4** (18), 3514–3524

[15] Pielou, E. C. (1969). *An introduction to mathematical ecology.* Wiley, 233.

[16] Purvis, A., Hector, A. (2000). *Getting the measure of biodiversity.* Nature. **405**, 212-219.

[17] Shannon, C. E. (1948). *A mathematical theory of communication.* The Bell System Technical Journal, 27, 379–423 and 623–656.

[18] Simpson, E. H. (1949). *Measurement of diversity.* Nature. **163**, 688.

[19] Stirling, G., B. Wilsey. (2001). *Empirical relationships between species richness, evenness, and proportional diversity.* Am. Nat. **158**, 286-299.

[20] Velmurugan, T., Santhanam, T. *Performance Analysis Of K-Means And K-Medoids Clustering Algorithms For A Randomly Generated Data Set* International Conference on Systemics, Cybernetics and Informatics. 578-583.

[21] Whittaker, R. H. (1972). *Evolution and measurement of species diversity.* Taxon. **21**, 213-251.