Spring 2010

# Rao's Quadratic Entropy and Some New Applications

Yueqin Zhao
*Old Dominion University*

# RAO'S QUADRATIC ENTROPY AND SOME NEW APPLICATIONS

by

Yueqin Zhao
B.S. July 2000, Shanghai University of Finance and Economics
M.S. May 2004, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirement for the Degree of

DOCTOR OF PHILOSOPHY

MATHEMATICS AND STATISTICS

OLD DOMINION UNIVERSITY
May 2010

Approved by:

Dayanand N. Naik (Director)

N. Rao Chaganty

Larry D. Lee

Michael J. Doviak

David O. Matson

# ABSTRACT

## RAO'S QUADRATIC ENTROPY AND SOME NEW APPLICATIONS

Yueqin Zhao
Old Dominion University, 2010
Director: Dr. Dayanand N. Naik

Many problems in statistical inference are formulated as testing the diversity of populations. The entropy functions measure the similarity of a distribution function to the uniform distribution and hence can be used as a measure of diversity. Rao (1982a) proposed the concept of quadratic entropy. Its concavity property makes the decomposition similar to ANOVA for categorical data feasible. In this thesis, after reviewing the properties and providing a modification to quadratic entropy, various applications of quadratic entropy are explored. First, analysis of quadratic entropy with the suggested modification to analyze the contingency table data is explored. Then its application to ecological biodiversity is established by constructing practically equivalent confidence intervals. The methods are applied on a real dinosaur diversity data set and simulation experiments are performed to study the validity of the intervals. Quadratic entropy is also used for clustering multinomial data. Another application of quadratic entropy that is provided here is to test the association of two categorical variables with multiple responses. Finally, the gene expression data inspires another application of quadratic entropy in analyzing large scale data, where a hill-climbing type iterative algorithm is developed based on a new minimum quadratic entropy criterion. The algorithm is illustrated on both simulated and real data.

# ACKNOWLEDGMENTS

I would never have been able to finish my dissertation without the guidance of my committee members, help from friends, and support from my family.

I would like to express my deepest gratitude to my advisor, Dr. Dayanand Naik, for his continual guidance and support throughout the entire process of my dissertation. His comments and suggestions lead to many important results in this research endeavor. Dr. Naik was my major source of inspiration during my graduate studies. Had it not been for his intellect and patience, I would not have been able to accomplish this, for which I am very grateful.

I am sincerely grateful to the members of my dissertation committee, Dr. Rao Chaganty, Dr. Larry Lee, Dr. Michael J. Doviak and Dr. David Matson, who have generously given their time and expertise to better my work. I thank them for their contribution and their good-natured support. I am also grateful to all the professors, staff members and my fellow students in the Department of Mathematics and Statistics for providing me with a pleasant and stimulating environment.

Moreover, I am heartily thankful to my supervisor, my colleagues, students and friends at Graduate Program in Public Health in Eastern Virginia Medical School. I especially want to thank Dr. Gavin Welch and Ms. Kay Cherry for their persistent support and encouragement over the years.

Last but not the least, I would like to express my gratefulness to my family. I am grateful to my parents for believing in me and my mother-in-law for taking care of my daughter. I also want to thank my husband for his unwavering support and remarkable patience through it all.

I dedicate this thesis to my parents.

# TABLE OF CONTENTS

APPENDICES

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER I

# INTRODUCTION

Many problems in statistical inference are formulated as testing the diversity of populations. When the variables involved are continuous then variance is generally used as a measure of diversity. However for categorical variables, there is no single measure of diversity. Entropy functions are generally used for this purpose. Entropy is a non-negative function defined on the space of distribution functions and attains the maximum when the distribution is uniform and attains minimum when the distribution is degenerate. The entropy measures the similarity of a distribution with the uniform distribution and hence it is used as a measure of diversity.

This chapter begins by introducing traditional diversity functions in Section I.1. In Section I.2 Rao's quadratic entropy will be introduced along with various examples and decomposition. In Section I.3, an overview of the thesis is presented.

## I.1 ENTROPY FUNCTIONS

There are several entropy functions defined in the literature. We will provide a list here. Let $\Pi = (\pi_1, \pi_2, ..., \pi_s)$ be a vector of relative frequencies in $s$ categories in a population, then the following are entropy functions:

- $H_S(\Pi) = -\sum \pi_i log \pi_i$, (Shannon entropy)
- $H_\alpha(\Pi) = \frac{1-\sum \pi_i^\alpha}{2^{\alpha-1}-1}, \alpha > 0, \alpha \neq 1$, ($\alpha$-order entropy of Havrda and Charvat)
- $H_R(\Pi) = \frac{log(\sum \pi_i{}^\alpha)}{1-\alpha}, \alpha > 0, \alpha \neq 1$, ($\alpha$-degree entropy of Renyi)
- $H_P(\Pi) = -\sum \pi_i log \pi_i - \sum (1-\pi_i) log(1-\pi_i)$, (paired Shannon entropy)
- $H_\gamma(\Pi) = \frac{1-(\sum \pi_i{}^{1/\gamma})^\gamma}{1-2^{\gamma-1}}, \gamma > 0, \gamma \neq 1$, ($\gamma$-entropy)
- $H_G(\Pi) = 1 - \sum \pi_i^2$, (Gini-Simpson entropy)

These entropy functions have been widely used in a variety of studies in genetics (Karlin, Kenett, and Bonne-Tamir, 1979), in anthropology (Rao, 1977), in biology Lewontin,

This dissertation follows the style of *Journal of Agricultural, Biological, and Environmental Statistics*.

1972), in ecology (Pielou, 1975), in economics (Sen, 1973) and in sociology (Agresti and Agresti, 1978), and so forth.

While some of these measures are derived from mathematically well postulated axioms, most are based on heuristic considerations and others are constructed assuming some models for genetic and environmental mechanisms causing differences between individuals and populations. However, these entropies, as shown in Rao (1982b), do not possess higher order convexity properties necessary for carrying out analysis of diversity (ANODIV) similar to analysis of variance (ANOVA). Rao (1982a) introduced a new measure called **Rao's quadratic entropy** which possesses these properties.

## I.2 RAO'S QUADRATIC ENTROPY

Rao (1982a) introduced a general diversity measure called Rao's quadratic entropy (QE):

$$H_Q(\Pi) = \sum\sum d_{ij}\pi_i\pi_j = \Pi'\Delta\Pi, \tag{I.2.1}$$

where $\Delta = (d_{ij})$, $d_{ij}$ is a nonnegative number representing the difference between the categories $i$ and $j$, so that $H_Q(\Pi)$ is the average difference between two individuals drawn at random from a population.

Let $d_{ij} = 1$, if $i \neq j$ and $d_{ii} = 0$; then

$$H_Q(\Pi) = 1 - \sum \pi_i^2 = H_G,$$

which is Gini-Simpson entropy.

Generally, Rao's quadratic entropy is determined by first choosing a non-negative symmetric function d($X_1$, $X_2$), which is a measure of difference between two individuals with $X$=$X_1$ and $X$=$X_2$. The quadratic entropy of any distribution function with $d(X_1,X_2)$ is defined as the function (Rao, 1982c):

$$H_Q = \int d(X_1,X_2)P(dX_1)P(dX_2). \tag{I.2.2}$$

This function $d(X_1,X_2)$ is a kernel function and satisfies the following properties: (Liu, 1991; Liu and Rao, 1995)

(1) $d(X_1,X_2)$ is symmetric and

$$d(x_1,x_2) \begin{cases} > 0 & \text{if } x_1 \neq x_2; \\ = 0 & \text{if } x_1 = x_2. \end{cases} \tag{I.2.3}$$

(2) It is conditionally negative definite, i.e.,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d(x_i,x_j)a_ia_j \leq 0, \tag{I.2.4}$$

for every integer $n$ and choices of $x_1,...,x_n$ and numbers $a_1,...,a_n$ such that $\sum_{i=1}^{n} a_i = 0$.

### I.2.1 Examples of Rao's Quadratic Entropy

In the following we provide two examples of Rao's quadratic entropy.

**Example 1.1:** Let $X \in R^m$, a real vector space of $m$ dimensions and A is a positive definite matrix. Then define

$$d(X_1,X_2) = (X_1 - X_2)'A(X_1 - X_2).$$

Let $X \sim (\mu_i,\Sigma_i)$, (i.e., X is distributed with mean vector $\mu_i$ and variance matrix $\Sigma_i$ and not necessarily multivariate normal). Then

$$H_i = 2tr(A\Sigma_i). \tag{I.2.5}$$

**Note 1.1:** Under univariate case, define a kernel function $d(x_1,x_2) = \frac{1}{2}(x_1 - x_2)^2$ which satisfies (I.2.3) and (I.2.4). Hence, $H = E[\frac{1}{2}(x_1 - x_2)^2]$ is a quadratic entropy for i.i.d. $x_1$, $x_2$. In this case, quadratic entropy is nothing but the variance.

**Example 1.2:** Let $X = (x_1,...,x_m)$, where $x_i$ can take only a finite number of values. In such a case the kernel function between $X_1$ and $X_2$ is $d(X_1,X_2) = m - \sum d_r$, where $d_r = 1$ if the $rth$ components of $X_1$ and $X_2$ agree and zero otherwise. Let $X_r$ take different values with probabilities $(p_{ir1}, p_{ir2}, ..., p_{irk_r})$ in population $\Pi_i$. Define

$$j_{ii}^{(r)} = E(d_r) = \sum_{s=1}^{k_r} p_{irs}p_{jrs},$$

when $X_1$ is drawn from $\Pi_i$ and $X_2$ is drawn from $\Pi_j$. Then

$$H_i = \sum_{r=1}^{m}(1 - j_{ii}^{(r)}) = m(1 - J_{ii}).$$

**Note 1.2:** When m=1, quadratic entropy is reduced to Gini-Simpson index,

$$H = 1 - \sum_{i=1}^{r} p_i^2.$$

From the examples above, it can be seen that the general approach in using quadratic entropy is first to define a function $d(X_1, X_2)$ measuring the difference between individuals $X_1$ and $X_2$ and use the probability distribution of $X_1$ and $X_2$ to find the average of $d(X_1, X_2)$. In practice, the function $d(X_1, X_2)$ can be chosen to reflect some intrinsic dissimilarity between individuals according certain investigation. This measure of entropy also is nonnegative, attains the maximum for the uniform distribution and has the minimum when the distribution is degenerate.

### I.2.2 Decomposition of Quadratic Entropy

The concavity of quadratic entropy can be easily verified (Rao, 1982c). In Equation (I.2.5) the quadratic entropy $H_i$ is defined as the average difference between two randomly drawn individuals from $\Pi_i$. Suppose that two individuals are from different populations, that is, one individual is drawn from $\Pi_i$ and another from $\Pi_j$.

$$H_{Q,i} = \int d(X_1,X_2)P_i(dX_1)P_i(dX_2);\ H_{Q,j} = \int d(X_1,X_2)P_j(dX_1)P_j(dX_2).$$

$$H_{Q,ij} = \int d(X_1,X_2)P_i(dX_1)P_j(dX_2),$$

$$D_{ij} = H_{Q,ij} - \frac{1}{2}(H_{Q,i} + H_{Q,j}).$$

For a mixed population $\Pi_\lambda$, where $\Pi_\lambda = \lambda\Pi_i + (1-\lambda)\Pi_j, 0 < \lambda < 1$ then

$$H_Q^\lambda = \int d(X_1,X_2)P_\lambda(dX_1)P_\lambda(dX_2) = \lambda^2 H_{Q,i} + (1-\lambda)^2 H_{Q,j} + 2\lambda(1-\lambda)H_{Q,ij}.$$

$$H_Q^\lambda - (\lambda H_{Q,i} + (1-\lambda)H_{Q,j}) = 2\lambda(1-\lambda)D_{ij},$$

$D_{ij} \geq 0$ ensures the concavity of $H_Q$ and vice versa (Rao, 1982c). $D_{ij}$ is also termed as the **Jensen difference** which is a measure of dissimilarity between $\Pi_i$ and $\Pi_j$.

**Note 1.3:** In the definition of quadratic entropy (Equation I.2.5) no condition is imposed on the function $d(X_1, X_2)$ except that it should be nonnegative. The logical requirement that the Jensen difference should be nonnegative restricts the choice of $d(X_1, X_2)$ to functions that induce a concave quadratic function.

The concavity property of Rao's quadratic entropy enables us to decompose the diversity in a mixed population in a natural way, as diversity between and within populations. If $P_1, P_2, ..., P_k$ are the distributions of X in $\Pi_1, \Pi_2, ..., \Pi_k$ and $\lambda_1, \lambda_2, ..., \lambda_k$ are the priori probabilities ($\sum \lambda_i = 1$), then the diversity in the mixture $\lambda_1 P_1 + \lambda_2 P_2 + ... + \lambda_k P_k$ can be decomposed as,

$$H_Q = H(\lambda_1 P_1 + \lambda_2 P_2 + ... + \lambda_k P_k) = \sum \lambda_i H_i + \sum\sum \lambda_i \lambda_j D_{ij} = SSW + SSB, \tag{I.2.6}$$

where $D_{ij} = H_{ij} - (H_i + H_j)/2$ is the Jensen difference between $\Pi_i$ and $\Pi_j$. *SSW* is the weighted average of the diversities within populations. *SSB* is the weighted average of the dissimilarity between all pairs of populations, which is nonnegative and vanishes only if $\Pi_1 = \Pi_2 = ... = \Pi_k$.

**Decomposition for Example 1.1:** Let us consider $k$ populations as in Example 1 of Section I.2.1. The $m$-vector variable $X \sim (\mu_i, \Sigma_i)$,

$$H_i = 2tr(A\Sigma_i),$$

$$H_{ij} = tr(A\Sigma_i) + tr(A\Sigma_j) + \delta_{ij}' A \delta_{ij},$$

where $\delta_{ij} = \mu_i - \mu_j$. The Jensen difference $D_{ij} = \delta_{ij}' A \delta_{ij}$ becomes Mahalanobis distance between $\Pi_i$ and $\Pi_j$ if $\Sigma_1 = \Sigma_2 = ... = \Sigma_k = \Sigma$ and $A = \Sigma^{-1}$. Further let $\Pi_Q$ be a mixture of $\Pi_1, ..., \Pi_k$ with a priori probabilities $\lambda_1, ..., \lambda_k$. Then using Equation (I.2.6), the decomposition becomes

$$H_Q = SSW + SSB = 2m + \sum\sum \lambda_i \lambda_j \delta_{ij}' \Sigma^{-1} \delta_{ij}. \tag{I.2.7}$$

Thus the diversity within population is $2m$ and the diversity between populations is the weighted combination of Mahalanobis $D^2$'s for all pairs of populations. Note here the normality of X is not required.

**Decomposition for Example 1.2:** For multinomially distributed variables $X = (x_1, ..., x_m)$, let the mixture of $\Pi_1, \Pi_2, ..., \Pi_k$ be denoted by $\Pi_Q$ with a priori probabilities $\lambda_1, \lambda_2, ..., \lambda_k$.

$$H_i = \sum_{r=1}^{m} (1 - j_{ii}^{(r)}) = m(1 - J_{ii}),$$

$$H_{ij} = \sum_{r=1}^{m} (1 - j_{ij}^{(r)}) = m(1 - J_{ij}),$$

Then the Jensen difference

$$D_{ij} = H_{ij} - \frac{1}{2}(H_i + H_j) = m[\frac{1}{2}(J_{ii} + J_{jj}) - J_{ij}] = \frac{1}{2}\sum_{r=1}^{m}\sum_{s=1}^{k_r}(p_{irs} - p_{jrs})^2.$$

In this case, Equation (I.2.6) becomes,

$$H_Q = m[\sum \lambda_i(1 - J_{ii}) + \sum\sum \lambda_i\lambda_j(\frac{1}{2}J_{ii} + \frac{1}{2}J_{jj} - J_{ij})],$$

which is the decomposition obtained by Nei (1973).

## I.3 OVERVIEW OF THESIS

The objective of this thesis is to provide modified methods to the analysis of diversity with Rao's quadratic entropy and then explore its new applications in analyzing categorical data in several scenarios. This thesis consists of six chapters.

After the introduction of quadratic entropy in Chapter I, several distance matrices are used to modify the quadratic entropy in Chapter II. The decomposition of quadratic entropy is proposed for analyzing categorical data similar to analysis of variance (ANOVA) for continuous data. Theoretically and empirically it is shown to have good performance.

The application of quadratic entropy in measuring and testing biodiversity is explored in Chapter III. Practically equivalent confidence intervals are constructed to compare biodiversity with bootstrap methods. The simulation is performed to compare the methods with those based on Shannon entropy. Simulation data and real dinosaur data are analyzed for illustrations of the methods.

In Chapter IV, a new distance is constructed based on quadratic entropy to cluster multinomially distributed data. Hierarchical methods are applied on both simulated and real data to compare with Euclidean distance and Bhattacharyya distance.

The application of quadratic entropy to the multi-response data is studied in Chapter V. A method based on bootstrap samples is proposed and compared with adjusted Pearson $\chi^2$ statistics. Both real and simulated data sets are used to illustrate and evaluate the method.

Chapter VI is another application of quadratic entropy in cluster analysis. Large scale data such as gene expression data is the focus of this chapter. A new minimum entropy criterion is developed based on quadratic entropy. A hill-climbing type iterative algorithm is applied to both simulation and real gene expression data. The quadratic entropy criteria is compared with other standard clustering methods by applying the adjusted Rand index as the measure of agreement.

# CHAPTER II

# ANALYSIS OF RAO'S QUADRATIC ENTROPY

In many statistical problems, the data can be formulated in the general factor-response framework, where one is interested in the estimation and testing of the individual as well as the interaction effects of the factors on the response variable. Practitioners familiar with analysis of variance (ANOVA) have well developed techniques available for the analysis of quantitative variables. However, for categorical variables they must use a completely different set of techniques. Let $\Pi = (\pi_1, \pi_2 ..., \pi_s)$ be the probability vector of a multinomial population with $s$ categories. Light and Margolin (1971) and Anderson and Landis (1980) used Gini-Simpson entropy

$$H_G(\Pi) = 1 - \Pi'\Pi = 1 - \sum \pi_i^2$$

to develop categorical analysis of variance (CATANOVA) for a nominal response variable. The Gini-Simpson entropy can be interpreted (Rao, 1982a) as the expected distance between two randomly selected individuals when the distance is defined as zero if they belong to the same category and unity otherwise. However in many applications, differences between different categories may not all be equal and hence in those cases it may not be appropriate to use Gini-Simpson entropy for the analysis. Since Rao's quadratic entropy (QE) is the expected distance between two randomly drawn individuals with a predefined distance matrix, this entropy seems like an appropriate choice. Nayak (1986a,b) generalized CATANOVA using Rao's QE,

$$H_Q(\Pi) = \Pi'\Delta\Pi, \tag{II.0.1}$$

where $\Delta_{s\times s} = (d_{ij})$ is a pre-determined distance matrix.

We will review the one-way analysis of diversity using Rao's quadratic entropy in Section II.1 and illustrate it with suggested $\Delta$ matrices proposed in Section II.2. The distribution of the modified quadratic entropy statistics is discussed in Section II.3. The performance of this modified statistics will be tested with real and simulated data in Sections II.4 and II.5.

## II.1 ONE-WAY ANALYSIS OF DIVERSITY USING RAO'S QUADRATIC ENTROPY

In Section I.2 we discuss the concavity properties of the Rao's quadratic entropy and the decomposition of total diversity within population and dissimilarity between populations. These properties apply to categorical case. Let $\Pi_1, \Pi_2, ..., \Pi_r$ be the probability vectors of $r$ multinomial populations and $\lambda_1, \lambda_2, ..., \lambda_r$ ($\sum \lambda_i = 1$) be the associated prior probabilities. Then for the mixed population $\bar{\Pi} = \sum \lambda_i \Pi_i$ we have the following decomposition of the total diversity $H(\bar{\Pi})$:

$$H(\bar{\Pi}) = \sum \lambda_i H(\Pi_i) + \sum \lambda_i (\Pi_i - \bar{\Pi})' \Delta (\Pi_i - \bar{\Pi}),$$

$$SST = SSW + SSB.$$

In practice, usually the population probabilities are not known and they are estimated from the sample observations. Nayak (1986a,b) derived standard errors and asymptotic distributions of sample diversities for one factor $X$. In particular, he proved that asymptotically: (i) $S\hat{S}T$ and $S\hat{S}B$ are independently distributed; and (ii) $S\hat{S}B$ is distributed as a linear combination of $\chi^2$ variables. Below we briefly describe the findings from Nayak (1986a,b).

Let $n_{ij}$, $i = 1, ..., r$, $j = 1, ..., s$, denote the number of responses in the $j$-th category for the $i$-th level of X;

$n_{i.} = \sum_j n_{ij}$, $n_{.j} = \sum_i n_{ij}$ and $n_{..} = \sum\sum n_{ij}$;

$V_i = (n_{i1}, ..., n_{is})'$, vector of frequencies in the $i$-th level of X;

$V = (n_{11}, ..., n_{1s}, n_{21}, ..., n_{2s}, ..., n_{rs})'$;

$\hat{\Pi}_i = n_{i.}^{-1} V_i$, the observed proportions in the $i$-th level of X;

$\hat{\bar{\Pi}} = n_{..}^{-1} \sum V_i$, the observed proportions in the combined sample;

$J$=matrix of unit elements;

$A \otimes B = (a_{ij} B)$, the Kronecker product of A and B.

For a vector $a = (a_1, ..., a_n)'$, we shall use $D_a$ to denote the diagonal matrix with elements $a_1, ..., a_n$.

For statistical inference, we assume that the responses in different levels of X are stochastically independent and $V_i$ follows multinomial law with parameters $n_{i.}$ and $\Pi_i = (\pi_{i1}, ..., \pi_{ik})'$.

With the above notations, the sample analogues of SST, SSW and SSB are as follows:

$S\hat{S}T = \hat{\bar{\Pi}}' \Delta \hat{\bar{\Pi}} = n_{..}^{-2} V' T V$, where $T = J_{r \times r} \otimes \Delta$.

$S\hat{S}W = n_{..}^{-1}\sum n_{i.}\hat{\Pi}_i'\Delta\hat{\Pi}_i = n_{..}^{-1}V'WV$, where $W = diag(1/n_{1.},...,1/n_{r.})\bigotimes\Delta$.
$S\hat{S}B = S\hat{S}T - S\hat{S}W = n_{..}^{-1}V'BV$, where $B = (n_{..}^{-1}T - W)$.

The sample diversities $S\hat{S}T$, $S\hat{S}W$ and $S\hat{S}B$ are the maximum likelihood estimators of the corresponding population diversities SST, SSW and SSB.

Nayak (1986a,b) derived the asymptotic distribution of the sample diversities and the results are given in the following two theorems:

**Theorem 1.1** Under $H_0 : \Pi_1 = \Pi_2 = ... = \Pi_r = \Pi$, asymptotically as $n_{i.} \to \infty$ and $n_{i.}/n_{..} \to \lambda_i$ (a fixed prior probability), $n_{..}S\hat{S}T$ and $n_{..}S\hat{S}B$ are independently distributed.

**Theorem 1.2** Under $H_0$, asymptotically as $n_{i.} \to \infty$ and $n_{i.}/n_{..} \to \lambda_i$ (a fixed prior probability),

$$n_{..}S\hat{S}B \sim \sum_{i=1}^{s-1}\alpha_i\chi^2_{i(r-1)}, \tag{II.1.1}$$

where $\alpha_i$, $i = 1,...,s-1$, are the possible nonzero eigenvalues of $(-\Delta\Sigma)$ and the $\{\chi^2_{i(r-1)}\}$ are independent $\chi^2$ random variables with $(r-1)$ d.f.. Here $\Sigma = D_\Pi - \Pi\Pi'$, where $D_\Pi = diag(\pi_1,\pi_2,...,\pi_s)$.

The asymptotic distribution of $S\hat{S}B$ given in Equation (II.1.1) depends on $\alpha_i$, which are functions of the unknown matrix $\Sigma$. Replacing $\alpha_i$ by $\bar{\alpha} = \sum\alpha_i/(k-1)$ in Equation (II.1.1) the distribution of $n_{..}S\hat{S}B$ can be approximated by $\bar{\alpha}\chi^2_{(r-1)(k-1)}$. Using an unbiased estimate of $\bar{\alpha} = tr(-\Delta\Sigma)/(k-1)$ as $n_{..}S\hat{S}T/[(k-1)(n_{..}-1)]$, the distribution of $C_\Delta = (k-1)(n_{..}-1)S\hat{S}B/S\hat{S}T$ can be approximated by $\chi^2_{(k-1)(r-1)}$. Thus a simple test for $H_0$ provided by Nayak (1986b) is $C_\Delta$, and reject $H_0$ at level $\alpha$ when

$$C_\Delta = (s-1)(n_{..}-1)S\hat{S}B/S\hat{S}T > \chi^2_{\alpha;(s-1)(r-1)}. \tag{II.1.2}$$

See Nayak (1986a,b) for proof of these results and more details.

In Nayak (1986a,b)'s attempt for using analysis of diversity with Rao's quadratic entropy, one of the unresolved issue is the choice of $\Delta$. In practice it is usually arbitrary and based on an individual's assessment of the differences with reference to the problem under investigation. This has restricted the applications of quadratic entropy. Here we have proposed several ways to select $\Delta$ based on the frequency table. However, it will make the derivation of the asymptotic distribution of statistics $S\hat{S}B/S\hat{S}T$ difficult. Alternatively, $S\hat{S}B$ can be used as the test statistics. $\Sigma$ can be estimated by its unbiased estimator $\hat{\Sigma}$ and $\alpha_i$ be replaced by its estimates. Then an algorithm proposed by Davis (1980) can be used to get the exact distribution of the linear combination of $\chi^2$ variables, that's to say, the distribution of $\sum_{i=1}^{k-1}\alpha_i\chi^2_{i(r-1)}$. However, as described later, we have resolved to using bootstrap method for determining the approximate distributions.

## II.2 VARIOUS CHOICES FOR MATRIX Δ

In this section we will discuss several methods that can be used to find the distance matrix Δ. Liu (1991) and Liu and Rao (1995) described that in constructing quadratic entropy the distance function $d(x_1, x_2) : X^2 \rightarrow \mathbf{R}$ has the properties:

- $d(x_1, x_2) > 0$ if $x_1 \neq x_2$; $d(x_1, x_2) = 0$ if $x_1 = x_2$;
- $d(\cdot, \cdot)$ is conditionally negative definite, i.e. $\sum_{i=1}^{n} \sum_{j=1}^{n} d(x_i, x_j) a_i a_j \leq 0$ for every integer n and choices $x_1, ..x_n$ in X and $a_1, ..., a_n$ in $\mathbf{R}$ such that $a_1 + a_2 + ... + a_n = 0$.

The distance matrix Δ satisfying these two properties can be constructed in following ways.

**1. Δ Based upon the Variables Measured Scores**

The item $d_{ij}$ in Δ is the distance between the $i$-th level and $j$-th level of the variable. So, we can use the scores to scale the ordinal variables and then compute the distance between different levels as $d_{ij} = |S_i - S_j|$ (Stokes, Davis, and Koch, 2005).

- Table Scores

  For the ordinal variables, table scores ($S_{1i}$) are the values of the ordered levels. If the variables are nominal, the table scores ($S_{1i}$) are the numeric value corresponding to that level;

- Rank Scores

  Rank scores, which are defined by the frequencies: $S_{2i} = \sum_{s<i} n_{s.} + (n_{i.} + 1)/2$;

- Ridit Scores

  Ridit scores are standardized by the sample size and can be derived from rank scores as $S_{3i} = S_{2i}/n$;

- Modified Ridit Scores

  Modified ridit scores represent the expected values of the order statistics for the uniform distribution on (0,1). Modified ridit scores are derived from rank scores as $S_{4i} = S_{2i}/(n+1)$.

**2. Δ Based upon Distances**

- Let $U = (\frac{n_{1.}}{n_{..}}, \frac{n_{2.}}{n_{..}}, \dots \frac{n_{r.}}{n_{..}})$, $D_U = Diag(U)$ and $C_i = (\frac{n_{1i}}{n_{1.}}, \frac{n_{2i}}{n_{2.}}, \dots, \frac{n_{ri}}{n_{r.}})'$,

  **Euclidean distance** is defined as,

$$d_{ij} = \sqrt{(C_i - C_j)'(C_i - C_j)}.$$

- Using the same notation as above, **Chi-square distance** between the $i$-th level and $j$-th level of the variable is defined as,

$$d_{ij} = \sqrt{(C_i - C_j)' D_U^{-1} (C_i - C_j)}$$

- **Nei's Distance** between the $i$-th and $j$-th category of the response variable is defined as

$$d_{ij}^2 = (Q_i - Q_j)'(Q_i - Q_j),$$

  where $Q_i = (\frac{n_{1i}}{n_{.j}}, \frac{n_{2i}}{n_{.i}}, \dots, \frac{n_{ri}}{n_{.i}})'$.

- **Ochiai's Distance** is suitable for binary data. When comparing the $i$-th and $j$-th level of the variable, let a(1,1), b(1,0), c(0,1) and d(0,0) be the number of pairs for value (1,1), (1,0), (0,1) and (0,0), where a+b+c+d=r, Ochiai's distance is defined as

$$d_{ij} = \sqrt{1 - t_{ij}},$$

  where $t_{ij} = \frac{a}{\sqrt{(a+b)(a+c)}}$.

**3. Δ Based On Probabilities**

We provide two choices for Δ here. Take $\Delta = (d_{ij})$, where

$$d_{ij} = \begin{cases} |\pi_{.i} - \pi_{.j}| + 1 & if \quad i \neq j \\ 0 & if \quad i = j \end{cases}, \tag{II.2.1}$$

and

$$d_{ij} = \begin{cases} 0 & if & i = j \\ 1 & if & \pi_{.i} = \pi_{.j} = 0 \\ |log(\pi_{.i})| + 1 & if & \pi_{.j} = 0 \\ |log(\pi_{.j})| + 1 & if & \pi_{.i} = 0 \\ |log(\pi_{.i}) - log(\pi_{.j})| + 1 & else. & \end{cases} \tag{II.2.2}$$

Here $\pi_{.i}$ and $\pi_{.j}$ are the corresponding probabilities at the $i$-th and $j$-th categories of $\bar{\Pi} = \sum_r \lambda \Pi_r$. In practice they can be replaced by their estimators $\hat{\pi}_{.i} = n_{.i}/n_{..}$ and $\hat{\pi}_{.j} = n_{.j}/n_{..}$.

TABLE 1. Distribution of parties in neighborhoods

| | Neighbor | | | |
|---|---|---|---|---|
| Party | Bayside | Highland | Longview | Sheffeld |
| Democrat | 221 | 160 | 360 | 140 |
| Independent | 200 | 291 | 160 | 311 |
| Republican | 208 | 106 | 316 | 97 |

## II.3 BOOTSTRAP FOR THE DISTRIBUTION OF $S\hat{S}B$

Since the proposed Δ's are to be estimated from the observed data, the asymptotic distribution of the modified statistics $S\hat{S}B$ is more complicated than that in Equation (II.1.1). However, if we base our tests on conditional distribution given the marginal frequencies, the asymptotic distribution is a linear combination of $\chi_r^2$'s with positive coefficients. The explicit expressions for these coefficients are very difficult to find. From the point of view of application, it is necessary to find a more computable approach for approximating the distribution of the statistic $S\hat{S}B$. We propose to use the bootstrap method for this.

## II.4 A REAL LIFE EXAMPLE

The data in Table 1 are from a study concerning the distribution of party affiliation in a city suburb (Stokes et al., 2005). The data consists of a factor: Neighborhood (X) with 4 levels (Bayside=1, Highland=2, Longview=3, and Sheffeld=4) and a response variable: Party (Y) with 3 levels (Democrat=1, Independent=2, and Republican=3). Researcher might be interested in whether there is an association between registered political party and the neighborhood they live in.

To determine the effects of X on Y, we perform an analysis of diversity using the following methods:

1. Pearson statistics
2. Fisher's exact test
3. CATANOVA

TABLE 2. Analysis of diversity for the political parties data

| Methods | P-value |
|---|---|
| Pearson $\chi^2$ | $< 0.0001$ |
| Fisher's Exact | 1.5182E-09 |
| CATANOVA | 1.601E-10 |
| $C_{\Delta_1}$ | 2.71E-10 |
| $C_{\Delta_2}$ | 2.5496E-09 |

4. $C_{\Delta_1} = (s-1)(n_{..}-1)S\hat{S}B_1/S\hat{S}T_1$ with

$$\Delta_1 = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$$

5. $C_{\Delta_2} = n_{..}S\hat{S}B_2$ and bootstrap approximation with $\Delta_2 = (d_{ij})$, where

$$d_{ij} = \begin{cases} 0 & if \quad i=j \\ 1 & if \quad \pi_i = \pi_j = 0 \\ |log(\pi_i)|+1 & if \quad \pi_j = 0 \\ |log(\pi_j)|+1 & if \quad \pi_i = 0 \\ |log(\pi_i)-log(\pi_j)|+1 & else. \end{cases}$$

which based upon the data becomes,

$$\Delta_2 = \begin{bmatrix} 0 & 1.088 & 1.192 \\ 1.088 & 0 & 1.280 \\ 1.192 & 1.280 & 0 \end{bmatrix}.$$

The distribution of $C_{\Delta_2}$ were simulated using a nonparametric bootstrap procedure with $B = 5000$ bootstrap samples.

All methods indicate strong evidence against independence as shown in Table 2. If a Bonferroni adjusted significance level of 0.05/6=0.0083 is used, the pair of Longview and Sheffeld neighbor are found significantly different from each other.

## II.5 EMPIRICAL NULL DISTRIBUTION AND POWER COMPARISON

In this section, we will examine the accuracy of the approximate asymptotic null distribution theory by using simulated data. The performance of Rao's quadratic entropy with the previously defined distances will be compared with Pearson $\chi^2$ test, Fisher Exact test, and Gini-Simpson test of Light and Margolin (1971).

Nayak (1986b) studied the empirical significance level of $C_\Delta$ test with reference to critical points $\chi^2_{\alpha;(s-1)(r-1)}$ for 13 populations with different distributions. See our Table 3. We use the same settings and two of the same distance matrices $\Delta_1$ and $\Delta_2$ used by Nayak (1986b) for easy comparisons. In the examples, there are 3 response categories and 2 levels of X. We assume a common probability distribution for both levels of X, given in the first column in Table 3. The second column in Table 3 gives the sample sizes, i.e., the values of $n_{1.}$ and $n_{2.}$. All the distances proposed in Section II.2 have been explored. However, the matrices $\Delta_3$ and $\Delta_4$ have produced more meaningful results. Hence, results corresponding to only those are presented in Table 3.

$$\Delta_1 = \begin{bmatrix} 0 & 1 & 1.5 \\ 1 & 0 & 1 \\ 1.5 & 1 & 0 \end{bmatrix}$$

$$\Delta_2 = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$$

$\Delta_3 = (d_{3,ij})$, where

$$d_{3,ij} = \begin{cases} |\pi_{.i} - \pi_{.j}| + 1 & if \quad i \neq j \\ 0 & if \quad i = j \end{cases} .$$

$\Delta_4 = (d_{4,ij})$, where

$$d_{4,ij} = \begin{cases} 0 & if \quad i = j \\ 1 & if \quad \pi_{.i} = \pi_{.j} = 0 \\ |log(\pi_{.i})| + 1 & if \quad \pi_{.j} = 0 \\ |log(\pi_{.j})| + 1 & if \quad \pi_{.i} = 0 \\ |log(\pi_{.i}) - log(\pi_{.j})| + 1 & else. \end{cases}$$

The distribution of $C_{\Delta_1}$ and $C_{\Delta_2}$ are approximated by $\chi^2_{\alpha,2}$ as stated in Equation (II.1.2). Because the distribution of $C_{\Delta_3} = S\hat{S}B_3$ and $C_{\Delta_4} = S\hat{S}B_4$ are very complicated, in that the asymptotic distribution cannot be easily determined, the nonparametric bootstrap procedure is used to determine the p-values. The algorithm is described in the following steps (Efron and Tibshirani, 1993):

(1) Take $B$ re-samples of size $n$ by randomly selecting subjects with replacement from the original data set independently within each group;

(2) for each re-sample, calculate the test statistic, $C^*_\Delta$, for $b = 1, ..., B$ and

(3) calculate the p-value as $B^{-1}\sum_b I(C^*_\Delta > C_\Delta)$.

### II.5.1 Empirical Level of Significance

The empirical type I errors are presented in Table 3 under test statistics as column headings. In each case 10,000 independent samples were generated and used to compute the rejection probabilities for $\alpha = 0.1, 0.05, 0.01$. For each sample, 500 bootstrap re-samples were generated for the computations.

In Table 3, we see that the level of significance of all statistics are all close to $\alpha$, except that CATANOVA, which is very liberal; $QE_{\Delta_1}$ is more accurate than $QE_{\Delta_2}$; $QE_{\Delta_4}$ is more accurate than $QE_{\Delta_3}$; among all the quadratic entropy statistics the empirical significance level of $QE_{\Delta_4}$ is most close to $\alpha$. Hence one should feel comfortable using the distance matrix $\Delta_4$ in practice, with p-values computed using the bootstrap method.

### II.5.2 Empirical Power

We also compared the empirical powers of $C_\Delta$ with Pearson $\chi^2$, Fisher's exact test and CATANOVA for 10 different alternatives in the case of two levels of X and 3 response categories. In each case $n_{1.}$ and $n_{2.}$ were fixed at 100. The probabilities associated with one level of X are $\Pi_1 = (1/3, 1/3, 1/3)$ and the probabilities for the other level $\Pi_2$ are given in Table 4, 5 and 6. In the first five cases $\Pi_2$ is of the form $(p, q, q)$ with $p > 1/3$ and the departure of $\Pi_2$ from $\Pi_1$ is towards a vertex of the simplex. For the last five cases the departure is towards a base of the simplex. Since, unlike $\chi^2$ and Gini-Simpson entropy, the powers of $C_\Delta$ are not symmetric in the arguments of $\Pi_2$ in our study. We have also considered the permutation of $\Pi_2$ in out study. In each case 1000 independent samples were used to estimate the empirical power for $\alpha$=0.01, 0.05 and 0.10. For each sample, 500 bootstrap re-samples were generated for the computation. The results are reported in Tables 4, 5 and 6.

In Tables 4, 5 and 6 we observe the following: (1) The powers of $QE_{\Delta_1}$, $QE_{\Delta_2}$, $QE_{\Delta_3}$ and $QE_{\Delta_4}$ are larger than CATANOVA; (2) The powers of $QE_{\Delta_4}$ are larger than $QE_{\Delta_1}$, $QE_{\Delta_2}$ and $QE_{\Delta_3}$; (3) $\chi^2$ and Fisher test usually perform better than $QE_{\Delta_1}$, $QE_{\Delta_2}$, $QE_{\Delta_3}$ and $QE_{\Delta_4}$ for the departures of $\Pi_2$ towards the base of the simplex.

TABLE 3. Empirical significance level of Rao's quadratic entropy statistics

| Probability Structure | Sample Size | $\alpha$ | $\chi^2$ | FISHER | CATANOVA | $QE_{\Delta_1}$ | $QE_{\Delta_2}$ | $QE_{\Delta_3}$ | $QE_{\Delta_4}$ |
|---|---|---|---|---|---|---|---|---|---|
| 0.33, 0.33, 0.34 | 30 30 | 0.1 | 0.11 | 0.095 | 0.106 | 0.109 | 0.117 | 0.104 | 0.102 |
| | | 0.05 | 0.047 | 0.042 | 0.05 | 0.051 | 0.057 | 0.044 | 0.044 |
| | | 0.01 | 0.011 | 0.011 | 0.009 | 0.011 | 0.013 | 0.009 | 0.009 |
| 0.30, 0.30, 0.40 | 25 40 | 0.1 | 0.121 | 0.111 | 0.127 | 0.12 | 0.124 | 0.121 | 0.113 |
| | | 0.05 | 0.059 | 0.059 | 0.056 | 0.056 | 0.062 | 0.052 | 0.052 |
| | | 0.01 | 0.008 | 0.007 | 0.008 | 0.011 | 0.014 | 0.01 | 0.011 |
| 0.25, 0.30, 0.45 | 30 30 | 0.1 | 0.112 | 0.103 | 0.119 | 0.125 | 0.124 | 0.118 | 0.111 |
| | | 0.05 | 0.051 | 0.046 | 0.059 | 0.055 | 0.062 | 0.058 | 0.055 |
| | | 0.01 | 0.008 | 0.008 | 0.006 | 0.012 | 0.015 | 0.006 | 0.006 |
| 0.25, 0.50, 0.25 | 15 15 | 0.1 | 0.108 | 0.086 | 0.091 | 0.105 | 0.096 | 0.09 | 0.099 |
| | | 0.05 | 0.043 | 0.04 | 0.041 | 0.042 | 0.05 | 0.04 | 0.044 |
| | | 0.01 | 0.007 | 0.011 | 0.014 | 0.011 | 0.013 | 0.015 | 0.012 |
| 0.20, 0.50, 0.30 | 25 30 | 0.1 | 0.104 | 0.089 | 0.095 | 0.094 | 0.096 | 0.09 | 0.092 |
| | | 0.05 | 0.047 | 0.043 | 0.049 | 0.051 | 0.059 | 0.047 | 0.053 |
| | | 0.01 | 0.008 | 0.008 | 0.014 | 0.008 | 0.01 | 0.014 | 0.016 |
| 0.50, 0.30, 0.20 | 50 50 | 0.1 | 0.096 | 0.093 | 0.093 | 0.092 | 0.093 | 0.094 | 0.097 |
| | | 0.05 | 0.053 | 0.048 | 0.051 | 0.05 | 0.051 | 0.056 | 0.055 |
| | | 0.01 | 0.01 | 0.01 | 0.014 | 0.018 | 0.023 | 0.013 | 0.016 |
| 0.60, 0.30, 0.10 | 30 50 | 0.1 | 0.085 | 0.093 | 0.098 | 0.102 | 0.102 | 0.099 | 0.092 |
| | | 0.05 | 0.034 | 0.053 | 0.056 | 0.058 | 0.058 | 0.056 | 0.054 |
| | | 0.01 | 0.006 | 0.012 | 0.021 | 0.026 | 0.026 | 0.025 | 0.022 |
| 0.60, 0.10, 0.30 | 25 25 | 0.1 | 0.097 | 0.069 | 0.107 | 0.114 | 0.115 | 0.108 | 0.1 |
| | | 0.05 | 0.042 | 0.046 | 0.074 | 0.081 | 0.087 | 0.072 | 0.06 |
| | | 0.01 | 0.004 | 0.009 | 0.015 | 0.02 | 0.022 | 0.013 | 0.009 |
| 0.30, 0.60, 0.10 | 70 70 | 0.1 | 0.095 | 0.094 | 0.097 | 0.104 | 0.108 | 0.097 | 0.101 |
| | | 0.05 | 0.048 | 0.044 | 0.052 | 0.048 | 0.052 | 0.05 | 0.051 |
| | | 0.01 | 0.01 | 0.01 | 0.017 | 0.014 | 0.019 | 0.019 | 0.012 |
| 0.70, 0.15, 0.15 | 60 60 | 0.1 | 0.096 | 0.089 | 0.098 | 0.102 | 0.11 | 0.101 | 0.109 |
| | | 0.05 | 0.049 | 0.047 | 0.06 | 0.062 | 0.072 | 0.067 | 0.066 |
| | | 0.01 | 0.009 | 0.01 | 0.008 | 0.016 | 0.019 | 0.016 | 0.019 |
| 0.10, 0.70, 0.20 | 60 50 | 0.1 | 0.103 | 0.095 | 0.099 | 0.097 | 0.095 | 0.098 | 0.106 |
| | | 0.05 | 0.052 | 0.05 | 0.059 | 0.056 | 0.053 | 0.059 | 0.064 |
| | | 0.01 | 0.01 | 0.014 | 0.014 | 0.011 | 0.013 | 0.015 | 0.013 |
| 0.10, 0.80, 0.10 | 30 30 | 0.1 | 0.105 | 0.115 | 0.111 | 0.102 | 0.11 | 0.113 | 0.124 |
| | | 0.05 | 0.044 | 0.07 | 0.046 | 0.04 | 0.046 | 0.06 | 0.065 |
| | | 0.01 | 0.007 | 0.01 | 0.013 | 0.015 | 0.009 | 0.016 | 0.018 |
| 0.03, 0.94, 0.03 | 80 80 | 0.1 | 0.084 | 0.115 | 0.096 | 0.091 | 0.091 | 0.107 | 0.107 |
| | | 0.05 | 0.037 | 0.069 | 0.065 | 0.046 | 0.044 | 0.083 | 0.082 |
| | | 0.01 | 0.001 | 0.021 | 0.011 | 0.006 | 0.004 | 0.018 | 0.022 |

TABLE 4. Empirical power comparison of $\chi^2$, Fisher's test, CATANOVA and QE tests

| Probability Structure | $\alpha$ | $\chi^2$ | FISHER | CATANOVA | $QE_{\Delta_1}$ | $QE_{\Delta_2}$ | $QE_{\Delta_3}$ | $QE_{\Delta_4}$ |
|---|---|---|---|---|---|---|---|---|
| 0.40, 0.30, 0.30 | 0.1 | 0.201 | 0.197 | 0.2 | 0.221 | 0.226 | 0.208 | 0.209 |
| | 0.05 | 0.134 | 0.133 | 0.135 | 0.158 | 0.167 | 0.133 | 0.142 |
| | 0.01 | 0.045 | 0.045 | 0.047 | 0.064 | 0.084 | 0.047 | 0.048 |
| 0.44, 0.28, 0.28 | 0.1 | 0.399 | 0.391 | 0.408 | 0.436 | 0.434 | 0.414 | 0.419 |
| | 0.05 | 0.256 | 0.251 | 0.268 | 0.304 | 0.311 | 0.273 | 0.277 |
| | 0.01 | 0.103 | 0.102 | 0.126 | 0.148 | 0.17 | 0.126 | 0.135 |
| 0.52, 0.24, 0.24 | 0.1 | 0.756 | 0.751 | 0.784 | 0.804 | 0.801 | 0.794 | 0.803 |
| | 0.05 | 0.674 | 0.67 | 0.72 | 0.743 | 0.748 | 0.735 | 0.744 |
| | 0.01 | 0.438 | 0.437 | 0.495 | 0.512 | 0.534 | 0.502 | 0.516 |
| 0.60, 0.20, 0.20 | 0.1 | 0.971 | 0.97 | 0.974 | 0.981 | 0.984 | 0.98 | 0.986 |
| | 0.05 | 0.945 | 0.944 | 0.96 | 0.966 | 0.969 | 0.968 | 0.97 |
| | 0.01 | 0.838 | 0.839 | 0.891 | 0.897 | 0.896 | 0.906 | 0.912 |
| 0.72, 0.14, 0.14 | 0.1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 0.05 | 0.998 | 0.998 | 0.999 | 1 | 1 | 1 | 1 |
| | 0.01 | 0.998 | 0.998 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| 0.36, 0.36, 0.28 | 0.1 | 0.189 | 0.185 | 0.182 | 0.194 | 0.194 | 0.186 | 0.189 |
| | 0.05 | 0.099 | 0.096 | 0.094 | 0.104 | 0.111 | 0.099 | 0.102 |
| | 0.01 | 0.027 | 0.025 | 0.021 | 0.04 | 0.052 | 0.021 | 0.025 |
| 0.39, 0.39, 0.22 | 0.1 | 0.467 | 0.462 | 0.441 | 0.477 | 0.481 | 0.449 | 0.468 |
| | 0.05 | 0.345 | 0.343 | 0.308 | 0.357 | 0.385 | 0.313 | 0.327 |
| | 0.01 | 0.168 | 0.164 | 0.151 | 0.181 | 0.204 | 0.153 | 0.163 |
| 0.42, 0.42, 0.16 | 0.1 | 0.822 | 0.816 | 0.791 | 0.821 | 0.83 | 0.805 | 0.827 |
| | 0.05 | 0.721 | 0.718 | 0.651 | 0.707 | 0.723 | 0.678 | 0.718 |
| | 0.01 | 0.521 | 0.514 | 0.415 | 0.496 | 0.551 | 0.449 | 0.518 |
| 0.45, 0.45, 0.10 | 0.1 | 0.986 | 0.985 | 0.969 | 0.975 | 0.979 | 0.976 | 0.984 |
| | 0.05 | 0.955 | 0.955 | 0.918 | 0.947 | 0.955 | 0.935 | 0.955 |
| | 0.01 | 0.889 | 0.886 | 0.786 | 0.846 | 0.866 | 0.828 | 0.892 |
| 0.48, 0.48, 0.04 | 0.1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 0.05 | 1 | 1 | 0.998 | 0.999 | 0.999 | 0.999 | 1 |
| | 0.01 | 0.996 | 0.996 | 0.984 | 0.988 | 0.989 | 0.989 | 0.994 |

TABLE 5. Empirical power comparison of $\chi^2$, Fisher's test, CATANOVA and QE tests (continued)

| Probability Structure | $\alpha$ | $\chi^2$ | FISHER | CATANOVA | $QE_{\Delta_1}$ | $QE_{\Delta_2}$ | $QE_{\Delta_3}$ | $QE_{\Delta_4}$ |
|---|---|---|---|---|---|---|---|---|
| 0.30, 0.30, 0.40 | 0.1 | 0.203 | 0.197 | 0.2 | 0.226 | 0.235 | 0.197 | 0.2 |
| | 0.05 | 0.131 | 0.13 | 0.127 | 0.152 | 0.16 | 0.133 | 0.137 |
| | 0.01 | 0.042 | 0.041 | 0.041 | 0.064 | 0.073 | 0.044 | 0.05 |
| 0.28, 0.28, 0.44 | 0.1 | 0.376 | 0.366 | 0.381 | 0.408 | 0.397 | 0.386 | 0.39 |
| | 0.05 | 0.248 | 0.24 | 0.263 | 0.295 | 0.31 | 0.27 | 0.276 |
| | 0.01 | 0.112 | 0.107 | 0.121 | 0.146 | 0.168 | 0.123 | 0.124 |
| 0.24, 0.24, 0.52 | 0.1 | 0.778 | 0.765 | 0.796 | 0.818 | 0.823 | 0.817 | 0.82 |
| | 0.05 | 0.676 | 0.665 | 0.715 | 0.748 | 0.752 | 0.733 | 0.743 |
| | 0.01 | 0.433 | 0.429 | 0.487 | 0.531 | 0.536 | 0.505 | 0.518 |
| 0.20, 0.20, 0.60 | 0.1 | 0.959 | 0.956 | 0.967 | 0.971 | 0.969 | 0.971 | 0.973 |
| | 0.05 | 0.943 | 0.941 | 0.959 | 0.963 | 0.962 | 0.959 | 0.963 |
| | 0.01 | 0.83 | 0.829 | 0.885 | 0.899 | 0.901 | 0.895 | 0.908 |
| 0.14, 0.14, 0.72 | 0.1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 0.05 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 0.01 | 0.996 | 0.996 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 |
| 0.28, 0.36, 0.36 | 0.1 | 0.177 | 0.177 | 0.172 | 0.189 | 0.198 | 0.173 | 0.173 |
| | 0.05 | 0.104 | 0.101 | 0.1 | 0.118 | 0.126 | 0.1 | 0.099 |
| | 0.01 | 0.021 | 0.021 | 0.02 | 0.027 | 0.037 | 0.02 | 0.022 |
| 0.22, 0.39, 0.39 | 0.1 | 0.475 | 0.468 | 0.441 | 0.466 | 0.472 | 0.448 | 0.47 |
| | 0.05 | 0.362 | 0.357 | 0.323 | 0.363 | 0.385 | 0.34 | 0.358 |
| | 0.01 | 0.153 | 0.154 | 0.129 | 0.172 | 0.198 | 0.135 | 0.146 |
| 0.16, 0.42, 0.42 | 0.1 | 0.836 | 0.832 | 0.79 | 0.83 | 0.837 | 0.812 | 0.83 |
| | 0.05 | 0.754 | 0.748 | 0.695 | 0.73 | 0.746 | 0.715 | 0.749 |
| | 0.01 | 0.51 | 0.508 | 0.425 | 0.5 | 0.541 | 0.459 | 0.505 |
| 0.10, 0.45, 0.45 | 0.1 | 0.986 | 0.985 | 0.976 | 0.984 | 0.985 | 0.986 | 0.987 |
| | 0.05 | 0.97 | 0.97 | 0.949 | 0.962 | 0.97 | 0.96 | 0.972 |
| | 0.01 | 0.893 | 0.89 | 0.8 | 0.853 | 0.873 | 0.839 | 0.89 |
| 0.04, 0.48, 0.48 | 0.1 | 1 | 0.984 | 1 | 1 | 1 | 1 | 1 |
| | 0.05 | 1 | 0.98 | 1 | 0.999 | 0.999 | 1 | 1 |
| | 0.01 | 0.998 | 0.978 | 0.979 | 0.987 | 0.988 | 0.992 | 0.998 |

TABLE 6. Empirical power comparison of $\chi^2$, Fisher's test, CATANOVA and QE tests (continued)

| Probability Structure | $\alpha$ | $\chi^2$ | FISHER | CATANOVA | $QE_{\Delta_1}$ | $QE_{\Delta_2}$ | $QE_{\Delta_3}$ | $QE_{\Delta_4}$ |
|---|---|---|---|---|---|---|---|---|
| | 0.1 | 0.24 | 0.237 | 0.241 | 0.207 | 0.174 | 0.243 | 0.247 |
| 0.30. 0.40, 0.30 | 0.05 | 0.143 | 0.14 | 0.149 | 0.121 | 0.094 | 0.152 | 0.155 |
| | 0.01 | 0.044 | 0.044 | 0.043 | 0.032 | 0.027 | 0.047 | 0.049 |
| | 0.1 | 0.378 | 0.37 | 0.382 | 0.319 | 0.24 | 0.395 | 0.399 |
| 0.28, 0.44, 0.28 | 0.05 | 0.254 | 0.249 | 0.266 | 0.208 | 0.152 | 0.272 | 0.281 |
| | 0.01 | 0.105 | 0.106 | 0.117 | 0.067 | 0.041 | 0.12 | 0.127 |
| | 0.1 | 0.791 | 0.785 | 0.813 | 0.743 | 0.636 | 0.825 | 0.833 |
| 0.24, 0.52, 0.24 | 0.05 | 0.677 | 0.667 | 0.715 | 0.607 | 0.439 | 0.738 | 0.744 |
| | 0.01 | 0.401 | 0.398 | 0.465 | 0.319 | 0.171 | 0.49 | 0.514 |
| | 0.1 | 0.967 | 0.966 | 0.977 | 0.957 | 0.914 | 0.979 | 0.982 |
| 0.20, 0.60, 0.20 | 0.05 | 0.938 | 0.939 | 0.957 | 0.912 | 0.842 | 0.961 | 0.962 |
| | 0.01 | 0.814 | 0.811 | 0.869 | 0.758 | 0.586 | 0.887 | 0.907 |
| | 0.1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0.14, 0.72, 0.14 | 0.05 | 1 | 1 | 1 | 1 | 0.997 | 1 | 1 |
| | 0.01 | 0.998 | 0.998 | 0.998 | 0.998 | 0.985 | 0.999 | 0.999 |
| | 0.1 | 0.165 | 0.161 | 0.156 | 0.144 | 0.125 | 0.157 | 0.162 |
| 0.36, 0.28, 0.36 | 0.05 | 0.101 | 0.1 | 0.092 | 0.075 | 0.072 | 0.093 | 0.099 |
| | 0.01 | 0.022 | 0.022 | 0.018 | 0.017 | 0.022 | 0.019 | 0.018 |
| | 0.1 | 0.494 | 0.484 | 0.459 | 0.364 | 0.261 | 0.465 | 0.482 |
| 0.39, 0.22, 0.39 | 0.05 | 0.344 | 0.335 | 0.31 | 0.215 | 0.14 | 0.322 | 0.329 |
| | 0.01 | 0.163 | 0.162 | 0.138 | 0.071 | 0.043 | 0.143 | 0.163 |
| | 0.1 | 0.803 | 0.802 | 0.763 | 0.654 | 0.503 | 0.786 | 0.8 |
| 0.42, 0.16, 0.42 | 0.05 | 0.712 | 0.706 | 0.639 | 0.484 | 0.335 | 0.669 | 0.702 |
| | 0.01 | 0.508 | 0.506 | 0.425 | 0.214 | 0.111 | 0.456 | 0.5 |
| | 0.1 | 0.988 | 0.986 | 0.973 | 0.926 | 0.809 | 0.979 | 0.989 |
| 0.45, 0.10, 0.45 | 0.05 | 0.965 | 0.964 | 0.937 | 0.847 | 0.675 | 0.95 | 0.968 |
| | 0.01 | 0.87 | 0.871 | 0.781 | 0.536 | 0.26 | 0.815 | 0.867 |
| | 0.1 | 0.999 | 0.975 | 0.997 | 0.995 | 0.98 | 0.999 | 0.999 |
| 0.48, 0.04, 0.48 | 0.05 | 0.999 | 0.982 | 0.997 | 0.982 | 0.905 | 0.999 | 0.999 |
| | 0.01 | 0.999 | 0.979 | 0.987 | 0.896 | 0.643 | 0.996 | 0.998 |

In this chapter, a new distance matrix is proposed to modify Rao's quadratic entropy statistics. Although it brings complication in computation, it makes the measure of diversity generalizable. Nonparametric bootstrap methods are used for the hypothesis testing. If the null hypothesis is rejected, a post-hoc test should be performed. It can be multiple comparisons by applying similar method of analysis of quadratic entropy to each pair of the groups; alternatively, confidence intervals can be constructed for the pairwise differences.

While Rao's quadratic entropy based analysis of diversity can be used to test the independence of response and factor(s), in some other data analysis problems, the entropy functions can be directly applied, especially in ecology data. In the next chapter, we present a case like that.

# CHAPTER III

# ANALYSIS OF BIODIVERSITY

Statistical tests of the equality of dinosaurs biodiversity of different era have been used for determining whether the extinction of the dinosaurs was sudden or gradual over a period of time. If the biodiversity of the community of dinosaur species was different from period to period, then there is a reason to believe that the extinction was gradual; On the other hand, if the biodiversity remained the same through different time periods, then there is a reason to believe that the extinction might have been sudden due to asteroid collision. Sheehan, Fastovsky, Hoffmann, Berghaus, and Gabriel (1991) and Fritsch and Hsu (1999) analyzed a data set on Dinosaurs to check this theory. We provide that data set from Sheehan et al. (1991) here, in Table 7.

## The Dinosaur Data

Dinosaur bones deposited about 2.2 million years were collected from sites in North Dakota and Montana. The formation was divided into three equal stratigraphic intervals, with each third representing approximately 730,000 years. Although it is difficult to distinguish individual species of dinosaurs, it is relatively easy for researchers to classify bones according to their family. In all, eight families were identified. Table 7 lists the name of all eight dinosaur families and the number of individual dinosaurs of each families identified from the two research sites. There are several measures of biodiversity that can be used for measuring the biodiversity of dinosaurs. Suppose in a biological community there are $s$ species and let $\Pi = (\pi_1, ..., \pi_s)'$ be the vector of proportions of these species in the community, then the two well known measures are:

- Shannon index ($H_S$): $H_S = -\sum \pi_i \ln \pi_i$,
- Gini-Simpson Index ($H_G$): $H_G = 1 - \sum \pi_i^2$.

Suppose in a biological community there are $N$ individuals from $s$ species. Let $n_1, ..., n_s$ be the abundance of each species and $\hat{\pi}_1, ..., \hat{\pi}_s$ be the proportions of these species, that is, $\hat{\pi}_i = n_i/N$. To measure the biodiversity, Sheehan et al. (1991) used the Shannon index and tested the hypothesis:

$$H_0 : H_{S_{upper}} = H_{S_{middle}} = H_{S_{lower}}.$$

TABLE 7. List of the dinosaur families and frequency in each intervals.

| Family Names | Upper Interval | Middle Interval | Lower Interval |
|---|---|---|---|
| Ceratopsidae | 50 | 53 | 19 |
| Hadrosauridae | 29 | 51 | 7 |
| Hypsilophodontidae | 3 | 2 | 1 |
| Pachycephalosauridae | 0 | 0 | 0 |
| Tyrannosauridae | 3 | 3 | 2 |
| Ornithominidae | 4 | 8 | 0 |
| Saurornithoididae | 1 | 6 | 3 |
| Dromaeosauridae | 0 | 0 | 0 |

where, $H_{S_{upper}}$ is the biodiversity of the upper time period measured by Shannon index $H_S = -\sum_{i=1}^{S} \hat{\pi}_i log(\hat{\pi}_i)$, and similarly, $H_{S_{middle}}$ and $H_{S_{lower}}$ are for the middle and lower time period. Shannon entropy was used to define the biodiversity and utilized in ANOVA and post-hoc test to analyze the dinosaur data and rejected the hypothesis that "the dinosaurian part of the ecosystem was deteriorating during the latest Cretaceous" (Sheehan et al., 1991). Fritsch and Hsu (1999) argued that Sheehan et al. (1991) misinterpret the large p-value and suggested that accepting null hypothesis may be caused by insufficient data. Instead they proposed to construct equivalence confidence intervals for the difference between two Shannon indices from two time periods (Fritsch and Hsu, 1999). For example:

$$H_0 : |H_{S_i} - H_{S_j}| > \delta \ for \ some \ i \neq j$$

$$H_a : |H_{S_i} - H_{S_j}| \leq \delta \ for \ all \ i \neq j$$

$\delta(> 0)$ is a predetermined limit to control the difference. Then, the bootstrap-t techniques were applied to determine confidence intervals.

However, Shannon index and Gini-Simpson index are based upon abundance of the species only and they do not take differences in the species into consideration. In the other hand, quadratic entropy (QE), as stated in Izsák and Papp (2000), "is the only ecological diversity index, the value of which reflects both the differences and abundances of the species." In this chapter, using the same dinosaur data, we show how one can analyze data for determining biodiversity using Rao's quadratic entropy. First, we will introduce Rao's quadratic entropy and its sampling distribution in Section III.1. In Section III.2 we will provide various confidence intervals for the entropy function and provide simulation results to show which of these intervals is the best. In Section III.4 we will provide confidence intervals for difference between the entropy's and once again provide simulation results. Finally, we will provide an analysis of dinosaurs data in Section III.4.

## III.1 QUADRATIC ENTROPY AND ITS SAMPLING DISTRIBUTION

Suppose in a biological community there are $s$ species and let $\Pi = (\pi_1, ..., \pi_s)'$ be the vector of proportions of these species in the community. A general diversity measure called Rao's quadratic entropy (QE) can be defined as (Rao 1982a,b,c):

$$H_Q = H_Q(\Pi) = \sum\sum d_{ij}\pi_i\pi_j = \Pi'\Delta\Pi, \tag{III.1.1}$$

where $\Delta = (d_{ij})$ and $d_{ij}$ is a nonnegative number representing the difference between the categories $i$ and $j$, so that $H_Q$ is the average difference between two individuals drawn at random from a population. In the special case, when $d_{ij} = 1$ if $i \neq j$ and $d_{ii} = 0$, that is, $\Delta = J_s - I_s$, where $J_s$ is an $s \times s$ matrix of all ones and $I_s$ is an $s \times s$ identity matrix, $H_Q = 1 - \sum \pi_i^2 = H_G$, which is the Gini-Simpson entropy function. QE can also be used to construct analysis of variance for categorical data, where the total diversity can be decomposed into diversities between and within populations (Nayak 1986a,b). This analysis has found some interesting applications in economics (Nayak and Gastwirth 1989).

Let $n_1, n_2, ..., n_s$ be the abundance of each species in a sample of size $N = \sum n_i$. Then, assuming multinomial probability model, we get the maximum likelihood estimate of $\Pi$ as $\hat{\Pi} = (\hat{\pi}_1, ..., \hat{\pi}_s)'$, where $\hat{\pi}_i = n_i/N$, $i = 1, ..., s$. Note that $E(\hat{\Pi}) = \Pi$, and $Var(\hat{\Pi}) = \frac{1}{N}[diag(\Pi) - \Pi\Pi'] = \frac{1}{N}V$. Here, $diag(\Pi)$ is the diagonal matrix with the elements of $\Pi$ as its diagonal elements. Let an estimate of $V$ be $\hat{V} = diag(\hat{\Pi}) - \hat{\Pi}\hat{\Pi}'$. Also, by the standard asymptotic theory, we have,

$$\hat{\Pi} \approx N_s(\Pi, \frac{1}{N}V).$$

Then, the maximum likelihood estimate of $H_Q$ is $\hat{H}_Q = \hat{\Pi}'\Delta\hat{\Pi}$. Also,

$$E(\hat{H}_Q) = tr(\Delta\frac{1}{N}V) + \Pi'\Delta\Pi = tr(\Delta \times \frac{1}{N}[diag(\Pi) - \Pi\Pi']) + H_Q = \frac{N-1}{N}H_Q,$$

and

$$Var(\hat{H}_Q) = \frac{1}{N^2}2tr(\Delta V)^2 + \frac{1}{N}4\Pi'\Delta V\Delta\Pi.$$

Then, by the delta theorem, as $N \rightarrow \infty$, we have,

$$\frac{N}{N-1}\hat{H}_Q \approx N(H_Q, \frac{1}{N}[\frac{2tr(\Delta V)^2}{N} + 4\Pi'\Delta V\Delta\Pi]). \tag{III.1.2}$$

For the Gini-Simpson index $H_G$, we have

$$\frac{N}{N-1}\hat{H}_G \approx N(H_G, \frac{1}{N}[\frac{2tr(VV)}{N} + 4\Pi' V\Pi]).$$

## III.2 INTERVAL ESTIMATION OF QUADRATIC ENTROPY

In this section, we provide various ways of constructing confidence intervals for $H_Q$. The first of which is based upon the asymptotic distribution of $\hat{H}_Q$. Secondly, we propose a variance stabilizing transformation and a method for constructing confidence intervals. Further, we will use various bootstrap based methods and compare all the methods using simulation.

### III.2.1 Confidence Interval Estimation

**Normal Confidence Intervals**

Based upon the asymptotic distribution of $\hat{H}_Q$ given in (III.1.2), one can provide an approximate confidence interval for $H_Q$, as follows. Suppose $z_{\frac{\alpha}{2}}$ is the standard normal upper $\alpha/2$ probability cutoff point and $L_1$ and $U_1$, respectively, are the lower and upper $100(1-\alpha)\%$ confidence limits for $H_Q$, then

$$L_1 = \frac{N}{N-1}\hat{H}_Q - z_{\alpha/2}\hat{\sigma}/\sqrt{N},$$

and

$$U_1 = \frac{N}{N-1}\hat{H}_Q + z_{\alpha/2}\hat{\sigma}/\sqrt{N}, \text{ where}$$

$$\hat{\sigma}^2 = \frac{1}{N}2tr(\Delta\hat{V})^2 + 4\hat{\Pi}'\Delta\hat{V}\Delta\hat{\Pi}. \qquad \text{(III.2.1)}$$

Our simulations have shown that the distribution of the sample entropy, although more closely centered around 0, does not agree well with the standard normal distribution in the tail regions.

**Variance Stabilizing Transformed Intervals**

Given a certain distance matrix $\Delta$, the QE reaches its minimum value when there is only one family in the community and reaches its maximum value at a certain diversity distribution. The maximum value can be calculated with an algorithm for certain choices of dissimilarity matrix (Pavoine, Ollier, and Pontier 2005). With the maximum value known, we can define a ratio index as

$$I(\Pi) = \frac{H_Q}{maxH_Q}.$$

Note that $I(\Pi)$ takes values between 0 and 1. If we estimate $I(\Pi)$ by $I(\hat{\Pi})$, then its asymptotic distribution is given by

$$I(\hat{\Pi}) \rightarrow N(I(\Pi), \frac{1}{N} I(\Pi)(1 - I(\Pi))).$$

By applying the usual variance stabilizing transformation for a binomial proportion, we get

$$arcsin\sqrt{I(\hat{\Pi})} \rightarrow N(arcsin\sqrt{I(\Pi)}, \frac{1}{4N})$$

or, in other form,

$$\sqrt{N}(arcsin(\sqrt{I(\hat{\Pi})}) - arcsin(\sqrt{I(\Pi)})) \rightarrow N(0, \frac{1}{4}).$$

Then, the $100(1-\alpha)$ confidence interval for $H_Q$ is given by

$$L_2 = maxH_Q \times sin^2[sin^{-1}\sqrt{I(\hat{\Pi})} - z_{\alpha/2}\frac{1}{\sqrt{4N}}]$$
$$U_2 = maxH_Q \times sin^2[sin^{-1}\sqrt{I(\hat{\Pi})} + z_{\alpha/2}\frac{1}{\sqrt{4N}}].$$

**Bootstrap-t Confidence Intervals**

By applying the bootstrap-t techniques (Efron and Tibshirani 1993) on the test statistic $\frac{\frac{N}{N-1}\hat{H}_Q - H_Q}{\hat{\sigma}/\sqrt{N}}$, we get a bootstrap value of the test statistic $\frac{\frac{N}{N-1}\hat{H}^*_Q - \hat{H}_Q}{\hat{\sigma}^*/\sqrt{N}}$ for each $B$ bootstrap samples. Here, $\hat{H}^*_Q$ is the entropy computed from a typical bootstrap sample and $\hat{\sigma}^*$ is its corresponding standard deviation estimate. So, the bootstrap lower and upper $100(1-\alpha)$ confidence limits are

$$L_b = \frac{N}{N-1}\hat{H}_Q - q^b_{1-\alpha/2}\hat{\sigma}/\sqrt{N}$$
$$U_b = \frac{N}{N-1}\hat{H}_Q - q^b_{\alpha/2}\hat{\sigma}/\sqrt{N},$$

where $q^b_{\alpha/2}$ and $q^b_{1-\alpha/2}$ are the $\lfloor B(\alpha/2) \rfloor + 1$ and $\lfloor B(1-\alpha/2) \rfloor + 1$ order statistics of the $B$ bootstrap quantiles and $\lfloor \cdot \rfloor$ is the greatest integer function.

This method and the following two bootstrap confidence intervals can also be applied to the variance stabilization transformation.

**Bootstrap Percentile Confidence Intervals**

Suppose $\hat{H}_Q^{*1},...,\hat{H}_Q^{*B}$ are $B$ bootstrap estimates of $H_Q$. Then, the lower and upper $100(1-\alpha)$ confidence limits are the $\lfloor B(\alpha/2)\rfloor$ and $\lfloor B(1-\alpha/2)\rfloor$ order statistics of the $B$ ordered values of $\hat{H}_Q^{*i}$.

**Bootstrap $BC_\alpha$ Confidence Intervals**

Let $\hat{H}_Q^{*(\gamma)}$ be the $100\gamma$ percentile of $\hat{H}_Q^{*1},...,\hat{H}_Q^{*B}$, then

$$\hat{z}_0 = \Phi^{-1}(\frac{\sharp\{\hat{H}_Q^{*i} < \hat{H}_Q\}}{B}),$$

where $\Phi$ is the standard normal distribution function. Then, the $100(1-\alpha)$ confidence intervals for H are $\hat{H}_Q^{*(\alpha_1)}$ and $\hat{H}_Q^{*(\alpha_2)}$, where

$$\alpha_1 = \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z_\alpha/2}{1-\hat{\alpha}/2(\hat{z}_0 + z_\alpha/2)}),$$

$$\alpha_2 = \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z_{(1-\alpha/2)}}{1-\hat{\alpha}/2(\hat{z}_0 + z_{(1-\alpha/2)})}),$$

and $\hat{\alpha}$ is the sample entropy computed without the $i$th observation, which is calculated from $\hat{H}_Q^{(i)}$ as

$$\hat{\alpha} = \frac{\sum_{i=1}^{N}(\hat{H}_Q^{(\cdot)} - \hat{H}_Q^{(i)})^3}{6\{\sum_{i=1}^{N}(\hat{H}_Q^{(\cdot)} - \hat{H}_Q^{(i)})^2\}^{1.5}}.$$

Here $\hat{H}_Q^{(\cdot)}$is the average of $\hat{H}_Q^{(i)}$.

### III.2.2 Selection of Difference Matrices

One of the first steps in computing QE is to identify an appropriate $\Delta$, the distance matrix. If we assume the distance between each pair of the eight dinosaur families is the same, then $\Delta$ is as given below. As noted earlier, in this case, the QE is same as the Gini-Simpson index.

TABLE 8. Diets of dinosaur families

| Family | Dietary |
|---|---|
| Ceratopsidae | Herbivores |
| Hadrosauridae | Herbivores |
| Hypsilophodontidae | Herbivores |
| Pachycephalosauridae | Herbivores/Omnivorous |
| Tyrannosauridae | Carnivorous |
| Ornithominidae | Omnivorous/Herbivorous |
| Saurornithoididae | Carnivorous |
| Dromaeosauridae | Carnivorous |

$$\Delta_0 = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

If dinosaur families with similar food chains have similar ecological characteristics, then they can be assigned relatively a shorter distance than those who do not. Using the dietary information about different dinosaur families given in Table 8, based upon Norman (1991), we can suggest the following $\Delta_1$ as an appropriate distance.

$$\Delta_1 = \begin{bmatrix} 0 & 1 & 1 & 2 & 3 & 2 & 3 & 3 \\ 1 & 0 & 1 & 2 & 3 & 2 & 3 & 3 \\ 1 & 1 & 0 & 2 & 3 & 2 & 3 & 3 \\ 2 & 2 & 2 & 0 & 2 & 1 & 2 & 2 \\ 3 & 3 & 3 & 2 & 0 & 2 & 1 & 1 \\ 2 & 2 & 2 & 1 & 2 & 0 & 2 & 2 \\ 3 & 3 & 3 & 2 & 1 & 2 & 0 & 1 \\ 3 & 3 & 3 & 2 & 1 & 2 & 1 & 0 \end{bmatrix}$$

We have experimented with different ways of finding $\Delta$ using data and have proposed the following:

$\Delta_2 = (d_{ij})$, where

$$d_{ij} = \begin{cases} 0 & if \qquad i = j \\ 1 & if \qquad \hat{\pi}_i = \hat{\pi}_j = 0 \\ |log(\hat{\pi}_i)| + 1 & if \qquad \hat{\pi}_j = 0 \\ |log(\hat{\pi}_j)| + 1 & if \qquad \hat{\pi}_i = 0 \\ |log(\hat{\pi}_i) - log(\hat{\pi}_j)| + 1 & otherwise. \end{cases} \tag{III.2.2}$$

### III.2.3 Comparisons of Confidence Intervals Based upon Simulated Data

The accuracy of the confidence intervals were compared by the noncoverage probability with simulated data. Data were generated from a multinomial (50, $\Pi$) distribution, where $\Pi$ is defined by various geometric models with $s = 8$. Tables 9, 10, 11 list the percentage of confidence bounds that fail to bound the true entropy value. "Below" and "Above" represent the probability that the true entropy value falls below the lower limit, i.e. $P(H < L)$ and the true entropy value falls above the upper limit, i.e. $P(H > U)$. The comparisons of these two probabilities with a true significance level $\alpha$ will show the performance of the confidence intervals. The noncoverage probabilities were based upon 5,000 simulated data with 7,500 bootstrap samples.

When $k = 0.8$, QE and the indices based upon QE have better simulated results than Shannon entropy. When comparing confidence intervals based upon normal distributions with those based upon bootstrap-t, bootstrap percentiles, bias adjusted and bias-corrected and accelerated ($BC_\alpha$) techniques, the $BC_\alpha$ confidence intervals appear to be the most liberal. Confidence bounds based upon bootstrap-t come closest to the desired $1 - 2\alpha$, when compared with normal intervals and other bootstrap intervals, however, the bootstrap-t interval based upon Shannon entropy and QE with $\Delta_0$ and the corresponding indices have some imbalance in that the lower bound appears to be conservative, whereas its upper bound appears to be liberal; and the imbalance is not found in confidence intervals based upon other quadratic entropies and indices. The closest coverage is reached by using the QE with $\Delta_2$ bootstrap-t confidence intervals; the QE index with $\Delta_2$ and the QE with $\Delta_0$ produce the next closest coverage. Similar results are also observed in the case of $k = 0.6$.

## TABLE 9. Noncoverage probability for single entropy with k=0.8

| $\alpha$ | Method | Shannon | | $QE_0$ | | $QE_1$ | | $QE_2$ | | $i\_QE_0$ | | $i\_QE_1$ | | $i\_QE_2$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above |
| 0.05 | Normal | 0.0608 | 0.1112 | 0.0668 | 0.0666 | 0.0906 | 0.0392 | 0.0916 | 0.0386 | 0.1094 | 0.0768 | 0.0442 | 0.0170 | 0.0292 | 0.0078 |
| | Bootstrap t | 0.1070 | 0.0320 | 0.0164 | 0.0688 | 0.0658 | 0.0532 | 0.0320 | 0.0520 | 0.0596 | 0.0952 | 0.0670 | 0.0748 | 0.0328 | 0.0776 |
| | *Percentile* | 0.1816 | 0.0148 | 0.0876 | 0.0352 | 0.1058 | 0.0294 | 0.1098 | 0.0320 | 0.0876 | 0.0352 | 0.1058 | 0.0294 | 0.1098 | 0.0320 |
| | Bias Adjusted | 0.0914 | 0.0650 | 0.0668 | 0.0580 | 0.1058 | 0.0364 | 0.1062 | 0.0324 | N/A | N/A | N/A | N/A | N/A | N/A |
| | $BC_a$ | 0.3312 | 0.0604 | 0.1600 | 0.0658 | 0.1786 | 0.0624 | 0.1766 | 0.0750 | 0.1600 | 0.0658 | 0.1786 | 0.0632 | 0.1768 | 0.0752 |
| 0.025 | Normal | 0.0402 | 0.0628 | 0.0470 | 0.0324 | 0.0532 | 0.0158 | 0.0682 | 0.0152 | 0.0958 | 0.0394 | 0.0158 | 0.0058 | 0.0146 | 0.0018 |
| | Bootstrap t | 0.0542 | 0.0122 | 0.0064 | 0.0292 | 0.0164 | 0.0254 | 0.0102 | 0.0220 | 0.0184 | 0.0592 | 0.0168 | 0.0426 | 0.0110 | 0.0408 |
| | Percentile | 0.1302 | 0.0056 | 0.0472 | 0.0160 | 0.0630 | 0.0110 | 0.0684 | 0.0120 | 0.0472 | 0.0160 | 0.0636 | 0.0110 | 0.0684 | 0.0120 |
| | Bias Adjusted | 0.0602 | 0.0240 | 0.0440 | 0.0240 | 0.0532 | 0.0138 | 0.0676 | 0.0120 | N/A | N/A | N/A | N/A | N/A | N/A |
| | $BC_a$ | 0.2252 | 0.0282 | 0.1120 | 0.0296 | 0.1258 | 0.0360 | 0.1088 | 0.0382 | 0.1126 | 0.0296 | 0.1262 | 0.0360 | 0.1094 | 0.0382 |
| 0.01 | Normal | 0.0216 | 0.0344 | 0.0216 | 0.0214 | 0.0398 | 0.0054 | 0.0422 | 0.0058 | 0.0546 | 0.0176 | 0.0106 | 0.0016 | 0.0046 | 0.0006 |
| | Bootstrap t | 0.0252 | 0.0042 | 0.0058 | 0.0136 | 0.0188 | 0.0116 | 0.0102 | 0.0106 | 0.0216 | 0.0328 | 0.0194 | 0.0252 | 0.0110 | 0.0276 |
| | Percentile | 0.0934 | 0.0010 | 0.0288 | 0.0050 | 0.0444 | 0.0026 | 0.0416 | 0.0026 | 0.0288 | 0.0050 | 0.0444 | 0.0026 | 0.0416 | 0.0026 |
| | Bias Adjusted | 0.0370 | 0.0102 | 0.0288 | 0.0072 | 0.0436 | 0.0042 | 0.0416 | 0.0028 | N/A | N/A | N/A | N/A | N/A | N/A |
| | $BC_a$ | 0.0188 | 0.0146 | 0.0722 | 0.0132 | 0.1028 | 0.0194 | 0.0926 | 0.0202 | 0.0724 | 0.0128 | 0.1028 | 0.0196 | 0.0942 | 0.0204 |
| 0.005 | Normal | 0.0106 | 0.0254 | 0.0184 | 0.0172 | 0.0162 | 0.0054 | 0.0252 | 0.0044 | 0.0486 | 0.0106 | 0.0052 | 0.0006 | 0.0040 | 0.0000 |
| | Bootstrap t | 0.0226 | 0.0016 | 0.0056 | 0.0104 | 0.0162 | 0.0086 | 0.0114 | 0.0054 | 0.0184 | 0.0362 | 0.0166 | 0.0240 | 0.0118 | 0.0202 |
| | Percentile | 0.0796 | 0.0002 | 0.0174 | 0.0032 | 0.0262 | 0.0016 | 0.0288 | 0.0018 | 0.0174 | 0.0032 | 0.0264 | 0.0016 | 0.0298 | 0.0016 |
| | Bias Adjusted | 0.0280 | 0.0062 | 0.0136 | 0.0062 | 0.0262 | 0.0028 | 0.0284 | 0.0018 | N/A | N/A | N/A | N/A | N/A | N/A |
| | $BC_a$ | 0.1370 | 0.0104 | 0.0616 | 0.0090 | 0.0636 | 0.0144 | 0.0658 | 0.0150 | 0.0616 | 0.0090 | 0.0640 | 0.0144 | 0.0658 | 0.0150 |

## TABLE 10. Noncoverage probability for single entropy with k=0.6

| $\alpha$ | Method | Shannon |  | $QE_0$ |  | $QE_1$ |  | $QE_2$ |  | $i\_QE_0$ |  | $i\_QE_1$ |  | $i\_QE_2$ |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above |
| 0.05 | Normal | 0.0722 | 0.0832 | 0.0390 | 0.0938 | 0.1306 | 0.0240 | 0.0984 | 0.0316 | 0.0872 | 0.0258 | 0.0568 | 0.0394 | 0.0324 | 0.0136 |
|  | Bootstrap t | 0.0794 | 0.0362 | 0.0236 | 0.1362 | 0.0826 | 0.0410 | 0.0630 | 0.0482 | 0.0342 | 0.0996 | 0.1020 | 0.0376 | 0.0670 | 0.0526 |
|  | Percentile | 0.2534 | 0.0046 | 0.0972 | 0.0262 | 0.1448 | 0.0176 | 0.1332 | 0.0198 | 0.0972 | 0.0262 | 0.1448 | 0.0176 | 0.1332 | 0.0198 |
|  | Bias Adjusted | 0.1136 | 0.0302 | 0.0666 | 0.0450 | 0.1270 | 0.0246 | 0.1044 | 0.0308 | N/A | N/A | N/A | N/A | N/A | N/A |
|  | $BC_a$ | 0.4522 | 0.0566 | 0.1416 | 0.0722 | 0.2344 | 0.0510 | 0.2098 | 0.0570 | 0.1422 | 0.0722 | 0.2358 | 0.0508 | 0.2100 | 0.0572 |
| 0.025 | Normal | 0.0440 | 0.0558 | 0.0168 | 0.0692 | 0.0902 | 0.1118 | 0.0636 | 0.0152 | 0.0536 | 0.0128 | 0.0276 | 0.0198 | 0.0156 | 0.0054 |
|  | Bootstrap t | 0.0432 | 0.0186 | 0.0094 | 0.0932 | 0.0564 | 0.0172 | 0.0360 | 0.0252 | 0.0096 | 0.0734 | 0.0640 | 0.0164 | 0.0358 | 0.0306 |
|  | Percentile | 0.1738 | 0.0030 | 0.0526 | 0.0164 | 0.1062 | 0.0090 | 0.0910 | 0.0100 | 0.0528 | 0.0162 | 0.1064 | 0.0090 | 0.0912 | 0.0098 |
|  | Bias Adjusted | 0.0752 | 0.0148 | 0.0370 | 0.0258 | 0.0878 | 0.0118 | 0.0710 | 0.0142 | N/A | N/A | N/A | N/A | N/A | N/A |
|  | $BC_a$ | 0.3536 | 0.0346 | 0.0840 | 0.0484 | 0.1832 | 0.0282 | 0.1522 | 0.0346 | 0.0844 | 0.0484 | 0.1836 | 0.0286 | 0.1530 | 0.0346 |
| 0.01 | Normal | 0.0240 | 0.0244 | 0.0084 | 0.0338 | 0.0620 | 0.0028 | 0.0368 | 0.0044 | 0.0292 | 0.0026 | 0.0106 | 0.0086 | 0.0046 | 0.0018 |
|  | Bootstrap t | 0.0246 | 0.0046 | 0.0020 | 0.0494 | 0.0394 | 0.0046 | 0.0176 | 0.0082 | 0.0046 | 0.0428 | 0.0334 | 0.0048 | 0.0134 | 0.0124 |
|  | Percentile | 0.1204 | 0.0006 | 0.0278 | 0.0042 | 0.0702 | 0.0016 | 0.0576 | 0.0020 | 0.0278 | 0.0042 | 0.0702 | 0.0016 | 0.0576 | 0.0020 |
|  | Bias Adjusted | 0.0486 | 0.0042 | 0.0192 | 0.0084 | 0.0580 | 0.0024 | 0.0452 | 0.0034 | N/A | N/A | N/A | N/A | N/A | N/A |
|  | $BC_a$ | 0.2604 | 0.0154 | 0.0488 | 0.0196 | 0.1296 | 0.0120 | 0.1026 | 0.0136 | 0.0492 | 0.0196 | 0.1298 | 0.0120 | 0.1038 | 0.0136 |
| 0.005 | Normal | 0.0158 | 0.0118 | 0.0042 | 0.0232 | 0.0500 | 0.0008 | 0.0294 | 0.0020 | 0.0172 | 0.0002 | 0.0062 | 0.0030 | 0.0028 | 0.0002 |
|  | Bootstrap t | 0.0132 | 0.0016 | 0.0002 | 0.0320 | 0.0388 | 0.0006 | 0.0122 | 0.0040 | 0.0012 | 0.0324 | 0.0250 | 0.0020 | 0.0064 | 0.0066 |
|  | Percentile | 0.0958 | 0.0000 | 0.0170 | 0.0016 | 0.0564 | 0.0006 | 0.0490 | 0.0008 | 0.0172 | 0.0016 | 0.0568 | 0.0006 | 0.0490 | 0.0008 |
|  | Bias Adjusted | 0.0396 | 0.0012 | 0.0118 | 0.0026 | 0.0480 | 0.0008 | 0.0360 | 0.0012 | N/A | N/A | N/A | N/A | N/A | N/A |
|  | $BC_a$ | 0.2176 | 0.0082 | 0.0318 | 0.0112 | 0.1150 | 0.0044 | 0.0870 | 0.0084 | 0.0322 | 0.0112 | 0.1158 | 0.0044 | 0.0872 | 0.0084 |

TABLE 11. Noncoverage probability for single entropy with k=0.4

| $\alpha$ | Method | Shannon | | $QE_0$ | | $QE_1$ | | $QE_2$ | | i_$QE_0$ | | i_$QE_1$ | | i_$QE_2$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above |
| 0.05 | Normal | 0.0828 | 0.0748 | 0.0244 | 0.1018 | 0.0878 | 0.0392 | 0.1038 | 0.0304 | 0.0602 | 0.0054 | 0.1050 | 0.0646 | 0.0406 | 0.0192 |
| | Bootstrap t | 0.0822 | 0.0414 | 0.0122 | 0.2612 | 0.0328 | 0.0392 | 0.0560 | 0.0454 | 0.0384 | 0.0996 | 0.0784 | 0.0500 | 0.0796 | 0.0472 |
| | Percentile | 0.3616 | 0.0020 | 0.1540 | 0.0062 | 0.1138 | 0.0204 | 0.1522 | 0.0146 | 0.1540 | 0.0062 | 0.1138 | 0.0204 | 0.1522 | 0.0146 |
| | Bias Adjusted | 0.1734 | 0.0102 | 0.0768 | 0.0238 | 0.0858 | 0.0300 | 0.1096 | 0.0256 | N/A | N/A | N/A | N/A | N/A | N/A |
| | $BC_a$ | 0.7600 | 0.0626 | 0.3112 | 0.0838 | 0.1924 | 0.0512 | 0.2592 | 0.0526 | 0.3126 | 0.0842 | 0.1924 | 0.0512 | 0.2592 | 0.0524 |
| 0.025 | Normal | 0.0528 | 0.0418 | 0.0130 | 0.0646 | 0.0596 | 0.0190 | 0.0696 | 0.0124 | 0.0308 | 0.0024 | 0.0674 | 0.0314 | 0.0210 | 0.0064 |
| | Bootstrap t | 0.0438 | 0.0172 | 0.0058 | 0.1892 | 0.0142 | 0.0132 | 0.0356 | 0.0192 | 0.0160 | 0.0644 | 0.0560 | 0.0236 | 0.0546 | 0.0218 |
| | Percentile | 0.2834 | 0.0004 | 0.0960 | 0.0038 | 0.0794 | 0.0076 | 0.1016 | 0.0050 | 0.0964 | 0.0038 | 0.0796 | 0.0076 | 0.1018 | 0.0050 |
| | Bias Adjusted | 0.1198 | 0.0040 | 0.0450 | 0.0110 | 0.0586 | 0.0128 | 0.0750 | 0.0086 | N/A | N/A | N/A | N/A | N/A | N/A |
| | $BC_a$ | 0.6500 | 0.0316 | 0.2044 | 0.0510 | 0.1480 | 0.0238 | 0.2004 | 0.0274 | 0.2054 | 0.0514 | 0.1480 | 0.0238 | 0.2004 | 0.0274 |
| 0.01 | Normal | 0.0248 | 0.0240 | 0.0022 | 0.0430 | 0.0316 | 0.0084 | 0.0404 | 0.0062 | 0.0124 | 0.0000 | 0.0316 | 0.0176 | 0.0054 | 0.0022 |
| | Bootstrap t | 0.0200 | 0.0070 | 0.0040 | 0.1318 | 0.0036 | 0.0036 | 0.0154 | 0.0082 | 0.0024 | 0.0424 | 0.0304 | 0.0112 | 0.0252 | 0.0112 |
| | Percentile | 0.2000 | 0.0000 | 0.0510 | 0.0002 | 0.0416 | 0.0074 | 0.0700 | 0.0014 | 0.0510 | 0.0002 | 0.0416 | 0.0040 | 0.0700 | 0.0014 |
| | Bias Adjusted | 0.0796 | 0.0004 | 0.0204 | 0.0016 | 0.0292 | 0.0054 | 0.0474 | 0.0038 | N/A | N/A | N/A | N/A | N/A | N/A |
| | $BC_a$ | 0.5000 | 0.0164 | 0.1212 | 0.0290 | 0.0978 | 0.0108 | 0.1426 | 0.0132 | 0.1218 | 0.0292 | 0.0980 | 0.0106 | 0.1428 | 0.0130 |
| 0.005 | Normal | 0.0190 | 0.0138 | 0.0014 | 0.0270 | 0.0262 | 0.0056 | 0.0300 | 0.0028 | 0.0080 | 0.0000 | 0.0226 | 0.0092 | 0.0026 | 0.0002 |
| | Bootstrap t | 0.0086 | 0.0012 | 0.0002 | 0.0934 | 0.0006 | 0.0002 | 0.0116 | 0.0048 | 0.0020 | 0.0284 | 0.0270 | 0.0076 | 0.0196 | 0.0058 |
| | Percentile | 0.1546 | 0.0000 | 0.0306 | 0.0000 | 0.0318 | 0.0022 | 0.0545 | 0.0004 | 0.0312 | 0.0000 | 0.0324 | 0.0022 | 0.0554 | 0.0004 |
| | Bias Adjusted | 0.0626 | 0.0002 | 0.0132 | 0.0002 | 0.0254 | 0.0026 | 0.0396 | 0.0012 | N/A | N/A | N/A | N/A | N/A | N/A |
| | $BC_a$ | 0.4338 | 0.0098 | 0.0766 | 0.0162 | 0.0738 | 0.0060 | 0.1118 | 0.0072 | 0.0772 | 0.0168 | 0.0744 | 0.0060 | 0.1118 | 0.0072 |

When $k = 0.4$, QEs and the indices based upon QEs continue to have better simulated results than Shannon entropy; the $BC_\alpha$ confidence intervals remains to be the most liberal. Bias adjustment improves the percentile method, especially for QE with $\Delta_0$, making its total coverage probability $1-(Below+Above)$ the closest to the desired $1-2\alpha$, when compared to other methods. The bootstrap-t confidence interval for QE with $\Delta_2$ produce next closest coverage. Imbalance is observed in the bias adjusted percentile confidence intervals for QE with $\Delta_0$ and $\Delta_2$ in that the lower bound appears to be liberal and its upper bound appears to be conservative.

Overall, the proposed QE and QE indices are more accurate than Shannon entropy in terms of coverage probability. The indices built on $\Delta_2$ with the bootstrap-t intervals exhibit the best performance when the distribution is set with larger $k$ values. As $k$ gets smaller, indices based upon QE with $\Delta_0$ will produce a better result. Among entropy based intervals, bias adjusted percentile intervals perform better than normal intervals and other bootstrap intervals. The $BC_\alpha$ confidence intervals appear to be the most liberal.

## III.3 ESTIMATION OF DIFFERENCE BETWEEN TWO QUADRATIC ENTROPIES

In this section, we will derive the confidence bounds of the differences between two quadratic entropies.

### III.3.1 Confidence Intervals of Difference between Two Quadratic Entropies

Suppose $H_{Qi}$ is the QE defined as in (III.1.1) for the $i$th population, each of which has $s$ species. Let $N_i$ be the number of observations from the $i$th population. Then, an estimate of $H_{Qi}$ as before can be obtained as $\hat{H}_{Qi}$. Let $\hat{\sigma}_i$ be the estimated standard deviation, as in (III.2.1) for the $i$th population. Then, we can provide the lower and upper limits of the $100(1-\alpha)\%$ confidence intervals for the difference $H_{Qi}-H_{Qj}$ based upon the asymptotic normal distribution as

$$L_{ij} = \frac{N_i}{N_i-1}\hat{H}_{Qi} - \frac{N_j}{N_j-1}\hat{H}_{Qj} - z_{\alpha/2}\sqrt{\hat{\sigma}_i^2/N_i + \hat{\sigma}_j^2/N_j},$$

and

$$U_{ij} = \frac{N_i}{N_i-1}\hat{H}_{Qi} - \frac{N_j}{N_j-1}\hat{H}_{Qj} + z_{\alpha/2}\sqrt{\hat{\sigma}_i^2/N_i + \hat{\sigma}_j^2/N_j}.$$

As before, we can provide bootstrap t-interval using the ordered bootstrap values,

$$\frac{\frac{N_i}{N_i-1}\hat{H}^*_{Qi} - \frac{N_j}{N_j-1}\hat{H}^*_{Qj} - (\hat{H}_{Qi} - \hat{H}_{Qi})}{\sqrt{\hat{\sigma}^2_{*i}/N_i + \hat{\sigma}^2_{*j}/N_j}}.$$

Also, the percentile based intervals can be obtained using the ordered bootstrap values

$$\frac{N_i}{N_i-1}\hat{H}^*_{Qi} - \frac{N_j}{N_j-1}\hat{H}^*_{Qj}.$$

Confidence intervals based on the index are also constructed as before.

### III.3.2 Empirical Simulation for Difference of Two Entropies

The performance of QE in constructing confidence intervals of pair differences were tested in simulated data. The data were generated from two multinomial (50, Π) distributions, where Π is defined by one of the three geometric models with $s = 8$. The noncoverage probabilities of confidence intervals were estimated based upon 5,000 sets of simulated bounds; each bootstrap bound was computed using $B = 7,500$ bootstrap samples. Tables 12, 13, 14 list the estimated noncoverage probabilities of confidence intervals of $H_{Qi} - H_{Qj}$.

If the distributions of two time intervals (groups) are similar, in other words, parameter $k$ is the same ($k_1 = 0.6$ and $k_2 = 0.6$), QE with $\Delta_0$ and the index based upon $\Delta_2$ produce more accurate results than other measures. While comparing confidence intervals based upon normal distributions with those based upon bootstrap-t, bootstrap percentiles, and $BC_\alpha$ techniques, the $BC_\alpha$ confidence intervals appear to be the most liberal. When $U_{ij}$ and $L_{ij}$ do not have coverage probabilities exactly equal to $1-\alpha$, the size of the test that rejects $H_0^{ij}$, when $[L^-_{ij}, U^+_{ij}] \subseteq [-\delta, \delta]$, is (Berger and Hsu 1996)

$$max\{sup_{H_i - H_j > \delta} P(U_{ij} < H_i - H_j), sup_{H_i - H_j < -\delta} P(L_{ij} > H_i - H_j)\}.$$

Therefore, the normal distribution bounds of the QE index based upon $\Delta_2$ appears to be the most accurate for assessing the practical equivalence of entropies because they have the smallest max{Below, Above} for all $\alpha$ values.

When the distributions of two time intervals (groups) differ, the normal confidence bounds of the QE index based upon distance $\Delta_2$ remain more accurate, because they have the smallest max{Below, Above}. The QE and QE indices based upon $\Delta_0$ have the next to the best coverage with bootstrap-t and bootstrap percentile confidence bounds. $BC_\alpha$ confidence bounds remain liberal throughout.

TABLE 12. Noncoverage probability for difference of two entropies $H_i - H_j$ with $k_1 = 0.6$ and $k_2 = 0.6$

| $\alpha$ | Method | Shannon | | $QE_0$ | | $QE_1$ | | $QE_2$ | | i.$QE_0$ | | i.$QE_1$ | | i.$QE_2$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above |
| 0.05 | Normal | 0.0776 | 0.0692 | 0.0576 | 0.0550 | 0.0624 | 0.0584 | 0.0580 | 0.0564 | 0.0554 | 0.0550 | 0.0526 | 0.0442 | 0.0212 | 0.0226 |
| | Bootstrap t | 0.0632 | 0.0590 | 0.0602 | 0.0564 | 0.0660 | 0.0592 | 0.0578 | 0.0564 | 0.0564 | 0.0524 | 0.0626 | 0.0592 | 0.0584 | 0.0576 |
| | Percentile | 0.0616 | 0.0572 | 0.0550 | 0.0506 | 0.0638 | 0.0584 | 0.0586 | 0.0558 | 0.0550 | 0.0506 | 0.0638 | 0.0582 | 0.0586 | 0.0558 |
| | $BC_a$ | 0.0480 | 0.0728 | 0.0482 | 0.0598 | 0.0506 | 0.0712 | 0.0456 | 0.0668 | 0.0472 | 0.0598 | 0.0536 | 0.0656 | 0.0504 | 0.0642 |
| 0.025 | Normal | 0.0468 | 0.0382 | 0.0296 | 0.0272 | 0.0314 | 0.0280 | 0.0342 | 0.0290 | 0.0290 | 0.0288 | 0.0232 | 0.0202 | 0.0094 | 0.0086 |
| | Bootstrap t | 0.0328 | 0.0294 | 0.0304 | 0.0276 | 0.0354 | 0.0316 | 0.0324 | 0.0284 | 0.0270 | 0.0246 | 0.0330 | 0.0294 | 0.0342 | 0.0280 |
| | Percentile | 0.0340 | 0.0306 | 0.0272 | 0.0252 | 0.0334 | 0.0300 | 0.0342 | 0.0276 | 0.0272 | 0.0250 | 0.0332 | 0.0300 | 0.0342 | 0.0276 |
| | $BC_a$ | 0.0270 | 0.0396 | 0.0234 | 0.0316 | 0.0214 | 0.0404 | 0.0262 | 0.0356 | 0.0232 | 0.0318 | 0.0238 | 0.0386 | 0.0308 | 0.0326 |
| 0.01 | Normal | 0.0266 | 0.0206 | 0.0126 | 0.0118 | 0.0148 | 0.0120 | 0.0134 | 0.0098 | 0.0108 | 0.0120 | 0.0098 | 0.0096 | 0.0022 | 0.0022 |
| | Bootstrap t | 0.0178 | 0.0122 | 0.0122 | 0.0108 | 0.0176 | 0.0142 | 0.0138 | 0.0108 | 0.0098 | 0.0088 | 0.0152 | 0.0120 | 0.0130 | 0.0092 |
| | Percentile | 0.0178 | 0.0142 | 0.0112 | 0.0110 | 0.0172 | 0.0140 | 0.0148 | 0.0110 | 0.0112 | 0.0108 | 0.0172 | 0.0140 | 0.0148 | 0.0110 |
| | $BC_a$ | 0.0118 | 0.0192 | 0.0092 | 0.0144 | 0.0106 | 0.0218 | 0.0088 | 0.0166 | 0.0088 | 0.0142 | 0.0122 | 0.0196 | 0.0110 | 0.0146 |
| 0.005 | Normal | 0.0124 | 0.0138 | 0.0058 | 0.0064 | 0.0056 | 0.0068 | 0.0070 | 0.0072 | 0.0084 | 0.0060 | 0.0070 | 0.0040 | 0.0012 | 0.0004 |
| | Bootstrap t | 0.0066 | 0.0076 | 0.0046 | 0.0060 | 0.0078 | 0.0084 | 0.0070 | 0.0066 | 0.0040 | 0.0050 | 0.0060 | 0.0066 | 0.0068 | 0.0074 |
| | Percentile | 0.0074 | 0.0090 | 0.0052 | 0.0058 | 0.0080 | 0.0082 | 0.0074 | 0.0070 | 0.0052 | 0.0058 | 0.0080 | 0.0082 | 0.0074 | 0.0070 |
| | $BC_a$ | 0.0056 | 0.0126 | 0.0046 | 0.0074 | 0.0032 | 0.0148 | 0.0044 | 0.0118 | 0.0046 | 0.0078 | 0.0044 | 0.0130 | 0.0056 | 0.0102 |

TABLE 13. Noncoverage probability for difference of two entropies $H_i - H_j$ with $k_1 = 0.6$ and $k_2 = 0.4$

| $\alpha$ | Method | Shannon | | $QE_0$ | | $QE_1$ | | $QE_2$ | | i.$QE_0$ | | i.$QE_1$ | | i.$QE_2$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above |
| 0.05 | Normal | 0.0672 | 0.0908 | 0.0520 | 0.0658 | 0.0576 | 0.0694 | 0.0664 | 0.0504 | 0.0416 | 0.0454 | 0.0602 | 0.0800 | 0.0334 | 0.0364 |
| | Bootstrap t | 0.0688 | 0.0616 | 0.0630 | 0.0618 | 0.0528 | 0.0560 | 0.0600 | 0.0546 | 0.0558 | 0.0668 | 0.0612 | 0.0706 | 0.0676 | 0.0518 |
| | Percentile | 0.0422 | 0.0926 | 0.0428 | 0.0670 | 0.0476 | 0.0742 | 0.0644 | 0.0526 | 0.0394 | 0.0754 | 0.0514 | 0.0728 | 0.0650 | 0.0516 |
| | $BC_a$ | 0.0198 | 0.0774 | 0.0302 | 0.0620 | 0.0356 | 0.0722 | 0.0588 | 0.0650 | 0.0184 | 0.0654 | 0.0398 | 0.0712 | 0.0614 | 0.0642 |
| 0.025 | Normal | 0.0336 | 0.0496 | 0.0218 | 0.0372 | 0.0298 | 0.0322 | 0.0372 | 0.0268 | 0.0188 | 0.0260 | 0.0304 | 0.0360 | 0.0138 | 0.0182 |
| | Bootstrap t | 0.0342 | 0.0312 | 0.0258 | 0.0402 | 0.0238 | 0.0254 | 0.0340 | 0.0294 | 0.0214 | 0.0392 | 0.0324 | 0.0326 | 0.0372 | 0.0252 |
| | Percentile | 0.0206 | 0.0484 | 0.0180 | 0.0370 | 0.0212 | 0.0356 | 0.0372 | 0.0284 | 0.0150 | 0.0428 | 0.0234 | 0.0344 | 0.0372 | 0.0278 |
| | $BC_a$ | 0.0076 | 0.0402 | 0.0134 | 0.0360 | 0.0128 | 0.0338 | 0.0320 | 0.0376 | 0.0090 | 0.0396 | 0.0162 | 0.0324 | 0.0334 | 0.0372 |
| 0.01 | Normal | 0.0216 | 0.0268 | 0.0104 | 0.0150 | 0.0192 | 0.0134 | 0.0186 | 0.0082 | 0.0084 | 0.0072 | 0.0180 | 0.0152 | 0.0050 | 0.0050 |
| | Bootstrap t | 0.0198 | 0.0126 | 0.0100 | 0.0184 | 0.0132 | 0.0098 | 0.0164 | 0.0094 | 0.0094 | 0.0148 | 0.0182 | 0.0142 | 0.0164 | 0.0074 |
| | Percentile | 0.0116 | 0.0232 | 0.0096 | 0.0146 | 0.0136 | 0.0162 | 0.0186 | 0.0096 | 0.0082 | 0.0152 | 0.0144 | 0.0150 | 0.0186 | 0.0090 |
| | $BC_a$ | 0.0036 | 0.0222 | 0.0068 | 0.0150 | 0.0092 | 0.0156 | 0.0168 | 0.0162 | 0.0040 | 0.0150 | 0.0106 | 0.0148 | 0.0172 | 0.0160 |
| 0.005 | Normal | 0.0124 | 0.0144 | 0.0050 | 0.0088 | 0.0102 | 0.0068 | 0.0122 | 0.0042 | 0.0040 | 0.0040 | 0.0086 | 0.0080 | 0.0016 | 0.0022 |
| | Bootstrap t | 0.0088 | 0.0060 | 0.0044 | 0.0110 | 0.0076 | 0.0048 | 0.0106 | 0.0044 | 0.0036 | 0.0080 | 0.0100 | 0.0080 | 0.0108 | 0.0042 |
| | Percentile | 0.0062 | 0.0116 | 0.0052 | 0.0072 | 0.0074 | 0.0074 | 0.0118 | 0.0040 | 0.0040 | 0.0076 | 0.0080 | 0.0070 | 0.0118 | 0.0036 |
| | $BC_a$ | 0.0260 | 0.0114 | 0.0040 | 0.0076 | 0.0038 | 0.0094 | 0.0106 | 0.0078 | 0.0018 | 0.0076 | 0.0050 | 0.0086 | 0.0106 | 0.0080 |

TABLE 14. Noncoverage probability for difference of two entropies $H_i - H_j$ with $k_1 = 0.6$ and $k_2 = 0.8$

| $\alpha$ | Method | Shannon | | $QE_0$ | | $QE_1$ | | $QE_2$ | | i.$QE_0$ | | i.$QE_1$ | | i.$QE_2$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above | Below | Above |
| 0.05 | Normal | 0.0942 | 0.0592 | 0.0560 | 0.0614 | 0.0872 | 0.0388 | 0.0606 | 0.0624 | 0.0830 | 0.0752 | 0.0374 | 0.0412 | 0.0296 | 0.0364 |
| | Bootstrap t | 0.0648 | 0.0618 | 0.0506 | 0.0502 | 0.0710 | 0.0514 | 0.0568 | 0.0598 | 0.0592 | 0.0460 | 0.0792 | 0.0418 | 0.0646 | 0.0510 |
| | Percentile | 0.0988 | 0.0358 | 0.0618 | 0.0456 | 0.0922 | 0.0350 | 0.0622 | 0.0582 | 0.0588 | 0.0476 | 0.0742 | 0.0442 | 0.0586 | 0.0654 |
| | $BC_a$ | 0.1208 | 0.0750 | 0.0660 | 0.0532 | 0.1106 | 0.0638 | 0.0590 | 0.0718 | 0.0600 | 0.0516 | 0.0806 | 0.0590 | 0.0548 | 0.0650 |
| 0.025 | Normal | 0.0554 | 0.0344 | 0.0274 | 0.0388 | 0.0526 | 0.0218 | 0.0286 | 0.0342 | 0.0438 | 0.0462 | 0.0150 | 0.0228 | 0.0132 | 0.0158 |
| | Bootstrap t | 0.0346 | 0.0342 | 0.0238 | 0.0240 | 0.0456 | 0.0292 | 0.0272 | 0.0298 | 0.0286 | 0.0242 | 0.0490 | 0.0226 | 0.0334 | 0.0256 |
| | Percentile | 0.0562 | 0.0228 | 0.0294 | 0.0270 | 0.0582 | 0.0192 | 0.0296 | 0.0314 | 0.0282 | 0.0288 | 0.0444 | 0.0260 | 0.0252 | 0.0356 |
| | $BC_a$ | 0.0700 | 0.0434 | 0.0324 | 0.0302 | 0.0666 | 0.0378 | 0.0242 | 0.0404 | 0.0300 | 0.0284 | 0.0464 | 0.0328 | 0.0214 | 0.0326 |
| 0.01 | Normal | 0.0284 | 0.0176 | 0.0104 | 0.0202 | 0.0294 | 0.0074 | 0.0118 | 0.0170 | 0.0222 | 0.0262 | 0.0052 | 0.0086 | 0.0022 | 0.0066 |
| | Bootstrap t | 0.0130 | 0.0166 | 0.0068 | 0.0090 | 0.0252 | 0.0092 | 0.0110 | 0.0136 | 0.0088 | 0.0118 | 0.0254 | 0.0096 | 0.0142 | 0.0118 |
| | Percentile | 0.0276 | 0.0096 | 0.0104 | 0.0112 | 0.0034 | 0.0062 | 0.0116 | 0.0132 | 0.0100 | 0.0122 | 0.0220 | 0.0110 | 0.0094 | 0.0182 |
| | $BC_a$ | 0.0322 | 0.0230 | 0.0110 | 0.0114 | 0.0372 | 0.0168 | 0.0080 | 0.0216 | 0.0110 | 0.0098 | 0.0228 | 0.0132 | 0.0076 | 0.0146 |
| 0.005 | Normal | 0.0192 | 0.0106 | 0.0068 | 0.0134 | 0.0196 | 0.0050 | 0.0066 | 0.0084 | 0.0126 | 0.0180 | 0.0024 | 0.0040 | 0.0004 | 0.0034 |
| | Bootstrap t | 0.0080 | 0.0088 | 0.0044 | 0.0048 | 0.0178 | 0.0058 | 0.0056 | 0.0072 | 0.0068 | 0.0096 | 0.0162 | 0.0046 | 0.0072 | 0.0072 |
| | Percentile | 0.0178 | 0.0046 | 0.0068 | 0.0064 | 0.0236 | 0.0042 | 0.0066 | 0.0078 | 0.0066 | 0.0074 | 0.0140 | 0.0058 | 0.0062 | 0.0094 |
| | $BC_a$ | 0.0210 | 0.0124 | 0.0074 | 0.0066 | 0.0238 | 0.0118 | 0.0054 | 0.0124 | 0.0068 | 0.0056 | 0.0150 | 0.0080 | 0.0054 | 0.0074 |

## III.4 ANALYSIS OF THE DINOSAUR DATA

Let us revisit the dinosaur data and apply QE and the proposed indices to construct confidence intervals for differences in entropy for all possible pairs of intervals.

First, we consider the simple distance matrix assuming equal distance between all pairs of eight dinosaur families. As noted earlier, in this case, the QE is same as the Gini-Simpson index. We also use two other distance matrices. The matrix $\Delta_1$ given earlier is based upon the diets of the dinosaur families listed in Table 8, which assumes that similar diets will have similar food chains and hence shorter distances between families. We also used $\Delta_2$ in Equation (III.2.2), which becomes:

$$\Delta_2 = \begin{bmatrix} 0 & 1.338 & 4.012 & 5.804 & 3.725 & 3.319 & 3.501 & 5.804 \\ 1.338 & 0 & 3.674 & 5.466 & 3.386 & 2.981 & 3.163 & 5.466 \\ 4.012 & 3.674 & 0 & 2.792 & 1.288 & 1.693 & 1.511 & 2.792 \\ 5.804 & 5.466 & 2.792 & 0 & 3.079 & 3.485 & 3.303 & 1.000 \\ 3.725 & 3.386 & 1.288 & 3.079 & 0 & 1.405 & 1.223 & 3.079 \\ 3.319 & 2.981 & 1.693 & 3.485 & 1.405 & 0 & 1.182 & 3.485 \\ 3.501 & 3.163 & 1.511 & 3.303 & 1.223 & 1.182 & 0 & 3.303 \\ 5.804 & 5.466 & 2.792 & 1.000 & 3.079 & 3.485 & 3.303 & 0 \end{bmatrix}$$

The practical equivalence confidence intervals can be used to analyze the dinosaur data in Table 7. Table 15 gives the 95% practical equivalence intervals for the difference in entropy for all pairs of upper, middle and lower intervals. Each bootstrap bound was computed using $B = 7,500$ bootstrap samples.

It is found that confidence intervals involving only upper and middle intervals are relatively narrower, while the intervals involving the lower intervals are quite a bit wider, regardless of standard normal or bootstrap techniques, Shannon entropy of Rao's QE. This is because of the relatively large sample sizes for the upper and middle intervals (90 and 123, respectively) and relatively small samples for the lower interval ($n$=32). Given the results in Section III.4, practical equivalence inference should be based upon standard normal confidence intervals of QE index with $\Delta_2$. Thus, to reject the null hypothesis

$$H_0 : |H_{Qi} - H_{Qj}| > \delta \ for \ some \ i \neq j$$

vs.

$$H_1 : |H_{Qi} - H_{Qj}| \leq \delta \ for \ all \ i \neq j,$$

TABLE 15. Dinosaur data revisit: 95 % confidence intervals for differences in entropy for all possible pairs of intervals

| Pairs of Intervals | Methods | Standard Normal | | Bootstrap-t | | Percentile | | $BC_a$ | |
|---|---|---|---|---|---|---|---|---|---|
| $H_{upper} - H_{middle}$ | Shannon | -0.303 | 0.096 | -0.309 | 0.126 | -0.318 | 0.093 | -0.299 | 0.073 |
| | $QE_0$ | -0.124 | 0.023 | -0.126 | 0.023 | -0.133 | 0.019 | -0.132 | 0.019 |
| | $QE_1$ | -0.424 | 0.066 | -0.424 | 0.086 | -0.421 | 0.068 | -0.416 | 0.061 |
| | $QE_2$ | -0.251 | 0.433 | -0.077 | 0.411 | -0.207 | 0.252 | -0.076 | 0.132 |
| | $i\_QE_0$ | -0.177 | 0.048 | -0.152 | 0.032 | -0.161 | 0.024 | -0.161 | 0.023 |
| | $i\_QE_1$ | -0.209 | 0.016 | -0.231 | 0.038 | -0.232 | 0.037 | -0.228 | 0.033 |
| | $i\_QE_2$ | -0.083 | 0.142 | -0.023 | 0.126 | -0.068 | 0.082 | -0.025 | 0.043 |
| $H_{middle} - H_{lower}$ | Shannon | -0.214 | 0.344 | -0.315 | 0.332 | -0.159 | 0.425 | -0.243 | 0.531 |
| | $QE_0$ | -0.100 | 0.171 | -0.106 | 0.189 | -0.058 | 0.215 | -0.070 | 0.237 |
| | $QE_1$ | -0.529 | 0.320 | -0.621 | 0.329 | -0.478 | 0.358 | -0.509 | 0.398 |
| | $QE_2$ | -0.603 | 0.233 | -0.620 | -0.095 | -0.270 | 0.208 | -0.538 | 0.652 |
| | $i\_QE_0$ | -0.071 | 0.193 | -0.135 | 0.198 | -0.076 | 0.257 | -0.092 | 0.281 |
| | $i\_QE_1$ | -0.189 | 0.075 | -0.311 | 0.159 | -0.273 | 0.196 | -0.297 | 0.219 |
| | $i\_QE_2$ | -0.192 | 0.073 | -0.188 | -0.032 | -0.088 | 0.068 | -0.178 | 0.222 |
| $H_{upper} - H_{lower}$ | Shannon | -0.333 | 0.255 | -0.411 | 0.264 | -0.296 | 0.316 | -0.354 | 0.388 |
| | $QE_0$ | -0.158 | 0.128 | -0.158 | 0.141 | -0.122 | 0.159 | -0.131 | 0.175 |
| | $QE_1$ | -0.713 | 0.147 | -0.802 | 0.153 | -0.660 | 0.176 | -0.686 | 0.203 |
| | $QE_2$ | -0.535 | 0.347 | -0.470 | 0.107 | -0.279 | 0.251 | -0.416 | 0.429 |
| | $i\_QE_0$ | -0.152 | 0.146 | -0.194 | 0.147 | -0.153 | 0.187 | -0.164 | 0.205 |
| | $i\_QE_1$ | -0.303 | -0.005 | -0.405 | 0.065 | -0.373 | 0.097 | -0.387 | 0.110 |
| | $i\_QE_2$ | -0.179 | 0.119 | -0.142 | 0.030 | -0.091 | 0.082 | -0.136 | 0.140 |

the quantity $\delta$ defining the boundary has be to at least 0.192, because the normal confidence intervals based upon $i\_QE_2$ are (-0.083, 0.142), (-0.192, 0.073) and (-0.179, 0.119). Note that the largest absolute value of the boundaries is 0.192. The value of $\delta$ defining practical equivalence of Shannon and Gini-Simpson entropies can be applied in a similar way.

The quadratic entropy index reflects both the differences and abundances of the species. When a species list is given without abundance data, using the QE index and postulating equal abundances, one derives the only biodiversity index from a traditional ecological index of diversity. Its extensive form is identical with the sum of differences or distances between the species present. The QE index trivially satisfies monotonicity, an important property for biodiversity indices.

As when constructing practical equivalence intervals for analyzing biodiversity, the challenge still remains on how to choose the rejection boundaries. One possibility is to consider an analogous community that has changed in biodiversity, then calculate the confidence boundaries based upon before and after data to derive such a $\delta$. Another approach is to build simulation geometric models and estimate the maximum boundaries for differences of two quadratic entropies to obtain the practical equivalence.

# CHAPTER IV

# CLUSTER ANALYSIS OF MULTINOMIAL DATA

Cluster analysis is a common data mining and analysis technique that has been used in many fields such as biology (Eisen, Spellman, Brown, and Botstein, 1998), medicine Romesburg, 2004), market research (Punj and Stewart, 1983) and social network analysis (Scott, 1988). The aim of cluster analysis is to cluster or group the observations into disjoint clusters. Data clustering algorithms can be hierarchical or partitional. Hierarchical algorithms find successive clusters using previously established clusters. These algorithms can be either agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters. Partitional algorithms typically determine all clusters at once.

An important step in any clustering is to select a distance measure, which will determine how the similarity of two elements is calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and may not according to another. Let $X = (x_1,...x_m)$ and $Y = (y_1,...y_m)$ be two vectors in real $m$-space. Commonly used distance functions include:

- Euclidean distance:

$$D_E = \sqrt{(x_1-y_1)^2+...+(x_m-y_m)^2} = \sqrt{\sum_{i=1}^{m}(x_i-y_i)^2}$$

- Mahalanobis distance

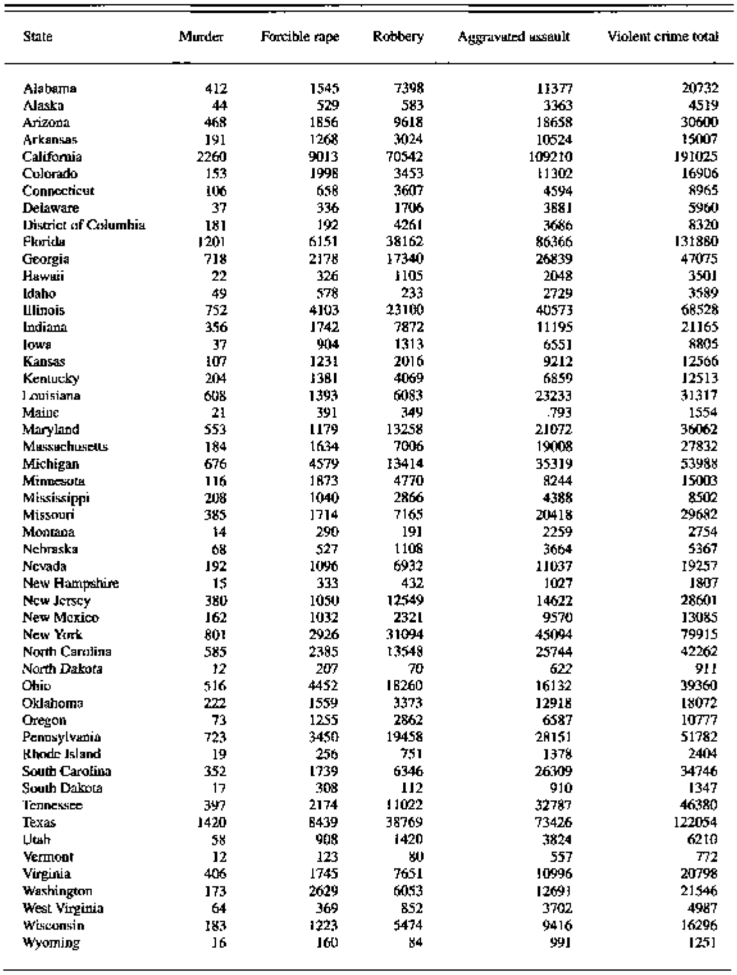
$$D_{Mah} = \sqrt{(X-Y)'\Sigma^{-1}(X-Y)}$$

- Manhattan distance

$$D_{Man} = ||X-Y|| = \sum_{i=1}^{m}|x_i-y_i|$$

All these distances are defined generally for continuous data. Not many methods have been proposed to define distances for quantitative data, which brings challenge in clustering categorical data. Suppose we are interested in clustering the states with similar violent crime statistics as in the next example. The data in Table 16 are on states crime rate in 2007 for 50 states and District of Columbia (From now on referred as 51 states). These are taken from Bureau of Justice Statistics web site (http://www.ojp.usdoj.gov/bjs) in the Department of Justice for illustrative purpose. The violent crimes include murder and non-negligent manslaughter, forcible rape, robbery and aggravated assault. The researcher may be interested in the clustering 51 states into several clusters based upon the similarity of violent crime rates.

TABLE 16. State violent crime statistics in 2007



| State | Murder | Forcible rape | Robbery | Aggravated assault | Violent crime total |
|---|---|---|---|---|---|
| Alabama | 412 | 1545 | 7398 | 11377 | 20732 |
| Alaska | 44 | 529 | 583 | 3363 | 4519 |
| Arizona | 468 | 1856 | 9618 | 18658 | 30600 |
| Arkansas | 191 | 1268 | 3024 | 10524 | 15007 |
| California | 2260 | 9013 | 70542 | 109210 | 191025 |
| Colorado | 153 | 1998 | 3453 | 11302 | 16906 |
| Connecticut | 106 | 658 | 3607 | 4594 | 8965 |
| Delaware | 37 | 336 | 1706 | 3881 | 5960 |
| District of Columbia | 181 | 192 | 4261 | 3686 | 8320 |
| Florida | 1201 | 6151 | 38162 | 86366 | 131880 |
| Georgia | 718 | 2178 | 17340 | 26839 | 47075 |
| Hawaii | 22 | 326 | 1105 | 2048 | 3501 |
| Idaho | 49 | 578 | 233 | 2729 | 3589 |
| Illinois | 752 | 4103 | 23100 | 40573 | 68528 |
| Indiana | 356 | 1742 | 7872 | 11195 | 21165 |
| Iowa | 37 | 904 | 1313 | 6551 | 8805 |
| Kansas | 107 | 1231 | 2016 | 9212 | 12566 |
| Kentucky | 204 | 1381 | 4069 | 6859 | 12513 |
| Louisiana | 608 | 1393 | 6083 | 23233 | 31317 |
| Maine | 21 | 391 | 349 | 793 | 1554 |
| Maryland | 553 | 1179 | 13258 | 21072 | 36062 |
| Massachusetts | 184 | 1634 | 7006 | 19008 | 27832 |
| Michigan | 676 | 4579 | 13414 | 35319 | 53988 |
| Minnesota | 116 | 1873 | 4770 | 8244 | 15003 |
| Mississippi | 208 | 1040 | 2866 | 4388 | 8502 |
| Missouri | 385 | 1714 | 7165 | 20418 | 29682 |
| Montana | 14 | 290 | 191 | 2259 | 2754 |
| Nebraska | 68 | 527 | 1108 | 3664 | 5367 |
| Nevada | 192 | 1096 | 6932 | 11037 | 19257 |
| New Hampshire | 15 | 333 | 432 | 1027 | 1807 |
| New Jersey | 380 | 1050 | 12549 | 14622 | 28601 |
| New Mexico | 162 | 1032 | 2321 | 9570 | 13085 |
| New York | 801 | 2926 | 31094 | 45094 | 79915 |
| North Carolina | 585 | 2385 | 13548 | 25744 | 42262 |
| North Dakota | 12 | 207 | 70 | 622 | 911 |
| Ohio | 516 | 4452 | 18260 | 16132 | 39360 |
| Oklahoma | 222 | 1559 | 3373 | 12918 | 18072 |
| Oregon | 73 | 1255 | 2862 | 6587 | 10777 |
| Pennsylvania | 723 | 3450 | 19458 | 28151 | 51782 |
| Rhode Island | 19 | 256 | 751 | 1378 | 2404 |
| South Carolina | 352 | 1739 | 6346 | 26309 | 34746 |
| South Dakota | 17 | 308 | 112 | 910 | 1347 |
| Tennessee | 397 | 2174 | 11022 | 32787 | 46380 |
| Texas | 1420 | 8439 | 38769 | 73426 | 122054 |
| Utah | 58 | 908 | 1420 | 3824 | 6210 |
| Vermont | 12 | 123 | 80 | 557 | 772 |
| Virginia | 406 | 1745 | 7651 | 10996 | 20798 |
| Washington | 173 | 2629 | 6053 | 12691 | 21546 |
| West Virginia | 64 | 369 | 852 | 3702 | 4987 |
| Wisconsin | 183 | 1223 | 5474 | 9416 | 16296 |
| Wyoming | 16 | 160 | 84 | 991 | 1251 |

The four categories of violent crimes can be assumed to be categories of a multinomial distribution. Since Euclidean distance, Manhattan distance and Mahalanobis distance are designed for the continuous data, none of them can catch the internal correlation of subcategories of crimes and hence cannot be applied here. Bhattacharyya distance was proposed by Bhattacharyya (1943) to measure the similarity of two discrete probability distributions and can be used for this purpose.

Let $X_i$ denote the categorical group variables (type of crime) with $s$ levels and $n_i$ be the total frequencies (number of crime for $i$-th state). Let $P_i = (p_{i1}, p_{i2}, ..., p_{is})$ be the vector of relative frequencies. The Bhattacharyya distance between $i$-th and $j$-th subject (state) can be calculated as (Bhattacharyya 1943),

$$D_{B,ij} = \sqrt{\sum_{m=1}^{s} (p_{im}^{1/2} - p_{jm}^{1/2})^2} \tag{IV.0.1}$$

In this chapter we will propose a new distance based upon Rao's quadratic entropy and use it to cluster the multinomially distributed data. Rao and Boudreau (1984) have used Gini-Simpson index for clustering blood group data in human populations. In Section IV.1 we define the distance based upon Rao's quadratic entropy; The performance of this new distance will be compared with Euclidean distance, Bhattacharyya distance and Gini-Simpson distance for simulated data in Section IV.2 and for state crime data in Section IV.3. In Section IV.4 this quadratic entropy distance will be generalized to multiple variables with clustering results on both simulated and state crime data. We will conclude this chapter with some remarks.

## IV.1 DEFINITION OF QUADRATIC ENTROPY DISTANCE

As discussed in Chapter I, the total quadratic entropy diversity ($SST$) can be decomposed into two parts: $SSW$ and $SSB$, where $SSW$ measures the similarity of diversities among populations and $SSB$ measures the difference of diversities between populations. Hence the quantity $\frac{SSB}{SST}$ can serve as a distance between two populations.

Let $X_i$ denote the categorical variable (type of crime) with $s$ levels and $n_i$ be the total frequencies (number of crimes for the $i$-th state). Let $P_i = (p_{i1}, p_{i2}, ..., p_{is})$ be the vector of relative frequencies. The quadratic entropy distance can be defined as,

$$D_{QE,ij} = \frac{SSB_{ij}}{SST_{ij}} = \frac{\bar{P}\Delta\bar{P} - \frac{n_i}{n_i+n_j}P_i'\Delta P_i - \frac{n_j}{n_i+n_j}P_j'\Delta P_j}{\bar{P}\Delta\bar{P}}, \tag{IV.1.1}$$

where $\bar{P} = (n_iP_i + n_jP_j)/(n_i + n_j)$ and $\Delta$ can be a predetermined matrix or one derived from data as proposed in Equation (II.2.1) or (II.2.2). It may be noted that this distance does not satisfy the triangle inequality. However, we can still use this for clustering purpose (Mardia, Kent and Bibby, 1979).

## IV.2 EMPIRICAL COMPARISONS

In this section, we generate a simulation data set to compare quadratic entropy distance with Euclidean Distance and Bhattacharyya distance. The data set consists of samples from geometric distributions (See Appendix A) with $n = 100$, $s = 4$ and $k$ = 0.4, 0.6, and 0.8 to produce three sets of samples. Each set of data consists of 1,000 samples. The quadratic entropy distances, along with Euclidean and Bhattacharyya distances are used in hierarchical clustering methods with complete linkage algorithms.

The two difference matrices $\Delta_1$ and $\Delta_2$ used for quadratic entropy are:

$$\Delta_1 = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix},$$

and $\Delta_2 = (d_{ij})$, where

$$d_{ij} = \begin{cases} 0 & if \quad i = j \\ 1 & if \quad \pi_{.i} = \pi_{.j} = 0 \\ |log(\pi_{.i})| + 1 & if \quad \pi_{.j} = 0 \\ |log(\pi_{.j})| + 1 & if \quad \pi_{.i} = 0 \\ |log(\pi_{.i}) - log(\pi_{.j})| + 1 & else. \end{cases}$$

To assess the quality of our algorithm, we need some objective external criteria. The external criteria could be the true class information. In order to compare clustering results against an objective external criteria, we employ the well known adjusted Rand index (Hubert and Arabie, 1985; Steinley, 2004) as the measure of agreement. Rand index is defined as the number of pairs of objects that are either in the same group or in different groups in both partitions divided by the total number of pairs of objects. The Rand index lies between 0 and 1. The adjusted Rand index is corrected-for-chance version of the Rand index; and it has the maximum value of 1 and its expected value is 0 in the case of random clusters. A larger adjusted Rand index means a higher agreement between two partitions. The adjusted Rand index is recommended for measuring agreement even when the partitions compared have different numbers of clusters. See Appendix B for details.

TABLE 17. Adjusted Rand index of the simulated data with $k_1 = 0.4$ and $k_2 = 0.6$

| Cluster | Euclidean | Bhattacharyya | $QE_{\Delta_1}$ | $QE_{\Delta_2}$ |
|---|---|---|---|---|
| 2 | 0.618472 | 0.248051 | 0.618046 | 0.629427 |
| 3 | 0.570018 | 0.216502 | 0.613762 | 0.614215 |
| 4 | 0.517917 | 0.271299 | 0.547501 | 0.553621 |
| 5 | 0.519915 | 0.194702 | 0.528180 | 0.530142 |
| 6 | 0.652023 | 0.295381 | 0.620194 | 0.660271 |
| 7 | 0.651715 | 0.259495 | 0.675896 | 0.676085 |
| 8 | 0.612986 | 0.198902 | 0.600519 | 0.635207 |
| 9 | 0.538132 | 0.148592 | 0.644494 | 0.645201 |
| 10 | 0.643204 | 0.225701 | 0.604661 | 0.643304 |

TABLE 18. Adjusted Rand index of the simulated data with $k_1 = 0.4$ and $k_2 = 0.8$

| Cluster | Euclidean | Bhattacharyya | $QE_{\Delta_1}$ | $QE_{\Delta_2}$ |
|---|---|---|---|---|
| 2 | 0.998881 | 0.999121 | 0.998961 | 0.999085 |
| 3 | 0.999440 | 0.999600 | 0.999240 | 0.999351 |
| 4 | 0.999121 | 0.999720 | 0.999041 | 0.999285 |
| 5 | 0.999080 | 0.999600 | 0.999001 | 0.999561 |
| 6 | 0.999400 | 0.999720 | 0.999441 | 0.999512 |
| 7 | 0.999041 | 0.999560 | 0.998722 | 0.999225 |
| 8 | 0.999282 | 0.999680 | 0.999241 | 0.999354 |
| 9 | 0.999200 | 0.999600 | 0.999043 | 0.999312 |
| 10 | 0.999081 | 0.999600 | 0.998801 | 0.999251 |

TABLE 19. Adjusted Rand index of the simulated data with $k_1 = 0.6$ and $k_2 = 0.8$

| Cluster | Euclidean | Bhattacharyya | $QE_{\Delta_1}$ | $QE_{\Delta_2}$ |
|---|---|---|---|---|
| 2 | 0.889204 | 0.931127 | 0.872781 | 0.892051 |
| 3 | 0.863560 | 0.927207 | 0.888591 | 0.889517 |
| 4 | 0.838906 | 0.890506 | 0.838802 | 0.852143 |
| 5 | 0.840396 | 0.913819 | 0.885124 | 0.886521 |
| 6 | 0.839969 | 0.905943 | 0.843705 | 0.853210 |
| 7 | 0.884875 | 0.922546 | 0.865229 | 0.895130 |
| 8 | 0.870603 | 0.914243 | 0.859382 | 0.886218 |
| 9 | 0.899273 | 0.893167 | 0.897176 | 0.899042 |
| 10 | 0.813119 | 0.913582 | 0.864634 | 0.882682 |

Table 17, 18 and 19 list the adjusted Rand index achieved by hierarchical algorithms for three type of distances. Quadratic entropy distance has a better partition than Euclidean distance and Bhattacharyya distance when $k_1 = 0.4$ and $k_2 = 0.6$. In the case of $k_1 = 0.4$ and $k_2 = 0.8$, or $k_1 = 0.6$ and $k_2 = 0.8$, Quadratic entropy distance have a better partition than Euclidean distance, but not better than Bhattacharyya distance. It is not surprising that all three distances obtain the best results when the specified number of clusters is correct since its model perfectly matches the data. However we often do not know the exact number of clusters in practice. When the specified number of clusters is not correct, quadratic entropy distance based method performs better than the methods based upon Euclidean distance and better than Bhattacharyya distance in some cases.

## IV.3 APPLICATION TO STATE VIOLENT CRIME DATA

Figure 1, 2 and 3 list the clustering results of state crime statistics data when different distances are used.

In Figure 1 Euclidean distance puts most other inner states as one group; industrial states in the north such as Ohio, Pennsylvania, Illinois, Michigan, New York, and south states such as Arizona, Georgia, North Carolina, Tennessee, and Washington state as another group; California, Texas, and Florida as three other separate groups.

In Figure 2 Bhattacharyya distance puts most mid-western states and northeastern states as as one group, Florida and Texas as another group; Illinois and New York as one group; California as a separate group and the rest as one group.

In Figure 3 Rao's quadratic entropy puts most inner states as one group; most states near the ocean as one group; North Dakota and South Dakota as one group; Idaho, Montana and Wyoming as one group; New Jersey, Ohio and District of Columbia as one group.

FIG. 1. Clusters of states based upon violent crime with Euclidean distance

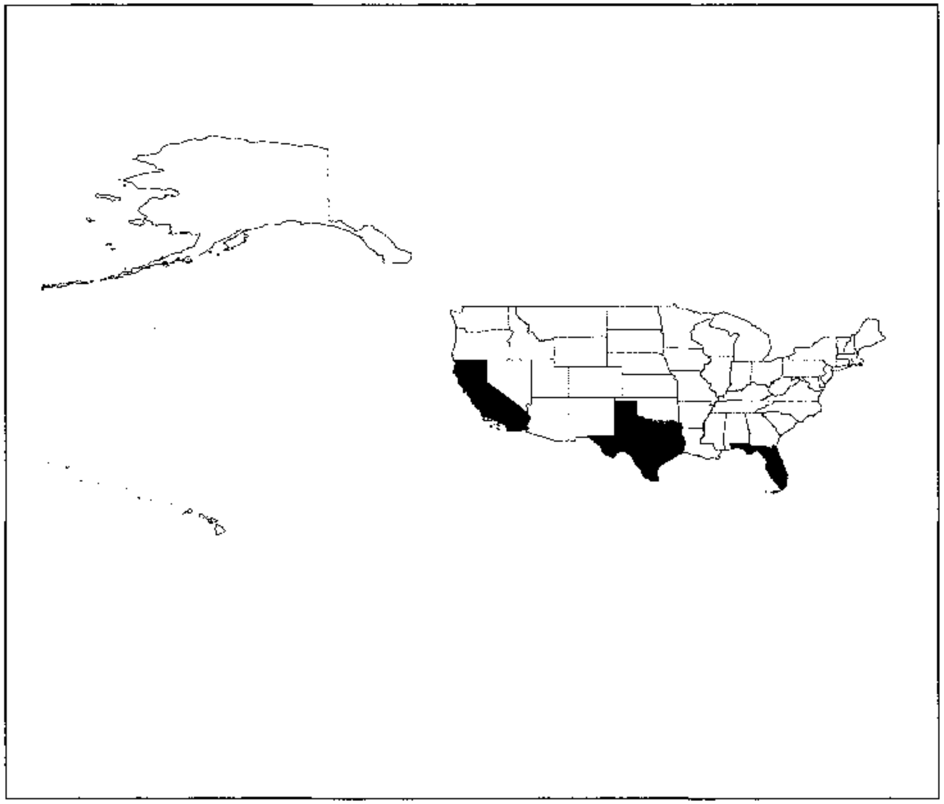


FIG. 2. Clusters of states based upon violent crime with Bhattacharyya distance

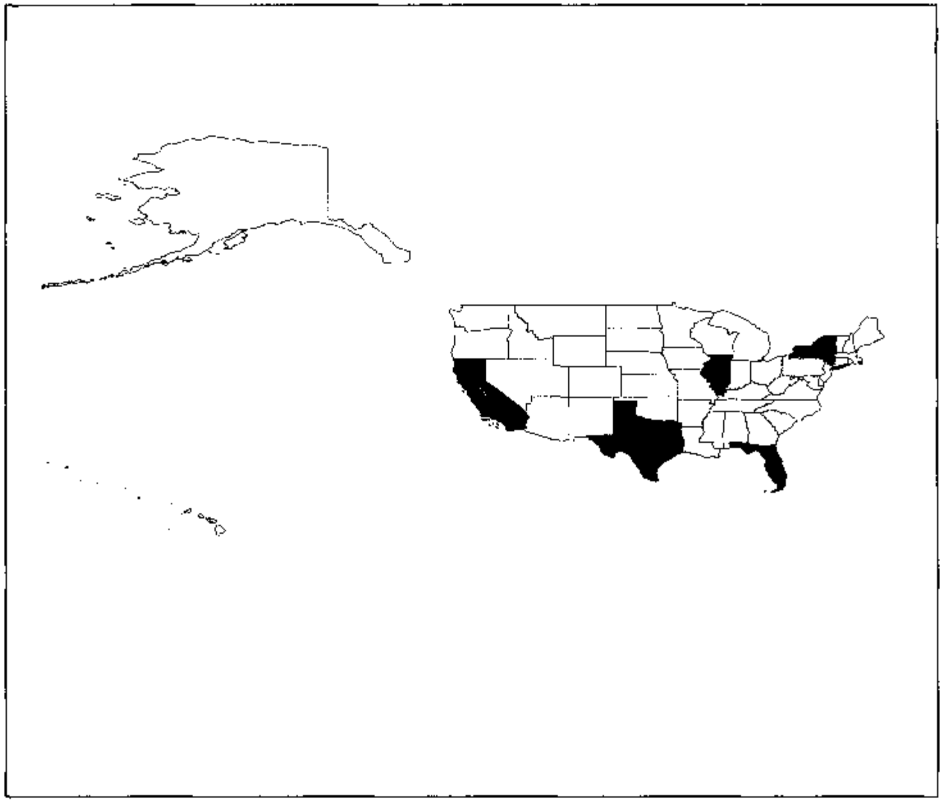
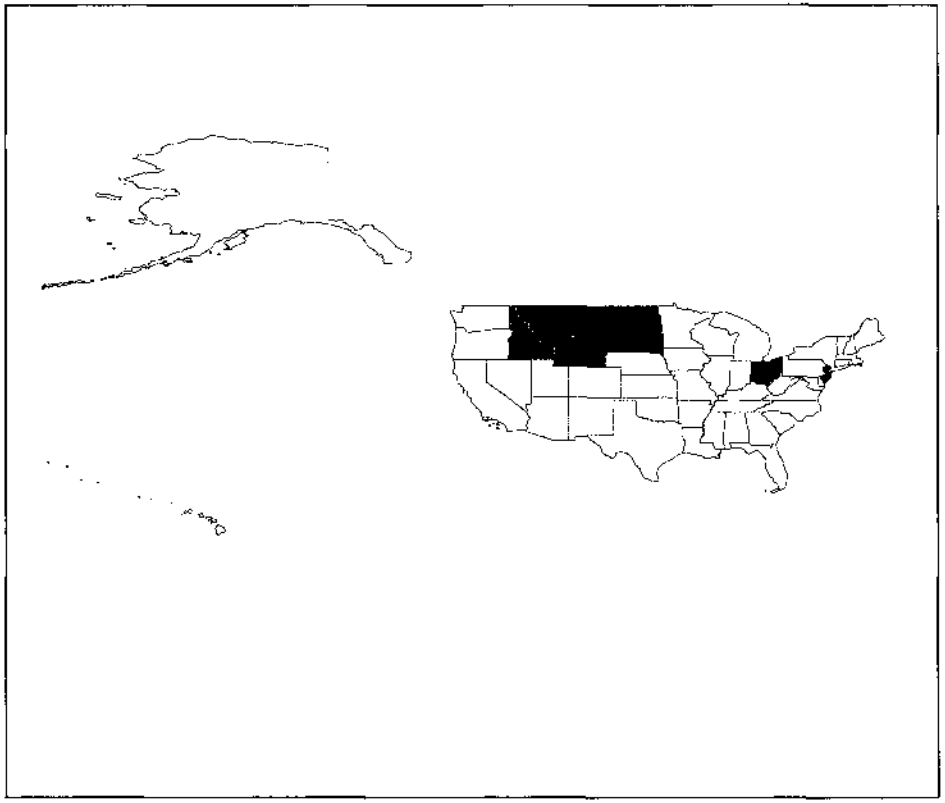


*FIG. 3. Clusters of states based upon violent crime with quadratic entropy distance*



## IV.4 CLUSTER ANALYSIS FOR MULTIPLE VARIABLES

In most of clustering problems there will always be data on more than one variable. The data in Table 20 on 51 states crime rate in 2007 (http://www.ojp.usdoj.gov/bjs) include two type of crimes, namely, violent crime and property crime. Violent crime includes murder and non-negligent manslaughter, forcible rape, robbery, and aggravated assault; and property crime includes three types: burglary, larceny-theft, and motor vehicle theft. The researcher may be interested in clustering 51 states into several clusters based upon the similarity of both violent and property crime rates.

### IV.4.1 Quadratic Entropy Distance for Multiple Variables Clustering

We can easily generalize the quadratic entropy distance to the case of multiple variables. Let $X$ and $Y$ denote the categorical variables (violent crime and property crime) with levels $s_1$ and $s_2$ and $n$ and $m$ be the total number (of crimes), respectively. Let $P = (p_1, p_2, ..., p_{s_1})$ be the vector of proportions of $s_1$ (violent crime) categories and $Q = (q_1, q_2, ..., q_{s_2})$ be the vector of proportions of $s_2$ (property crime) categories.

For characteristic variables X and Y, the Bhattacharyya distance between the $i$-th and $j$-th subject (state) can be calculated as,

$$D_{B,x} = \sqrt{\sum_{m=1}^{s_1} (p_{im}^{1/2} - p_{jm}^{1/2})^2},$$

$$D_{B,y} = \sqrt{\sum_{m=1}^{s_2} (q_{im}^{1/2} - q_{jm}^{1/2})^2}.$$

The overall distance between $i$-th and $j$-th subject(state) is constructed as,

$$D_B = \frac{n_i + n_j}{2n} D_{B,x} + \frac{m_i + m_j}{2m} D_{B,y}, \tag{IV.4.1}$$

where $n_i$ and $m_i$ are the frequencies in the $i$-th class for the two variables, respectively.

The between subjects quadratic entropy is measuring the dissimilarity between subjects and can be used for calculating the distance between $i$-th and $j$-th subject as well.

$$D_{QE,x} = \bar{P}\Delta\bar{P} - \frac{n_i}{n_i + n_j} P_i' \Delta P_i - \frac{n_j}{n_i + n_j} P_j' \Delta P_j, \tag{IV.4.2}$$

$$D_{QE,y} = \bar{Q}\Delta\bar{Q} - \frac{n_i}{n_i + n_j} Q_i' \Delta Q_i - \frac{n_j}{n_i + n_j} Q_j' \Delta Q_j, \tag{IV.4.3}$$

TABLE 20. State violent and property crime statistics in 2007

| State | Murder | Forcible rape | Robbery | Aggravated assault | Violent crime total | Burglary | Larceny-theft | Motor vehicle theft | Property crime total |
|---|---|---|---|---|---|---|---|---|---|
| Alabama | 412 | 1545 | 7398 | 11377 | 20732 | 45329 | 124238 | 14231 | 183798 |
| Alaska | 44 | 529 | 583 | 3363 | 4519 | 3682 | 16998 | 2418 | 23098 |
| Arizona | 468 | 1856 | 9618 | 18658 | 30600 | 57822 | 173581 | 48390 | 279794 |
| Arkansas | 191 | 1268 | 3024 | 10524 | 15007 | 32073 | 72978 | 7010 | 112061 |
| California | 2260 | 9013 | 70542 | 109210 | 191025 | 237011 | 652256 | 219393 | 1108660 |
| Colorado | 153 | 1998 | 3453 | 11302 | 16906 | 28751 | 100599 | 16792 | 146141 |
| Connecticut | 106 | 658 | 3607 | 4594 | 8965 | 15162 | 59724 | 9166 | 84052 |
| Delaware | 37 | 336 | 1706 | 3881 | 5960 | 6341 | 20486 | 2316 | 29143 |
| District of Columbia | 181 | 192 | 4261 | 3686 | 8320 | 3926 | 17382 | 7600 | 28908 |
| Florida | 1201 | 6151 | 38162 | 86366 | 131880 | 181837 | 490848 | 73662 | 746347 |
| Georgia | 718 | 2178 | 17340 | 26839 | 47075 | 90692 | 239053 | 42597 | 372342 |
| Hawaii | 22 | 326 | 1105 | 2048 | 3501 | 9097 | 38416 | 6715 | 54228 |
| Idaho | 49 | 578 | 233 | 2729 | 3589 | 6977 | 24482 | 2227 | 33685 |
| Illinois | 752 | 4103 | 23100 | 40573 | 68528 | 75521 | 267909 | 33892 | 377322 |
| Indiana | 356 | 1742 | 7872 | 11195 | 21165 | 46917 | 149052 | 19556 | 215526 |
| Iowa | 37 | 904 | 1313 | 6551 | 8805 | 16942 | 56327 | 4885 | 78154 |
| Kansas | 107 | 1231 | 2016 | 9212 | 12566 | 20262 | 73293 | 8564 | 102120 |
| Kentucky | 204 | 1381 | 4069 | 6859 | 12513 | 27684 | 70455 | 8674 | 106813 |
| Louisiana | 608 | 1393 | 6083 | 23233 | 31317 | 44602 | 115208 | 15181 | 174991 |
| Maine | 21 | 391 | 349 | 793 | 1554 | 6676 | 24057 | 1259 | 31992 |
| Maryland | 553 | 1179 | 13258 | 21072 | 36062 | 37093 | 127308 | 28395 | 192796 |
| Massachusetts | 184 | 1634 | 7006 | 19008 | 27832 | 35662 | 103594 | 14990 | 154246 |
| Michigan | 676 | 4579 | 13414 | 35319 | 53988 | 75429 | 191195 | 42151 | 308775 |
| Minnesota | 116 | 1873 | 4770 | 8244 | 15003 | 29669 | 115634 | 12527 | 157829 |
| Mississippi | 208 | 1040 | 2866 | 4388 | 8502 | 27959 | 58084 | 7382 | 93424 |
| Missouri | 385 | 1714 | 7165 | 20418 | 29682 | 43447 | 152527 | 23784 | 219759 |
| Montana | 14 | 290 | 191 | 2259 | 2754 | 3027 | 21707 | 1755 | 26489 |
| Nebraska | 68 | 527 | 1108 | 3664 | 5367 | 9047 | 41854 | 5201 | 56102 |
| Nevada | 192 | 1096 | 6932 | 11037 | 19257 | 24840 | 49745 | 22331 | 96916 |
| New Hampshire | 15 | 333 | 432 | 1027 | 1807 | 4986 | 18611 | 1299 | 24896 |
| New Jersey | 380 | 1050 | 12549 | 14622 | 28601 | 37481 | 132795 | 21950 | 192226 |
| New Mexico | 162 | 1032 | 2321 | 9570 | 13085 | 18992 | 45463 | 8939 | 73394 |
| New York | 801 | 2926 | 31094 | 45094 | 79915 | 64858 | 288919 | 28039 | 381816 |
| North Carolina | 585 | 2385 | 13548 | 25744 | 42262 | 108799 | 233592 | 27963 | 370354 |
| North Dakota | 12 | 207 | 70 | 622 | 911 | 2164 | 9010 | 914 | 12088 |
| Ohio | 516 | 4452 | 18260 | 16132 | 39360 | 98510 | 263918 | 33781 | 396209 |
| Oklahoma | 222 | 1559 | 3373 | 12918 | 18072 | 34121 | 79981 | 13460 | 127562 |
| Oregon | 73 | 1255 | 2862 | 6587 | 10777 | 22822 | 94773 | 14548 | 132143 |
| Pennsylvania | 723 | 3450 | 19458 | 28151 | 51782 | 56022 | 211097 | 26457 | 293577 |
| Rhode Island | 19 | 256 | 751 | 1378 | 2404 | 5236 | 19281 | 3226 | 27743 |
| South Carolina | 352 | 1739 | 6346 | 26309 | 34746 | 45214 | 126041 | 17027 | 188282 |
| South Dakota | 17 | 308 | 112 | 910 | 1347 | 2378 | 10043 | 735 | 13156 |
| Tennessee | 397 | 2174 | 11022 | 32787 | 46380 | 61715 | 168350 | 21659 | 251724 |
| Texas | 1420 | 8439 | 38769 | 73426 | 122054 | 228310 | 662937 | 93896 | 985142 |
| Utah | 58 | 908 | 1420 | 3824 | 6210 | 15541 | 68241 | 8812 | 92594 |
| Vermont | 12 | 123 | 80 | 557 | 772 | 3106 | 10683 | 641 | 14430 |
| Virginia | 406 | 1745 | 7651 | 10996 | 20798 | 31689 | 144469 | 14051 | 190209 |
| Washington | 173 | 2629 | 6053 | 12691 | 21546 | 52705 | 170404 | 37620 | 260729 |
| West Virginia | 64 | 369 | 852 | 3702 | 4987 | 10814 | 31447 | 3492 | 45753 |
| Wisconsin | 183 | 1223 | 5474 | 9416 | 16296 | 27840 | 117686 | 13433 | 158959 |
| Wyoming | 16 | 160 | 84 | 991 | 1251 | 2348 | 11840 | 796 | 14984 |

where

$$\bar{P} = (n_i P_i + n_j P_j)/2,$$

$$\bar{Q} = (n_i Q_i + n_j Q_j)/2,$$

and $\Delta$ can be a predetermined matrix or derived from data. The overall distance between the $i$-th and $j$-th subject (state) is constructed as,

$$D_{QE} = (D_{QE,x} + D_{QE,y})/2 \tag{IV.4.4}$$

The between subjects quadratic entropy measures the dissimilarity between subjects and the ratio of $\frac{SSB}{SST}$ can be used to calculate the distance between $i$-th and $j$-th subject as well.

$$SST_{QE,xy} = \frac{n_i + n_j}{2n} \bar{P} \Delta_X \bar{P} + \frac{m_i + m_j}{2m} \bar{Q} \Delta_Y \bar{Q},$$

$$SSB_{QE,xy} = \frac{n_i + n_j}{2n} (\bar{P} \Delta_X \bar{P} - \frac{n_i}{n_i + n_j} P_i' \Delta_X P_i - \frac{n_j}{n_i + n_j} P_j' \Delta_X P_j) +$$
$$\frac{m_i + m_j}{2m} (\bar{Q} \Delta_Y \bar{Q} - \frac{m_i}{m_i + m_j} Q_i' \Delta_Y Q_i - \frac{m_j}{m_i + m_j} Q_j' \Delta_Y Q_j),$$

where

$$\bar{P} = (n_i P_i + n_j P_j)/(n_i + n_j),$$

$$\bar{Q} = (m_i Q_i + m_j Q_j)/(m_i + m_j).$$

$\Delta_X$ and $\Delta_Y$ can be a predetermined matrix or derived from data. The overall distance between $i$-th and $j$-th subject (state) is constructed as,

$$D_{QE} = SSB_{QE,xy}/SST_{QE,xy}. \tag{IV.4.5}$$

### IV.4.2 Application to State Violent and Property Crime Data

We illustrate these clustering methods with state violent and property crime data.

Figure 4, 5 and 6 list the clustering results of state crime statistics by Euclidean distance, Bhattacharyya distance and quadratic entropy distance, respectively.

In Figure 4 Euclidean distance clusters most west-mid inner state as one group; east states as another group; Illinois and New York as one group; California, Texas and Florida as three other separate groups.

In Figure 5 Bhattacharyya distance puts most mid-western states and northeastern states as as one group, Florida and Texas as another group; Illinois and New York as one group; California as a separate group and the rest as one group.

In Figure 6 quadratic entropy distance puts most inner states as one group; most north east states as one group; Montana and Wyoming as one group; Nevada as one group; District of Columbia as one group.

In this chapter, we proposed quadratic entropy distance for clustering multinomially distributed data. The simulation results show that our new method performs significantly better than Euclidean distance and Bhattacharyya distance, especially when the number of clusters is incorrectly specified. We were also able to generalize the methods to more than one categorical variables. Our future work involves exploring these methods to clustering data with both continuous and discrete variables.

FIG. 4. Clusters of states based upon violent and property crime with Euclidean distance

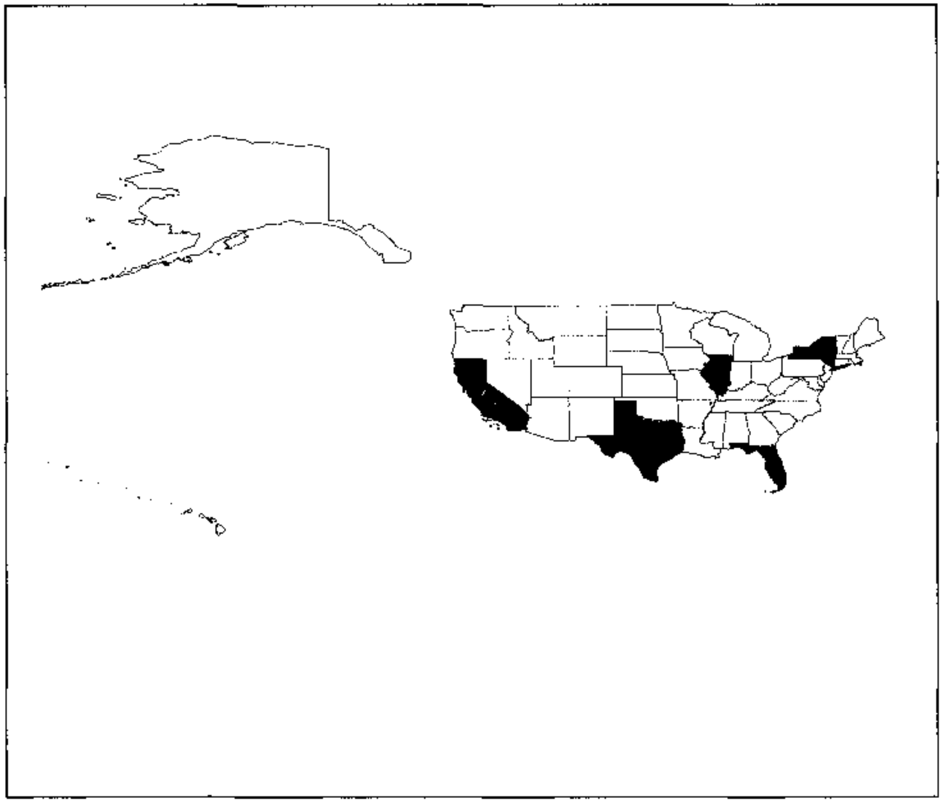
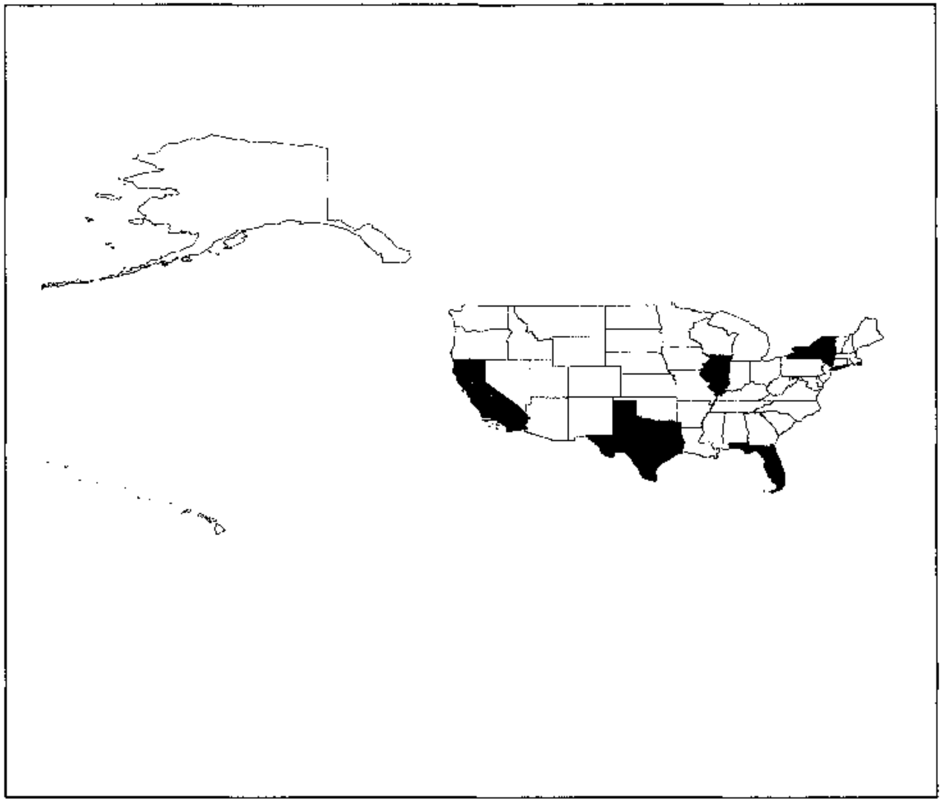


*FIG. 5. Clusters of states based upon violent and property crime with Bhattacharyya distance*



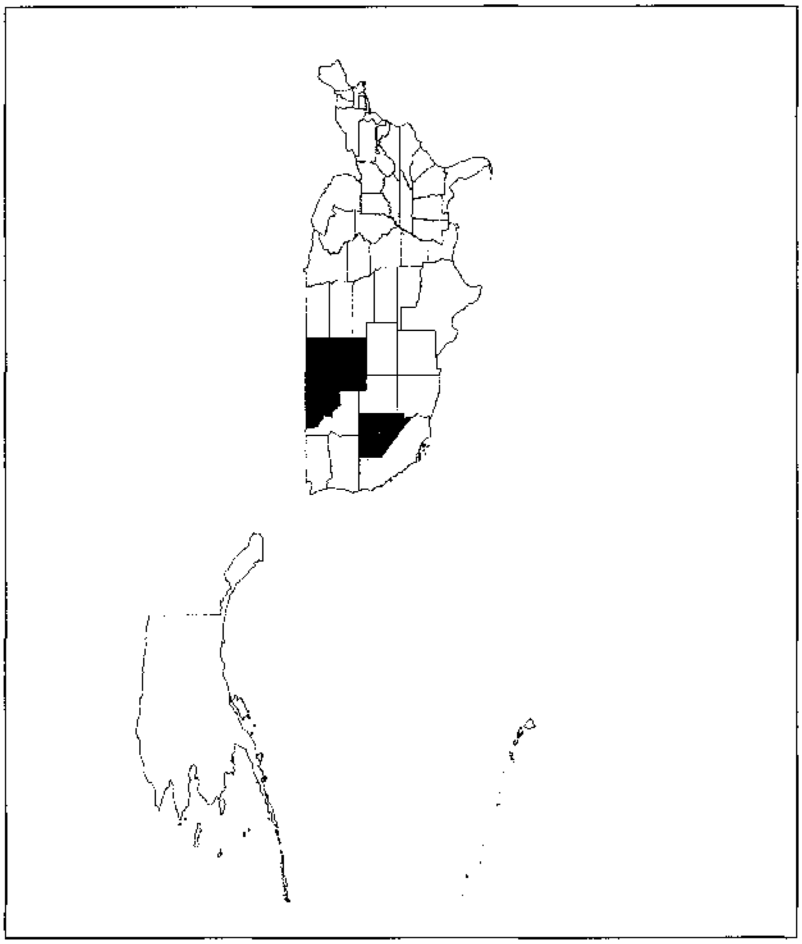


*FIG. 6. Clusters of states based upon violent and property crime with quadratic entropy distance*

# CHAPTER V

# ANALYSIS OF MULTI-RESPONSE DATA

Surveys and other studies often result in categorical response measurements being made on members of different populations or treatment groups. This arises often where individuals may mark all answers that apply when responding to a multiple-choice question such as "What criminal offenses have you been arrested?" "What type of diseases have you been diagnosed?" "What are your races?" These are all example questions appearing on surveys where the respondent is supposed to choose maybe more than one responses from a predefined list items. Survey data arising from questions of this type raise a unique challenge for analysis because of the dependence among responses provided by individual subjects.

To test the independence of two categorical variables where at least one of the categorical variables can have multiple responses, many familiar tests, such as, the Pearson Chi-square test and Fisher's exact test should not be used because of the within-subject dependency among responses. Loughin and Scherer (1998) proposed a modified chi-square test to test the multiple marginal independence (MMI) between one single response and one multiple-response categorical variable by bootstrapping. They also examined a test for conditional marginal independence (CMMI), where the conditioning is on a third single-response variable. Agresti and Liu (1999) examined the association between two multi-response categorical variables by testing simultaneous pairwise marginal independence (SPMI). Bilder and Loughin (2004) suggested bootstrapping a modified Pearson $\chi^2$ test to perform the test; Agresti and Liu (1999, 2001) and and Bilder and Loughin (2007, 2009)suggested generalized log-linear models to test for SPMI. Little research has been done applying Rao's quadratic entropy method on testing SPMI. In this chapter we develop new approaches to test marginal independence between two multi-response categorical variables with Rao's quadratic entropy.

We use the same data set from Bilder and Loughin (2004), which is from a survey conducted by the department of animal science at Kansas State University. In this survey two questions asked Kansas farmers about their "sources of veterinary information" and their "swine waste storage methods". For these questions, the farmers were permitted to select as many responses as applied from a list of items. Two hundred and seventy-nine farmers participated in the survey. Table 21 summarizes the data in a $4 \times 5$ table. For example, 34 farmers picked professional consultant as a source of veterinary information and lagoon as a waste storage method. A researcher may be interested in determining the association of waste storage methods and sources of veterinary information.

The traditional Pearson chi-square test for independence cannot be used here because of the within-subjects dependency of responses. Instead, a test for marginal independence should be performed. Specifically, $4 \times 5 = 20$ different $2 \times 2$ tables can be formed to marginally summarize all possible responses to item pairs. Table 22 is the table for responses with professional consultant and Lagoon. A "1" denotes a farmer picked that item and a "0" denotes the farmer did not pick that item. Instead of testing the independence of $4 \times 5$ table, the independence of 20, $2 \times 2$ tables is tested simultaneously. If this test is rejected, examination of the individual $2 \times 2$ tables can be followed to determine why the rejection occurs. This is analogous to the post-hoc pairwise test in analysis of variance. Rao's quadratic entropy can be applied to perform the testing.

## V.1 DERIVATION OF THE QUADRATIC ENTROPY TEST

Let W and Y denote the multiple-response categorical variables for an $r \times c$ table's row and column variables, respectively. The derivation of the Rao's quadratic entropy statistics requires consideration of two different contingency table representation of groups and responses. In the first, referred to as the original table, the $r$ groups of units correspond to rows of the table and the $c$ responses correspond to the columns, as in Table 21. Denote the cell counts in this table by $m_{ij}$, $i = 1, ..., r$; $j = 1, ..., c$. Marginal counts are denoted by $+$ subscripts: $m_{i+}$ is the total number of responses in row $i$, and $m_{+j}$ is the total number of responses in column $j$. Define $\pi_{ij}$ to be the probability that a unit chosen at random from the population falls into group $i$ and responds positively to category $j$, and let $\pi_{+j}$ be the marginal probability that a randomly chosen unit provides response category $j$.

As a second representation, let there be $R = 2^r$ rows and $C = 2^c$ columns, corresponding to all possible combinations of responses. This table is referred to as the expanded table. Counts in this table are denoted by $n_{hk}$, $h = 1..., R$; $k = 1, ..., C$. Marginal counts are again denoted by $+$ subscripts, with $n_{h+}$ being the number of responses in row $i$, $n_{+k}$ be the number of responses in column $k$ and $n_{++}$ being the total number of units in the study. Joint probabilities are denoted by $\tau_{hk} = P(W_h = 1, Y_k = 1)$.

TABLE 21. Marginal table for the Kansas farmer data

| | Sources of Veterinary information | | | | |
|---|---|---|---|---|---|
| Waste Storage Methods | Professional Consultant | Veterinarian | Local service | Magazines | Feed companies |
| Lagoon | 34 12.19% | 54 19.35% | 50 17.92% | 63 22.58% | 41 14.70% |
| Pit | 17 6.09% | 33 11.83% | 34 12.19% | 43 15.41% | 37 13.26% |
| Natural drainage | 6 2.15% | 23 8.24% | 30 10.75% | 49 17.56% | 34 12.19% |
| Holding tank | 1 0.36% | 4 1.43% | 4 1.43% | 6 2.15% | 2 0.72% |

TABLE 22. Professional consultant and lagoon $2 \times 2$ table

| | | Professional consultant | |
|---|---|---|---|
| | | 1 | 0 |
| Lagoon | 1 | 34 | 109 |
| | 0 | 10 | 126 |

There exists a special relationship between

$$\pi_{ij}, i = 1, ..., r; \ j = 1, ...c,$$

and

$$\tau_{hk}, h = 1, ...R; \ k = 1, ..., C$$

that is,

$$\pi_{ij} = \sum_{h,k:W_i=1\&Y_j=1} \tau_{hk}. \tag{V.1.1}$$

A similar relationship holds between

$$m_{ij}, i = 1, ..., r; \ j = 1, ..., c$$

and

$$n_{hk}, h = 1, ..., R; \ k = 1, ..., C.$$

A "joint table" gives the cross-classification of responses to each possible set of item responses for W and Y. This is similar to the joint table described in Bilder, Loughin, and Nettleton (2000); Bilder and Loughin (2004). Table 23 gives the joint table for the Kansas farmer data. For example, 15 farmers picked professional consultant as their only source of veterinary information and lagoon as their only waste storage method. Cell counts in the joint table are denoted by $n_{hk}$ and the corresponding probability is denoted by $\tau_{hk}$. Multinomial sampling is assumed within the entire joint table; thus $\sum_{hk} \tau_{hk} = 1$.

TABLE 23. Joint table for the Kansas farmer data. The $Y_j$ and $W_i$ items correspond to the same ordering of the column and row items listed in Table 21

| | | | $Y_1$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $Y_2$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | | | $Y_3$ | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| | | | $Y_4$ | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| | | | $Y_5$ | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 2 | 3 | 0 | 2 | 0 | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 2 | 12 | 19 | 7 | 12 | 0 | 1 | 0 | 4 | 3 | 0 | 0 | 0 | 0 | 3 | 5 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 4 | 7 | 7 | 2 | 4 | 0 | 6 | 2 | 4 | 2 | 1 | 1 | 0 | 1 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 |
| 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 11 | 9 | 22 | 0 | 10 | 3 | 2 | 2 | 13 | 1 | 1 | 3 | 2 | 0 | 3 | 3 | 15 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 2 | 3 |
| 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 0 | 0 | 2 | 1 | 0 | 2 | 0 | 1 | 0 | 2 | 4 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $W_1$ | $W_2$ | $W_3$ | $W_4$ | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Sparseness is usually the norm for the joint tables. The number of cells in the joint table is $2^{r+c}$, which can be quite large even for small values of $r$ and $c$. For the Kansas farmer data example, there are $2^9 = 512$ cells and 434 have zeros in them. This table sparseness can have a detrimental effect on model based testing approaches that need to estimate all $\tau_{hk}$ from the joint table. Even when the model based approaches converged after lengthy iteration, the interpretation of joint table is very complicated and not much of interest.

Thus we focus on the marginal table constructed by $m_{ij}$ as the number of observed responses to $W_i = 1$ and $Y_j = 1$. Table 21 is an example of marginal table. The marginal probability of $\pi_{ij} = \{W_i = 1; Y_j = 1\}$ can be estimated by its maximum likelihood estimate (MLE) as

$$\hat{\pi}_{ij} = \frac{m_{ij}}{n},$$

where $n = \sum\sum m_{ij}$. The hypotheses for test of marginal independence are

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \; for \; i = 1, \ldots r \; and \; j = 1, \ldots, c,$$

vs

$$H_1 : At\ least\ one\ equality\ does\ not\ hold.$$

Here $\pi_{ij} = P(W_i = 1, Y_j = 1)$, $\pi_{i+} = P(W_i = 1)$ and $\pi_{+j} = P(Y_j = 1)$. This specifies marginal independence between each $W_i$ and $Y_j$ pair. The hypotheses can also be written in the way of odds ratio. Consider the $rc$, $2 \times 2$ pairwise item response tables formed for each $W_i$ and $Y_j$ pair (analogous to Table 22) and suppose the cells contain probabilities for each $W_i$ and $Y_j$ pair; i.e., $P(W_i = 1, Y_j = 1) = \pi_{ij}$, $P(W_i = 1, Y_j = 0) = \pi_{i+} - \pi_{ij}$, $P(W_i = 0, Y_j = 1) = \pi_{+j} - \pi_{ij}$ and $P(W_i = 0, Y_j = 0) = 1 - \pi_{i+} - \pi_{+j} + \pi_{ij}$. If none of these cells have 0 probability, the pairwise marginal independence hypotheses can be written as,

$$OR_{WY,ij} = 1, \; for \; i = 1, \ldots r \; and \; j = 1, \ldots, c,$$

where

$$OR_{WY,ij} = \frac{\pi_{ij}(1 - \pi_{i+} - \pi_{+j} + \pi_{ij})}{(\pi_{i+} - \pi_{ij})(\pi_{+j} - \pi_{ij})}.$$

Therefore, SPMI represents simultaneous independence in the $rc$, $2 \times 2$ pairwise item response tables formed for each $W_i$ and $Y_j$ pair. The MLE for $\pi_{i+}$ and $\pi_{+j}$ are $\hat{\pi}_i = \frac{m_{i+}}{n}$ and $\hat{\pi}_j = \frac{m_{+j}}{n}$.

Let $\mathbf{m} = (m_{11}, m_{12}, \ldots, m_{rc})'$ and $\mathbf{n} = (n_{11}, n_{12}, \ldots, n_{2^r2^c})'$. Also, let $\mathbf{G}$ be a $r \times 2^r$ matrix with columns containing all possible values of $(W_1, \ldots, W_r)'$, and let $\mathbf{H}$ be a $c \times 2^c$ matrix with columns containing all possible values of $(Y_1, \ldots, Y_c)'$.

TABLE 24. $2 \times 2$ margin table

| | | Column Response $j$ | | |
|---|---|---|---|---|
| | | 1 | 0 | |
| Row Response $i$ | 1 | $\pi_{ij}$ | $\pi_{i+} - \pi_{ij}$ | $\pi_{i+}$ |
| | 0 | $\pi_{+j} - \pi_{ij}$ | $1 - \pi_{i+} - \pi_{+j} + \pi_{ij}$ | $1 - \pi_{i+}$ |
| | | $\pi_{+j}$ | $1 - \pi_{+j}$ | 1 |

For example, the column headers in Table 23 form $\mathbf{H}$ for the sources of veterinary information multiple-response categorical variable. Then $(\mathbf{G} \bigotimes \mathbf{H})\mathbf{n} = \mathbf{m}$, where $\bigotimes$ denotes the Kronecker product. This can also be written equivalently as $(\mathbf{G} \bigotimes \mathbf{H})\hat{\boldsymbol{\tau}} = \hat{\boldsymbol{\pi}}$, where $\hat{\boldsymbol{\tau}} = \mathbf{n}/n$ and $\hat{\boldsymbol{\pi}} = \mathbf{m}/n$. Define $\hat{\pi}^R = (\hat{\pi}_{1+}, \ldots, \hat{\pi}_{r+})'$ and $\hat{\pi}^C = (\hat{\pi}_{+1}, \ldots, \hat{\pi}_{+c})'$. For each marginal Table 24, let

$$V_{ij}^R = (\hat{\pi}_{ij}, \hat{\pi}_{i+} - \hat{\pi}_{ij}, \hat{\pi}_{+j} - \hat{\pi}_{ij}, 1 - \hat{\pi}_{i+} - \hat{\pi}_{+j} + \hat{\pi}_{ij})$$

$$V_{ij}^C = (\hat{\pi}_{ij}, \hat{\pi}_{+j} - \hat{\pi}_{ij}, \hat{\pi}_{i+} - \hat{\pi}_{ij}, 1 - \hat{\pi}_{i+} - \hat{\pi}_{+j} + \hat{\pi}_{ij})$$

The test statistics is constructed as

$$C_{ij\Delta} = (n-1)SS\hat{B}_{ij}^R / SS\hat{T}_{ij}^R + (n-1)SS\hat{B}_{ij}^C / SS\hat{T}_{ij}^C,$$

and $SS\hat{T}^R$, $SS\hat{W}^R$, $SS\hat{B}^R$ and $SS\hat{T}^C$, $SS\hat{W}^C$, $SS\hat{B}^C$ are defined as:
$SS\hat{T}_{ij}^R = V_{ij}^{R'} T_{ij} V_{ij}^R$, $SS\hat{T}_{ij}^C = V_{ij}^{C'} T_{ij} V_{ij}^C$, with $T_{ij} = J_{2\times 2} \bigotimes \Delta$;
$SS\hat{W}_{ij}^R = V_{ij}^{R'} W_{ij} V_{ij}^R$, $SS\hat{W}_{ij}^C = V_{ij}^{C'} W_{ij} V_{ij}^C$, with $W_{ij} = diag(1/\pi_{i+}, 1/(1-\pi_{i+})) \bigotimes \Delta$;
$SS\hat{B}_{ij}^R = V_{ij}^{R'} B_{ij} V_{ij}^R$, $SS\hat{B}_{ij}^C = V_{ij}^{C'} B_{ij} V_{ij}^C$, with $B_{ij} = n^{-1} T_{ij} - W_{ij}$
and then the overall statistics is,

$$C_\Delta = \sum_{i=1}^{r} \sum_{j=1}^{c} C_{ij\Delta}. \tag{V.1.2}$$

Using the joint asymptotic normality of $\hat{\boldsymbol{\tau}}$ and the delta method, it can be shown that $C_\Delta$ has an asymptotic $\chi^2_{2rc(n-rc)}$ distribution.

## V.2 DISTRIBUTION OF TEST STATISTICS BY BOOTSTRAP METHODS

The asymptotic distribution of $C_\Delta$, based upon the convergence of the multinomial distribution in the expanded table, is a multivariate normal. Because this table is of dimension

$2^r \times 2^c$, which can be very large even for relative small values of $r$ and $c$, extremely large sample sizes may be required to make the asymptotic distribution a reasonable approximation. Alternative methods need to be considered to estimate the distribution of $C_\Delta$. The bootstrap (Efron and Tibshirani, 1993) is a computational technique that can be used to estimate the finite-sample sampling distribution of a statistic. In this section, nonparametric bootstrapping and other alternatives are explored for estimating the p-values.

### V.2.1 Nonparametric Bootstrap

The sampling distribution of $C_\Delta$ can be approximated using a nonparametric bootstrap method. To re-sample under independence of W and Y, $W_s$ and $Y_s$ are independently re-sampled with replacement from the data set. The test statistics calculated for the $b^{th}$ re-sample of size $n$ is denoted by $C^*_\Delta$. The p-value is calculated as $B^{-1}\sum_b I(C^*_\Delta \geq C_\Delta)$, where $B$ is the number of re-samples taken and $I()$ is the indicator function.

### V.2.2 Bootstrap P-Value Combination Methods

Each $C_{ij\Delta}$ gives a test for independence between each $W_i$ and $Y_j$ pair for $i = 1,...,r, j = 1,...,c$. The p-values from each of these tests can be combined to form a new statistic, $\tilde{p}$. Combination methods can be the product of the p-values or the minimum of the p-values. Since the $rc$ different tests are likely to be correlated, the usual p-values combination methods based upon the independence of the p-values are not appropriate. The bootstrap can be used to approximate the sampling distribution of $\tilde{p}$. Resamples for the bootstrap procedure are taken the same way as described before. The p-value for the combined test is calculated as $B^{-1}\sum_b I(\tilde{p}^*_b \leq \tilde{p})$, where $\tilde{p}^*_b$ is the combined p-value calculated for the $b^{th}$ re-sample.

### V.2.3 Bonferroni Adjustment

As an alternative to the bootstrap procedures, a Bonferroni adjustment can be applied to $C_\Delta$. $H_0$ is rejected if any $C_{ij\Delta}$ is greater than the $1-\alpha/(rc)$ quantile of a $\chi^2_1$ distribution. A Bonferroni adjusted p-value can also be calculated by multiplying the minimum of the $rc$ p-values by $rc$. The advantage of a Bonferroni adjustment approach is that it can be calculated without knowing the joint table of responses. The disadvantage of this approach is that for moderate to large $r$ and $c$ values, the Bonferroni adjustment to the critical value may be severe leading to a conservative test.

### V.2.4 Post-hoc Test

If the hypothesis of marginal independence is rejected, one would want to know why it is rejected. Since $C_\Delta$ in Equation (V.1.2) is written as the sum of $rc$ different Chi-square test statistics, each $C_{\Delta ij}$ can be used to determine where the rejection occurs. The individual tests can be performed using chi-square approximation or the estimated sampling distribution in the proposed bootstrap procedures. This is similar to the post-hoc test in the analysis of variance for continuous data where a significant F-test is followed by multiple comparison tests.

## V.3 EMPIRICAL COMPARISONS

Thomas and Dacady (2000) suggested a Pearson statistic,

$$X_S^2 = n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\hat{\pi}_{ij}-\hat{\pi}_{i.}\hat{\pi}_{.j})}{\hat{\pi}_{i.}\hat{\pi}_{.j}(1-\hat{\pi}_{i.})(1-\hat{\pi}_{.j})}. \qquad \text{(V.3.1)}$$

Bilder et al. (2000) showed that the second order adjustment $rc\chi_S^2/\sum_{p=1}^{rc}\lambda_p^2$ can be approximated by a $\chi^2$ random variable with degree freedom of $r^2c^2/\sum_{p=1}^{rc}\lambda_p^2$, where $\lambda_i{}'s$ are eigenvalues of certain matrix. This Pearson statistics is used as a reference statistics for comparisons.

We have performed a simulation study to compare which test in Section V.2 holds the correct size under a range of different situations and has power to detect various alternative hypotheses. Each simulation uses 500 data sets and bootstrap method uses B=1000 bootstrap samples. The significance level is set as 0.05.

### V.3.1 Type I Error

For simulating of data under the null hypothesis, the $OR_{WY,ij}$ are set to 1 for each pair of $W_i$ and $Y_j$, $i=1,...,r; j=1,...,c$. Odds ratios between $W_i$ and $W_i'$ pair and each $Y_j$, $Y_j'$ pair are calculated as

$$OR_{W,ii'} = \frac{P(W_i=1 \; and \; W_{i'}=1)/P(W_i=1 \; and \; W_{i'}=0)}{P(W_i=0 \; and \; W_{i'}=1)/P(W_i=1 \; and \; W_{i'}=0)},$$

and

$$OR_{Y,jj'} = \frac{P(Y_j=1 \; and \; Y_{j'}=1)/P(Y_j=1 \; and \; Y_{j'}=0)}{P(Y_j=0 \; and \; Y_{j'}=1)/P(Y_j=1 \; and \; Y_{j'}=0)}.$$

The values of $OR_{W,ii'}$ and $OR_{Y,jj'}$ are set at values of 2 and 25 in the simulations to represent weak and strong pairwise dependence.

Table 25 and Table 26 show the estimated type I error rates for 2 × 2 marginal table simulations. The 95% expected range of estimated type I error rates for testing methods holding the correct size is $0.05 \pm 2(0.05(1-0.05)/500)^{1/2} = (0.0305, 0.0695)$.

Pearson statistics by Bilder et al. (2000) in Equation (V.3.1) mostly holds the correct size for the $OR_{W,ii'} = OR_{Y,jj'} = 2$ but rejects too often when an odds ratio of 25 is present. Quadratic entropy statistics holds the correct size for the $OR_{W,ii'} = OR_{Y,jj'} = 25$ but rejects too often when an odds ratio of 2 is present. All of the bootstrap methods generally hold the correct size at most of the times. Bonferroni adjustment holds the correct size most of the time but are too conservative sometimes.

### V.3.2 Power

A limited simulation study was performed to examine the power of the quadratic entropy statistics. We have excluded Modified Pearson's $\chi^2$ test and quadratic entropy with chi-square distribution from the power comparisons since they did not meet size conditions in Tables 25 and 26. Data were simulated with marginal probabilities of $\pi^R = (0.4, 0.5)'$, $\pi^C = (0.2, 0.3)'$; the sample size was set at n=100.

Table 27 for comparison of the empirical power indicates that the Bootstrap product of p-values has larger power than the others in most cases. Nonparametric bootstrap tends to have similar power to the bootstrap product p-value method. Bonferroni adjusted method tends to have similar power to mininmum p-value method because of their statistics' similar construction.

TABLE 25. Estimated type I errors.$\pi^R = (0.4, 0.5)'$, $\pi^C = (0.2, 0.3)'$. Bold cells correspond to estimated type I error rates outside of the 95% expected range.

| OR | n | Modified $\chi_s^2$ | QE | Bootstrap QE | Bootstrap Product P-values | Minimum P-values | Bonferroni Adjusted |
|---|---|---|---|---|---|---|---|
| 2 | 12 | **0.096** | **0.008** | 0.038 | 0.046 | 0.068 | 0.036 |
| | 25 | **0.078** | 0.034 | 0.038 | 0.042 | 0.052 | 0.038 |
| | 50 | 0.068 | **0.028** | 0.046 | 0.05 | 0.05 | 0.042 |
| | 100 | 0.05 | 0.064 | 0.052 | 0.044 | 0.046 | 0.042 |
| 25 | 12 | **0.136** | **0.008** | **0.022** | 0.042 | 0.05 | **0.018** |
| | 25 | **0.074** | 0.034 | 0.032 | **0.03** | 0.032 | **0.02** |
| | 50 | 0.066 | 0.048 | 0.054 | 0.056 | 0.058 | 0.046 |
| | 100 | 0.05 | 0.06 | 0.044 | 0.042 | 0.056 | 0.046 |

TABLE 26. Estimated type I errors. $\pi^R = (0.4, 0.5)'$, $\pi^C = (0.4, 0.5)'$. Bold cells correspond to estimated type I error rates outside of the 95% expected range.

| OR | n | Modified $\chi^2_s$ | QE | Bootstrap QE | Bootstrap Product P-values | Minimum P-values | Bonferroni Adjusted |
|---|---|---|---|---|---|---|---|
| 2 | 12 | 0.058 | **0.008** | 0.044 | 0.04 | 0.052 | **0.028** |
| | 25 | 0.048 | **0.026** | 0.048 | 0.04 | 0.048 | 0.044 |
| | 50 | 0.066 | **0.028** | 0.054 | 0.058 | 0.046 | 0.05 |
| | 100 | 0.052 | 0.064 | 0.044 | 0.038 | **0.08** | **0.08** |
| 25 | 12 | **0.08** | **0.014** | 0.048 | 0.062 | **0.086** | 0.04 |
| | 25 | **0.096** | 0.030 | 0.058 | 0.062 | 0.06 | 0.048 |
| | 50 | 0.054 | 0.046 | 0.05 | 0.04 | 0.04 | 0.036 |
| | 100 | **0.074** | 0.068 | 0.06 | 0.07 | 0.05 | 0.042 |

TABLE 27. Empirical power. $\pi^R = (0.4, 0.5)'$, $\pi^C = (0.2, 0.3)'$; n=100. Bold cells correspond to the highest power at each circumstance.

| $OR_{W,ii'} = OR_{Y,jj'}$ | $OR_{W,Y}$ | Bootstrap QE | Bootstrap Product P-values | Minimum P-values | Bonferroni Adjusted |
|---|---|---|---|---|---|
| 2 | 1.5 | 0.07 | 0.068 | **0.074** | 0.066 |
| | 2 | 0.088 | **0.1** | 0.096 | 0.092 |
| | 2.5 | 0.118 | 0.128 | **0.136** | 0.124 |
| | 3 | 0.132 | **0.152** | 0.148 | 0.136 |
| | 3.5 | 0.218 | **0.244** | 0.194 | 0.19 |
| 25 | 20 | 0.568 | **0.658** | 0.582 | 0.554 |
| | 25 | 0.574 | **0.65** | 0.582 | 0.562 |
| | 30 | 0.68 | **0.732** | 0.608 | 0.602 |
| | 35 | 0.64 | **0.71** | 0.644 | 0.622 |
| | 40 | 0.606 | **0.67** | 0.598 | 0.588 |

## V.4 APPLICATION TO THE KANSAS FARMER DATA

The testing procedures of Section V.2 are applied to the Kansas farmer data and the corresponding p-values are shown in Table 28. We have used 10,000 re-samples for the bootstrap methods. All methods indicated strong evidence against marginal independence. Using the post-hoc test outlined in Section V.2.4, the significant pairwise combinations are $(W_1,Y_1)$, $(W_2,Y_2)$, $(W_2,Y_3)$, $(W_3,Y_3)$ and $(W_3,Y_1)$ at the 0.05 significance level. If Bonferroni adjusted significance level of $0.05/20 = 0.0025$ is used instead, only $(W_1,Y_1)$=(Lagoon, Professional consultant) is significant.

TABLE 28. Testing p-values for the Kansas farmer data

| Testing Methods | P-Values |
|---|---|
| Modified Pearson's $\chi^2$ | $3.07 \times 10^{-5}$ |
| Quadratic Entropy | $2.11 \times 10^{-6}$ |
| Nonparametric Bootstrap | 0.0003 |
| Bootstrap Product of P-values | 0.0001 |
| Bootstrap Minimum P-values | 0.0027 |
| Bonferroni Adjustment | 0.0034 |

In this chapter, we provide a method to analyze multi-response data based upon Rao's quadratic entropy. The proposed methods of quadratic entropy for testing independence are counterparts to the already developed methods for single-response categorical variables. While the bootstrap methods may be the most computationally intensive of the testing methods, they most consistently hold the correct size and have higher power to detect the significance. Bonferroni adjustment provide simpler methods but they can be conservative at times. Model-based approaches to testing multiple-response data will be the focus of our future study.

# CHAPTER VI

# CLUSTERING GENE EXPRESSION DATA

DNA microarray technology has now made it possible to simultaneously monitor the gene expression levels during biological process. Elucidating the patterns hidden in the gene expression data offers a tremendous opportunity for understanding how genes are affected by disease states and cellular environments. However, the high dimensional genes and the complexity of biological structure brings great difficulty in interpreting the mass of data. A preliminary and common methodology towards addressing this challenge is the clustering technique.

As described in Chapter IV, clustering is a process of seeking a partition of given data set based upon certain features so that the data points within a group are more similar to each other than the points in different groups. Clustering can also be used to group genes according to their expressions in a set of samples. The second type of clustering is to cluster samples into homogeneous groups that may correspond to clinical syndromes or cancer types. Clustering of samples can be challenging due to the small sample volume and high genes dimensionality. The third type is subspace clustering, which is to capture the coherence exhibited by the "blocks" with gene expression matrices. Here a "block" is a sub-matrix defined by a subset of genes on a subset of samples.

There is a rich literature on cluster analysis and various techniques have been developed. Many conventional clustering methods such as *k*-means, hierarchical clustering have been adopted or directly applied to gene expression data, and also new algorithm such as graph-theoretical approaches, machine learning and neural network techniques have been proposed specifically aiming at gene expression data. Jiang, Tang, and Zhang (2004) have reviewed most of these techniques.

Although these clustering methods are often applied to clustering gene expression data, they face several new challenges in practice (Jiang et al., 2004). First, cluster analysis is typically the first step in data mining and knowledge discovery. Therefore, a good clustering algorithm should depend as little as possible on prior knowledge. However, most algorithms (except hierarchical clustering) require that the user specifies the "true" number of clusters in advance, which is usually not available before a cluster analysis is performed. Although hierarchical clustering does not need the number of clusters, it is still up to the researcher to decide where to cut the tree of clustering and decide how many groups to cluster. Second, due to the complex procedures of microarray experiments, gene expression data often contain a large amount of noise. $k$-means algorithm forces each gene into a cluster, which cause the algorithm to be sensitive to noise.

It is well known that entropy is a measure of information and uncertainty of a random variable. Hence it is natural to use entropy to measure the closeness within the cluster and minimize the overall entropy for clustering. While simply minimizing the entropy will cluster all the sample into one group, Li, Zhang, and Jiang (2004) proposed a minimum entropy clustering algorithm. First, a minimum entropy criterion was constructed on posteriori probabilities and then generalized to Havrda-Charvat's structural $\alpha$-entropy,

$$H^{\alpha}(x) = (2^{1-\alpha} - 1)^{-1}[\sum_{x} p^{\alpha}(x) - 1],$$

where $p(x)$ is the probability of variable $x$. With a nonparametric approach for estimating a posteriori probabilities, a hill-climbing iterative algorithm was then established to minimize the entropy. When $\alpha = 2$, Havrda-Charvat's structural $\alpha$-entropy becomes Gini-Simpson index. As stated in Chapter I, Rao's quadratic entropy can catch more information of clusters by implementing difference of groups in $\Delta$. The distance matrix $\Delta$ can be estimated from data as discussed in Chapter II. Using more "information" of clusters, this algorithm has the potential to have better performance than the traditional $k$-means, hierarchical methods, and the self learning minimum entropy algorithm in terms of adjusted Rand index.

We introduce minimum entropy criterion for clustering and modify it for quadratic entropy in in the next section. In Section VI.2, we estimate the posteriori probabilities following a nonparametric approach, and then propose an iterative algorithm to minimize the posteriori quadratic entropy. Section VI.3 compares the results of minimum entropy algorithm with $k$-means and hierarchical methods on both simulated and real data. We will end the chapter with some final comments.

## VI.1 MINIMUM QUADRATIC ENTROPY CLUSTERING CRITERION

In information theory, entropy is an important measure of information and uncertainty. Both Shannon entropy and Gini-Simpson entropy measure the amount of disorder in a system. Recall that

$$H_S(x) = -\sum_{x} p(x) log p(x),$$

and

$$H_G(x) = 1 - \sum_x p^2(x).$$

As observed in previous chapters, quadratic entropy is a generalized form of Gini-Simpson entropy and is defined as,

$$H_Q(x) = \sum_x \sum_{x'} d(x,x')p(x)p(x').$$

The measurements of entropy are functional of the distribution of $x$ and they do not depend on the actual values of random variable $x$ but only on the probabilities. In fact, Li and Vitányi (1997) shows that entropy is the minimum descriptive complexity of a random variable. In gene expression clustering we hope that each cluster has a low entropy so that data points in the same cluster would look similar. Hence, a straightforward minimum entropy criterion could be defined as,

$$\sum_{i=1}^{c} H(x|C_i), \tag{VI.1.1}$$

where $H(x|C_i)$ is the entropy of cluster $C_i$. This conventional minimum entropy clustering strategy seems a reasonable criterion. However it is actually not adequate for clustering because it neglects the semantic aspects of data. Data usually contain some hidden meaning, which is suggesting a modular structure in the gene regulation system. In clustering, the semantic information that we are interested in is the categories of genes. Hence we naturally assume that in cluster analysis that data are drawn from a mixed source made up with several components within each it is homogeneously statistically structured.

Li et al. (2004) proposed minimum entropy clustering criterion to reflect the relationship between data points and clusters, which is measured on a posteriori probabilities. For each cluster $C_i$, a posterior entropy can be defined as $H_x(C)$ where $C$ is the random variable of category taking values in $C_1, C_2, ..., C_c$, and $x$ is one object. For Rao's quadratic entropy, this posteriori measure becomes

$$H_{Q,x}(C) = \sum_{i=1}^{c} \sum_{j=1}^{c} d_{ij|x} p(C_i|x)p(C_j|x), \tag{VI.1.2}$$

where $\Delta_{C|x} = (d_{ij|x})$ is the distance matrix between clusters given the information of $x$. Here we compute posteriori probabilities $p(C_i|x), i = 1, ..., c$ to determine how much information has been gained. $H_{Q,x}(C)$ is maximized when $p(C_1|x), p(C_2|x), ..., p(C_c|x)$ reach certain level. In this case, the object $x$ could come from any clusters and we do not know which cluster the object $x$ should belong to. On the other hand, $H_{Q,x}(C)$ is minimized to 0 when one of the $p(C_1|x), p(C_2|x), ..., p(C_c|x)$ has value one but all the others are zero. Thus, $H_{Q,x}(C)$ can assess the dependence between objects $x$ and clusters $C$.

Li et al. (2004) suggested to integrate $x$ on the whole data space to find the clustering criterion. If using Rao's quadratic entropy, it becomes,

$$J = \int H_{Q,x}(C)p(x)dx. \qquad \text{(VI.1.3)}$$

The above quantity is actually the entropy of the random variable $C$ given the random variable $x$ and it measures how uncertain we are of $C$ on the average when we know $x$.

It is easy to prove that, for either Shannon, Gini-Simpson or quadratic entropy,

$$H(C|x) \leq H(C)$$

with equality if and only if $x$ and $C$ are independent (Li et al., 2004), which says that knowing the random variable $x$ can reduce the uncertainty in $C$ on the average unless $x$ and $C$ are independent. In the case of quadratic entropy,

$$H_Q(C) = \sum_{i=1}^{c} \sum_{j=1}^{c} d_{ij} p(C_i) p(C_j),$$

where $\Delta_C = (d_{ij})$ is the distance matrix between clusters without the information of $x$. This indicates that the minimum of $H(C|x)$ can be a good clustering criterion. This clustering criterion has been illustrated for Shannon entropy and Havrda-Charvat's structural $\alpha$-entropy in Li et al. (2004). As discussed in Chapter I and II, Rao's quadratic entropy allows to specify the distance between clusters and brings in "extra" self-learning information to the clustering algorithm, and eventually improves the clustering performance. Given a data set $X = x_1, ..., x_c$, the minimum quadratic entropy clustering (MQEC) criterion is defined as,

$$J = \int H_{Q,x}(C)p(x)dx = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} \sum_{j=1}^{c} d_{ij|k} p(C_i|x_k) p(C_j|x_k), \qquad \text{(VI.1.4)}$$

where $\Delta_{C|x_k} = (d_{ij|k})$ is the distance matrix between clusters given the information of $x_k$. Besides bringing in prior information of clusters by specifying $\Delta_{C|x_k}$, quadratic entropy has another merit of recursivity. Suppose random variable $C$ has the distribution $P = (p_1, p_2, ..., p_c)$. Let us write $H_Q(C|x)$ as $H_{Q,c}(p_1, p_2, ..., p_c)$, then

$$H_{Q,c}(p_1, p_2, ..., p_c) = H_{Q,c-1}(p_1, p_2, ..., p_c) + g(p_1, p_2) H_{Q,2}(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2})$$

holds for all $c \geq 3$ (Kapur, 1994). This recursive property allows us to develop clusters when there exhibits a nesting relationship between different clusters.

## VI.2 MINIMUM QUADRATIC ENTROPY CLUSTERING ALGORITHM

By employing Equation (VI.1.4) as the clustering criterion, the clustering algorithm can be developed at three steps: (1) estimating $p(C|x)$; (2) defining the matrix $\Delta_{C|x}$; and (3) minimizing $J = \int H_{Q,x}(C)p(x)dx$.

### VI.2.1 Estimation of Posterior Probabilities

To estimate the posterior probability $p(C|x)$, we could employ some parametric method. However, the choice of any particular distribution could lead to a very poor representation of the data if the data have a complex structure. We therefore apply a nonparametric method for estimating the posterior probability. There are two kinds of nonparametric techniques, Parzen density estimation and $k$-nearest neighbor density estimation (Devroye and Gyötfi, 1985). They are fundamentally similar with some different statistical properties. In what follows, we give a brief overview of Parzen density estimate and $k$-nearest neighbor density estimate.

Consider estimating the value of a density function at a point $x$; a small window $R(x)$ can be set up around $x$ and the probability mass of $R(x)$ can be approximated by $p(x) \cdot v$, where $v$ is the volume of $R(x)$. On the other hand, the probability of $R(x)$ can also be estimated by drawing a large number (say $n$) of sample $p(x)$, counting the number of samples falling in $R(x)$, say $m$ and computing as $m/n$. Equating these two probabilities, we obtain an estimate of the density function as

$$p(x) = \frac{m}{n \cdot v}. \tag{VI.2.1}$$

If we fix the volume $v$ and let $m$ be a function of $x$, we obtain Parzen density estimate; if we fix $m$ and let $v$ be a function of $x$, we have the $k$-nearest neighbor density estimate.

By Bayes's rule, we have

$$p(C_i|x) = \frac{p(C_i)p(x|C_i)}{p(x)}.$$

We may use $n_i/n$ as an estimator of $p(C_i)$, where $n_i$ is the number of points in cluster $C_i$. If Parzen density estimate is employed, we have the posterior probability as,

$$p(C_i|x) = \frac{\frac{n_i}{n} \cdot \frac{m(x|C_i)}{n_i \cdot v}}{\frac{m(x)}{n \cdot v}} = \frac{m(x|C_i)}{m(x)}. \tag{VI.2.2}$$

Figure 7 is an illustration of Parzen density estimation. For data point $x = a$, for a small window R(x) around $x = a$,

$$p(C_1|x = a) = \frac{m(x = a|C_1)}{m(x = a)} = \frac{8}{12},$$

and

$$p(C_2|x = a) = \frac{m(x = a|C_2)}{m(x = a)} = \frac{4}{12}.$$

Thus the estimate of $p(C_i|x)$ is just the ratio between the number of samples from cluster $C_i$ and the number of all samples in the local region $R(x)$. The MQEC becomes

$$J = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{c}\sum_{j=1}^{c} d_{ij|k} p(C_i|x_k) p(C_j|x_k) = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{c}\sum_{j=1}^{c} d_{ij|k} \frac{m(x|C_i)}{m(x)} \frac{m(x|C_j)}{m(x)}. \qquad \text{(VI.2.3)}$$

If $k$-nearest neighbor estimate is used, we obtain

$$p(C_j|x) = \frac{\frac{n_j}{n} \cdot \frac{m}{n_j \cdot v(x|C_j)}}{\frac{m}{n \cdot v(x)}} = \frac{v(x)}{v(x|C_j)} \qquad \text{(VI.2.4)}$$

Similarly, we can get a corresponding MQEC criterion.

### VI.2.2 Estimation of $\Delta_{C|x}$

In this section, we propose methods to calculate distance matrix $\Delta_{C|x} = (d_{ij|x})$ so that MQEC in Equation (VI.1.4) can be estimated. This matrix $\Delta$ should be a measure of distance between clusters given each data point $x$. Traditional distance measures such as Euclidean distance and Manhattan distance can be used to measure the distance between each data point; and to measure the distance between two clusters $C_1$ and $C_2$, we can use one of the following distances:

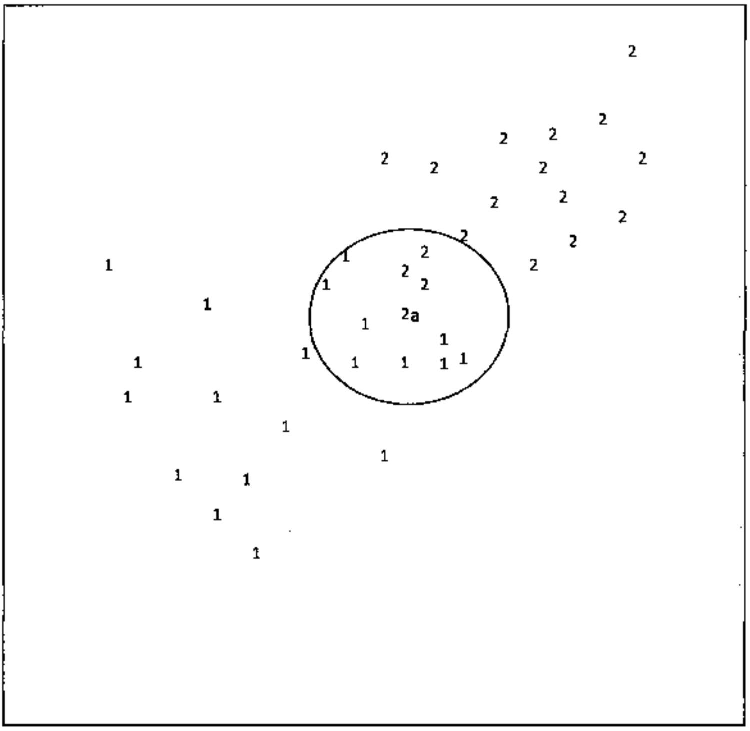
- The maximum distance between elements of each cluster (also called complete linkage clustering):

$$max\{d(x,y) : x \in C_1, y \in C_2\};$$

- The minimum distance between elements of each cluster (also called single linkage clustering):

$$min\{d(x,y) : x \in C_1, y \in C_2\};$$

*FIG. 7. An illustration of Parzen density estimation*



- The average distance between elements of each cluster (also called average linkage clustering):

$$\frac{1}{n_1 n_2} \sum_{x \in C_1} \sum_{y \in C_2} d(x,y);$$

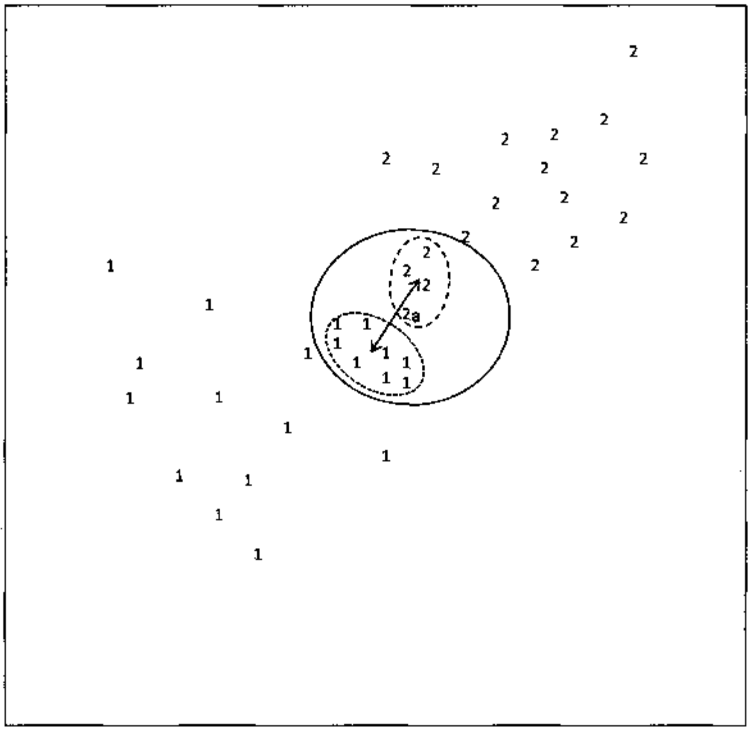
- The sum of all intra-cluster variance.

In what follows, we use Euclidean distance to measure the distance between data points and use average linkage to define the distance between clusters. Figure 8 is an illustration for measuring $\Delta$ in the previous example.

### VI.2.3 Minimization of Quadratic Entropy

In this section, we develop a clustering algorithm to optimize the MQEC in Equation (VI.2.3) with Parzen density estimation. However it is not suitable for directly clustering the data because we can minimize $H_Q(C|x)$ to 0 by simply clustering all data points into one group. Such a solution generally interferes with finding the practically useful partitions. Hence, instead of directly clustering, we use an iterative algorithm to reduce the entropy of an initial partition given by another clustering methods (e.g. $k$-means, hierarchical clustering). This hill-climbing type algorithm starts with some initial configuration, and a standard rearrangement is applied to the data set such that the objective function is improved (the MQEC is reduced); the rearranged partition then becomes the new configuration and the process is continued until no further improvement can be made. This process is illustrated in Algorithm 1.

*FIG. 8. An illustration of $\Delta_{C|x}$ estimation*



```
Input: A data set containing $n$ objects, the number of clusters $c$ and an initial
        partition given by $k$-means clustering method.
Output: A set of at most $c$ clusters that locally minimizes the entropy.
repeat
    for every objects x in the data set do
        if $C_j$ containing most of the neighbors of x is different from the current
        cluster $C_i$ of x then
            $h \leftarrow \sum_y (H'_{Q,y}(C) - H_{Q,y}(C))$
            where y are neighbors of x, and x is also regarded as the neighbor of
            itself. $H_{Q,y}(C)$ and $H'_{Q,y}(C)$ are the entropy associated with y before and
            after assigning x to the cluster $C_j$, respectively.
        end
        if $h < 0$ then
            assign x to the cluster $C_j$
        end
    end
until no change ;
```

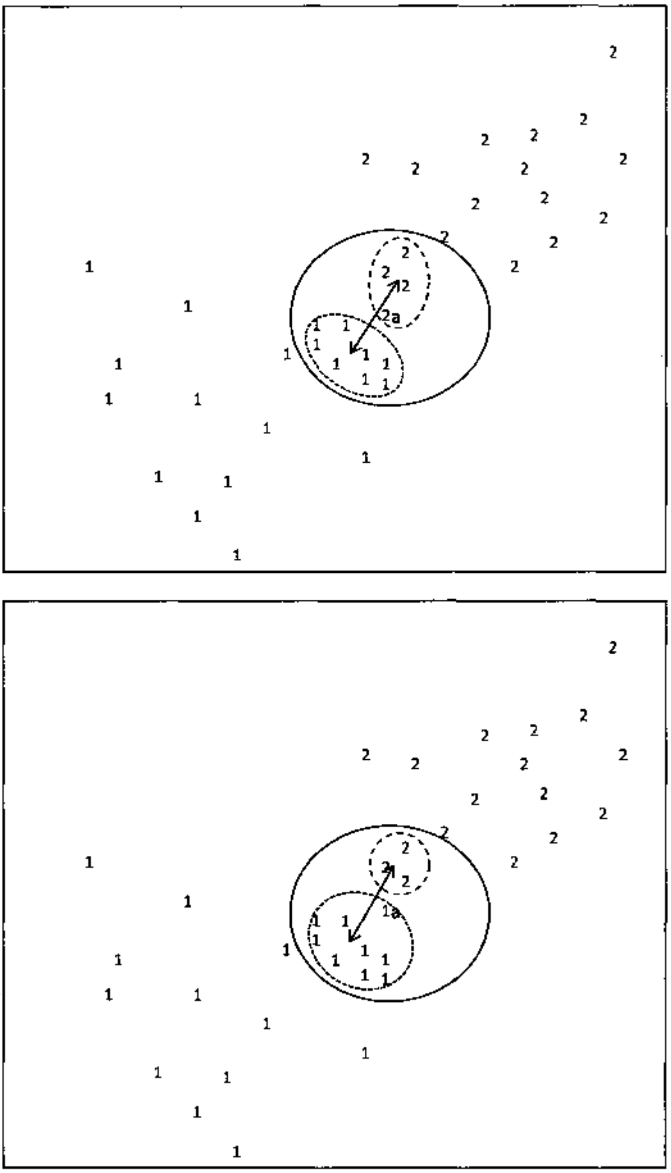
**Algorithm 1**: Minimum Quadratic Entropy Clustering Algorithm

Since the total entropy decreases in every step and the quadratic entropy is bounded by 0, the Algorithm 1 converges after a sufficient number of iterations. In the experiments of both simulation data and real gene expression data, it was found that the number of iteration is often very small, usually less than 10.

Note that this algorithm could give a set of fewer than $c$ clusters when a cluster migrate into another cluster to reduce MQEC during the iterations. This is different from most other clustering methods, which always return a given number of clusters.

Figure 9 is an illustration of one iteration in the hill-climbing algorithm. For point $x = a$, we can estimate the posterior probability with Parzen estimation, estimate the matrix $\Delta_{C|x}$, and then we can calculate the entropy measure of $H(C_1|x = a)$ and $H(C_2|x = a)$. Repeat this for each data point, we will get the $J$ as in Equation (VI.2.3). As cluster $C_1$ contains most of the neighbors of $x = a$, which is different from the current cluster $C_2$ of x, then we assign $x = a$ to cluster $C_1$, recalculate $H(C_1|x = a)$ and $H(C_2|x = a)$ and also the value of $J'$ in Equation (VI.2.3). If $J \leq J'$, then keep $x$ as in original cluster $C_2$; if $J' < J$, then assign $x$ to cluster $C_1$. And the same process can be repeated to each data point in the data set until there is no reduction of $J$ can be found.

*FIG. 9. An illustration of MQEC iteration algorithm*



## VI.3 EXPERIMENTS AND RESULTS

In this section, we report the results of using minimum quadratic entropy criterion on a simulated data and two real gene expression data sets. To assess the quality of algorithm, we compare the clustering results with true class information or gene functional categories by adjusted Rand index (Hubert and Arabie, 1985; Steinley, 2004) as the measure of agreement.

The adjusted Rand index lies between 0 and 1. When the clustering results perfectly agree with true clusters, the adjusted Rand index is 1; when the clustering is random, it has the minimum value of 0. A larger adjusted Rand index means a higher agreement between new cluster with true clusters. Another advantage is that adjusted Rand index can be used to measure the agreement even when the number of cluster results $D$ is different from number of true clusters $C$ (See Appendix B).

### VI.3.1 Simulated Data

To illustrate of the new algorithm, we generate data similar to Li et al. (2004). Given means as $[0,0]$ and $[2,2]$, variance-covariance matrices as $[\begin{array}{cc} 1 & 0.3 \\ 0.3 & 1 \end{array}]$ and $[\begin{array}{cc} 1 & -0.3 \\ -0.3 & 1 \end{array}]$, a two-dimensional data are simulated to follow Gaussian distribution. Then we compare the adjusted Rand index between minimum quadratic entropy algorithm and the $k$-means clustering method.

TABLE 29. Adjusted Rand index on the simulation data

| Cluster | Hierarchical | $k$-means | MQEC_Gini | MQEC_QE |
|---|---|---|---|---|
| 2 | 0.560538 | 0.571655 | 0.712 | 0.735 |
| 3 | 0.293405 | 0.408945 | 0.625 | 0.674 |
| 4 | 0.267715 | 0.309896 | 0.365 | 0.374 |
| 5 | 0.369879 | 0.241148 | 0.425 | 0.457 |
| 6 | 0.288235 | 0.208554 | 0.497 | 0.516 |
| 7 | 0.245126 | 0.170305 | 0.631 | 0.678 |
| 8 | 0.211185 | 0.153727 | 0.597 | 0.624 |
| 9 | 0.185435 | 0.139683 | 0.511 | 0.589 |
| 10 | 0.166678 | 0.13313 | 0.487 | 0.512 |

Table 29 lists the adjusted Rand index achieved by Hierarchical clustering, $k$-means, and minimum quadratic entropy algorithm based upon two distance matrices. Both quadratic entropy algorithm improve the initial partitions given by $k$-means. When the specified number of clusters is correct, the minimum quadratic entropy have some improvement from the $k$-means. When thc specified number of clusters are not correct, which is often the case, the minimum quadratic entropy still performs much better than the $k$-means and hierarchical clustering methods.

### VI.3.2 Real Example I: Yeast Galactose Data

We used two gene expression data to test MQEC algorithm. The first data is the yeast galactose data with 205 genes on 20 experiments from Yeung, Medvedovic, and Bumgarner (2003), whose expression categories correspond to four functional categories in the Gene Ontology listing. We used the four categories as the external knowledge to test the clustering methods. Before clustering, we normalized the data for each gene to have mean 0 and and variance 1 across experiments.

TABLE 30. Adjusted Rand index on the yeast galactose data

| Cluster | Hierarchical | $k$-means | MQEC_Gini | MQEC_QE |
|---|---|---|---|---|
| 4 | 0.769948 | 0.705322 | 0.938605 | 0.948752 |
| 5 | 0.867503 | 0.937955 | 0.94702 | 0.956223 |
| 6 | 0.859209 | 0.831913 | 0.946917 | 0.956121 |
| 7 | 0.860414 | 0.757277 | 0.937789 | 0.945148 |
| 8 | 0.855887 | 0.745214 | 0.937789 | 0.945148 |
| 9 | 0.689918 | 0.673045 | 0.842092 | 0.863605 |
| 10 | 0.660022 | 0.665898 | 0.842092 | 0.863605 |

The experimental results are listed in Table 30. Clearly the minimum entropy algorithm based upon Gini-Simpson entropy performs better than $k$-means and hierarchical clustering methods. When the specified number of clusters are far from the true number of clusters, quadratic entropy criterion is even better than Gini-Simpson entropy. The minimum entropy criterion algorithm with quadratic entropy achieves a very high adjusted Rand index ($> 0.9$), which indicates that this algorithm can effectively cluster genes into the same functional category according the expression levels. This algorithm can produce a reasonable clustering even when the specified number of clusters is larger than the true number (i.e. 4 in this case). One possible reason is, when the specified number of clusters is larger than the correct number, the minimum quadratic entropy algorithm can use the "extra" clusters to identify outliers and thus improve the quality of the final partition. In this sense, this algorithm is capable of extracting useful information and detect outliers.

### VI.3.3 Real Example II: Yeast Cell Cycle Data

The second data set is the yeast cell cycle data set which contains approximately 6000 genes expressions data over two cell cycles. Yeung and Ruzzo (2001) extracted 384 genes according to the peak time of genes, which were categorized into five phases of cell cycles by peak times. Again, the data was normalized to have mean 0 and variance 1 across each cell cycle. We took the five phases as the external knowledge and did the clustering. The results are listed in Table 31.

TABLE 31. Adjusted Rand index on the yeast cell cycle data

| Cluster | Hierarchical | $k$-means | MQEC_Gini | MQEC_QE |
|---|---|---|---|---|
| 5 | 0.482783 | 0.493835 | 0.483573 | 0.489987 |
| 6 | 0.480931 | 0.45666 | 0.470363 | 0.478038 |
| 7 | 0.478222 | 0.479874 | 0.485095 | 0.491022 |
| 8 | 0.480002 | 0.358697 | 0.453583 | 0.461202 |
| 9 | 0.364068 | 0.44738 | 0.466551 | 0.476355 |
| 10 | 0.343015 | 0.322426 | 0.486782 | 0.490753 |

For yeast cell cycle data, the MQEC algorithm with Gini-Simpson entropy and quadratic entropy still work better than $k$-means and hierarchical clustering methods, especially when the specified number of clusters is far from the true number of clusters. Quadratic entropy criterion achieves higher adjusted Rand index than Gini-Simpson entropy. However, all of them achieved low adjusted Rand indexes and quadratic entropy does not improve the performance significantly. This does not necessarily mean that the MQEC algorithm performed poorly, but maybe because that the peak time may not be the best external criterion due to its lack of strong correlation with expression level (functional categories). It is used here because no better external information is available about the subset of genes.

In this chapter we proposed MQEC method in clustering gene expression data by implementing Rao's quadratic entropy in minimum entropy criterion proposed by Li et al. (2004). With a nonparametric approach for estimating a posteriori probabilities and a locally estimated difference matrix, an efficient iterative algorithm is used to minimize the entropy. The simulated data and two real gene expression data sets show that our new method performs significantly better than $k$-means, hierarchical clustering, and also better than minimum entropy criterion with Gini-Simpson entropy. It is seen that this algorithm performs very well even when the correct number of clusters is unknown and it is also capable of effectively identifying outliers.

# BIBLIOGRAPHY

Agresti, A. and Agresti, B. F. (1978), "Statistical analysis of qualitative variation," *Sociological Methodology*, 9, 204–237.

Agresti, A. and Liu, I. (1999), "Modeling a categorical variable allowing arbitrary many category choices," *Biometrics*, 55, 936–943.

— (2001), "Strategies for modeling a categorical variable allowing multiple category choices," *Sociological Methods and Research*, 29, 403–434.

Anderson, R. J. and Landis, J. T. (1980), "Cataonva for multidimensional contigency tables: nominal scale response," *Communication in Statistics*, 9, 1191–1206.

Berger, R. L. and Hsu, J. C. (1996), "Bioequivalence trials, intersection-union tests and equivalence confidence sets," *Statistical Science*, 11, 283–319.

Bhattacharyya, A. (1943), "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, 35, 99–109.

Bilder, C. R. and Loughin, T. M. (2002), "Testing for conditional multiple marginal independence," *Biometrics*, 58, 200–208.

— (2004), "Testing for marginal independence between two categorical variables with multiple responses," *Biometrics*, 60, 241–248.

— (2007), "Modeling association between two or more categorical variables that allow for multiple category choices," *Communications in Statistics. Theory and Methods*, 36, 433–451.

— (2009), "Modeling multiple-response categorical data from complex surveys," *The Canadian Journal of Statistics*, 37, 553–570.

Bilder, C. R., Loughin, T. M., and Nettleton, D. (2000), "Multiple marginal independence testing for pick any/c variables," *Communications in Statistics: Simulation and Computation*, 29, 1285–1316.

Davis, R. B. (1980), "Algorithm as 155: The distribution of a linear combination of $\chi^2$ random variables," *Applied Statistics*, 29, 323–333.

Devroye, L. and Gyötfi (1985), *Nonparametric density estimation: The L1 View (Wiley Series in Probability and Statistics)*, New York: John Wiley & Sons.

Efron, B. and Tibshirani, R. J. (1993), *An introduction to the bootstrap*, vol. 57 of *Monographs on Statistics and Applied Probability*, New York: Chapman and Hall.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, 95, 14863–14868.

Fritsch, K. S. and Hsu, J. C. (1999), "Multiple comparison of entropies with application to dinosaur biodiversity," *Biometrics*, 55, 1300–1305.

Hubert, L. and Arabie, P. (1985), "Comparing partitions," *Journal of Classification*, 2, 193–218.

Izsák, J. and Papp, L. (2000), "A link between ecological diversity indices and measures of biodiversity," *Ecological Modelling*, 130, 151–156.

Jiang, D., Tang, C., and Zhang, A. (2004), "Cluster analysis for gene expression data: a survey," *Knowledge and Data Engineering, IEEE Transactions on*, 16, 1370 – 1386.

Kapur, J. N. (1994), *Measures of information and their applications.*, John Wiley & Sons: New York.

Karlin, S., Kenett, R., and Bonne-Tamir, B. (1979), "Analysis of biochemical genetic data on jewish populations ii. results and interpretations of heterogeneity indices and distance measures with respect to standards," *American Journal of Human Genetics*, 31, 341–365.

Lewontin, R. C. (1972), "The appointionment of human diversity," *Evolutionary Biology*, 6, 381–398.

Li, H., Zhang, K., and Jiang, T. (2004), "Minimum entropy clustering and applications to gene expression analysis," *Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE*, 142 – 151.

Li, M. and Vitányi, P. (1997), *An introduction to Kolmogorov complexity and its applications*, 3rd ed., New York: Springer.

Light, R. J. and Margolin, B. H. (1971), "An analysis of variance for categorical data," *Journal of the American Statistical Association*, 66, 534–544.

Liu, Z. J. (1991), "Bootstrapping one way analysis of Rao's quadratic entropy," *Communications in Statististics*, 20, 1683–1703.

Liu, Z. J. and Rao, C. R. (1995), "Asymptotic distribution of statistics based on quadratic entropy and bootstrapping," *Journal of Statistical Planning and Inference*, 43, 1 – 18.

Loughin, T. M. and Scherer, P. N. (1998), "Testing for association in contigency tables with multiple column responses," *Biometrics*, 54, 630–637.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, London: Academic Press.

Nayak, T. K. (1986a), "Sampling distributions in analysis of diversity," *Sankhyā (Statistics). The Indian Journal of Statistics. Series B*, 48, 1–9.

— (1986b), "An analysis of diversity using Rao's quadratic entropy," *Sankhyā (Statistics). The Indian Journal of Statistics. Series B*, 48, 315–330.

Nayak, T. K. and Gastwirth, J. L. (1989), "The use of diversity analysis to assess the relative influence of factors affecting the income distribution," *Journal of Business and Economic Statistics*, 7, 453–460.

Nei, M. (1973), "Analysis of gene diversity in subdivided populations," *Proceedings of Natural Academic Science*, 70, 3321–3323.

Norman, D. (1991), *Dinosaur!*, New York: Prentice Hall.

Pavoine, S., Ollier, S., and Pontier, D. (2005), "Measuring diversity from dissimilarities with Rao's quadratic entropy: Are any dissimilarities suitable?" *Theoretical Population Biology*, 67, 231–239.

Pielou, E. C. (1975), *Ecological Diversity*, New York: John Wiley.

Punj, G. and Stewart, D. W. (1983), "Cluster analysis in marketing research: reviews and suggestions for application," *Journal of markteting Research*, 20, 134–148.

Rao, C. R. (1977), "Cluster analysis applied to a study of race mixture in human populations," *Proceedings of Michigan University Symposium*, 175–197.

— (1982a), "Gini-Simpson index of diversity: a characterization, generalization and applications," *Utilitas Mathematica*, 21, 273–282.

— (1982b), "Diversity: Its measurement, decomposition, apportionment and analysis," *Sankhyā: The Indian Journal of Statistics, Series A*, 44, 1–22.

— (1982c), "Diversity and dissimilarity coefficients: a unified approach," *Theoretical Population Biology*, 21, 24–43.

Rao, C. R. and Boudreau, R. M. (1984), "Diversity and Cluster Analyses of Blood Group Data on Some Human Populations," *Human Population Genetics: The Pittsburgh Symposium (1984)*, 331–362.

Romesburg, C. (2004), *Cluster Analysis for Researchers*, North Carolina: Lulu Press.

Scott, J. (1988), "Social network analysis," *Sociology*, 22, 109–127.

Sen, A. (1973), *On Economic Inequality*, London: Oxford Univerisity Press (Clarendon).

Sheehan, P. M., Fastovsky, D. E., Hoffmann, R. G., Berghaus, C. B., and Gabriel, D. L. (1991), "Sudden extinction of the dinosaurs: Latest cretaceous, Upper Great Plains, U.S.A." *Science*, 254, 835–839.

Steinley, D. (2004), "Properties of the hubert-arabie adjusted rand index," *Psychological Methods*, 9, 386–396.

Stokes, M. E., Davis, C. S., and Koch, G. G. (2005), *Categorical Data Analysis Using The SAS®System*, 2nd ed., North Carolina: SAS Publishing.

Thomas, D. R. and Dacady, Y. J. (2000), "Analyzing categorical data with multiple responses per subject," *Statistical Society of Canada Annual Meeting 2000: Proceedings of the Survey Methods Section*, 121–130.

Yeung, K. Y., Medvedovic, M., and Bumgarner, R. E. (2003), "Clustering gene-expression data with repeated measurements," *Genome Biology*, 4, R34.

Yeung, K. Y. and Ruzzo, W. L. (2001), "Pringcipal component analysis for clutering gene expression data," *Bioinformatics*, 17, 763–774.

# APPENDIX A

# GEOMETRIC MODEL

Among ecological models for species distribution, the geometric model is one of the most compatible with the observed dinosaur data. If $p_i$ is the proportion of dinosaurs in the $i$th family, then the model is

$$p_m = \frac{k(1-k)^{m-1}}{1-(1-k)^S}, \; m = 1, ..., s \; (0 < k < 1)$$

Fritsch and Hsu (1999) showed that the sample proportions from the dinosaur data among the three stratigraphic intervals, upper and lower intervals are very similar to the geometric probability with $k = 0.6$. So we generated data from a range of geometric models ($k$=0.4, 0.6 and 0.8) to assess the accuracy of single biodiversity as well as the biodiversity difference of two intervals.

# APPENDIX B

# ADJUSTED RAND INDEX

Denote the data matrix as $X = \{x_{ij}\}_{N\times M}$, where $N$ is the number of samples and $M$ is the number of variables. The $N$ samples coming from $C$ true clusters are partitioned into $D$ groups. Let $t_{cd}$ represent the number of subjects that were classified in thc $d$-th cluster that actually belongs to $c$-th cluster. Table 32 can be formed to indicate the clustering results.

The adjusted Rand index is to measure the agreement between the new cluster results and true cluster based upon how pairs of subjects are classified in Table 32. Letting $\binom{N}{2}$ represent the total number of pairs results in four different types of pairs: (a) subjects in a pair coming from same true cluster are placed into same group; (b) subjects in a pair coming from same true cluster are placed into different groups; (c) subjects in a pair coming from different true clusters are placed into same groups; (d) subjects in a pair coming from different true clusters are placed into different groups. This leads to an alternative representation of the Table 32 as a $2\times 2$ contingency table based upon (a), (b), (c) and (d). The four cells of Table 33 are calculated as,

$$a = \frac{\sum_{c=1}^{C}\sum_{d=1}^{D} t_{cd}^2 - N}{2}, \tag{B.0.1}$$

$$b = \frac{\sum_{c=1}^{C} t_{c+}^2 - \sum_{c=1}^{C}\sum_{d=1}^{D} t_{cd}^2}{2}, \tag{B.0.2}$$

$$c = \frac{\sum_{d=1}^{D} t_{+d}^2 - \sum_{c=1}^{C}\sum_{d=1}^{D} t_{cd}^2}{2}, \tag{B.0.3}$$

TABLE 32. Data structure for calculating adjusted Rand index

| | | Clusters Results | | | | |
|---|---|---|---|---|---|---|
| | Group | 1 | 2 | ... | D | Total |
| True Cluster | 1 | $t_{11}$ | $t_{12}$ | ... | $t_{1D}$ | $t_{1+}$ |
| | 2 | $t_{21}$ | $t_{22}$ | ... | $t_{2D}$ | $t_{2+}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ |
| | C | $t_{C1}$ | $t_{c2}$ | ... | $t_{CD}$ | $t_{C+}$ |
| | | $t_{+1}$ | $t_{+2}$ | ... | $t_{+D}$ | $t_{++} = N$ |

TABLE 33. 2 × 2 contingency table representation

| | Cluster Results | |
|---|---|---|
| True Cluster | Pairs Placed in Same Group | Pairs Placed in Different Groups |
| Pairs Coming from Same Clusters | a | b |
| Pairs Coming from Different Clusters | c | d |

$$d = \frac{\sum_{c=1}^{C} \sum_{d=1}^{D} t_{cd}^2 + N^2 - \sum_{c=1}^{C} t_{c+}^2 - \sum_{d=1}^{D} t_{+d}^2}{2}. \tag{B.0.4}$$

Hubert and Arabie (1985) defined the adjusted Rand index as:

$$ARI = \frac{\binom{N}{2}(a+d) - [(a+b)(a+c) + (c+d)(b+d)]}{\binom{N}{2}^2 - [(a+b)(a+c) + (c+d)(b+d)]}. \tag{B.0.5}$$

# APPENDIX C

# THE COLLECTION OF SAS PROGRAMS

## C.1 SUBROUTINES FOR CALCULATING DISTANCE BASED UPON PROBABILITY

```
/*******************************************************************************
/* The following subroutine calculates the distance based upon probability, */
/* which is estimated from the margin proportions. */
/*******************************************************************************
START GETDELTA(MAT);
ROW=NROW(MAT);
COL=NCOL(MAT);
PII=J(ROW, COL, 0);
PI=J(COL,1,0);
LAMDA=J(ROW,1,0);
N_I=MAT*J(COL,1,1);
N_J=J(1,ROW,1)*MAT;
N=SUM(MAT);
DELTA=J(COL,COL,0);
DO J=1 TO COL;
DO I=1 TO ROW;
IF N_I[I]=0 THEN PII[I,J]=0;
ELSE PII[I,J]=MAT[I,J]/N_I[I];
IF PII[I,J]=. THEN PRINT MAT;
END;
END;
DO I=1 TO COL;
DO J=1 TO COL;
IF I=J THEN DELTA[I,J]=0;
ELSE IF (N_J[I]^=0) & (N_J[J]^=0)
THEN DELTA[I,J]=ABS(LOG(N_J[I])-LOG(N_J[J]))+1;
ELSE IF (N_J[I]=0) & (N_J[J]^=0) THEN DELTA[I,J]=LOG(N_J[J])+1;
ELSE IF (N_J[I]^=0) & (N_J[J]=0) THEN DELTA[I,J]=LOG(N_J[I])+1;
ELSE DELTA[I,J]=1;
END;
END;
RETURN(DELTA);
FINISH GETDELTA;
```

## C.2 SUBROUTINE FOR CALCULATING LINEAR COMBINATION OF $\chi^2$ DISTRIBUTIONS

```
/******************************************************************************
/* The following subroutine calculates the linear combination of */
/*$\chi^2$ distribution.*/
/******************************************************************************

START QF(LB, NC, N, R, SIGMA, C, LIM, ACC) GLOBAL(_LB, _NC, _N,
_R, _SIGMA, _C, _LIM, _ACC,QF, TRACE, IFAULT,PI,LN28,SIGSQ,INTL1,INTL2,
ERSM1,ERSM2,LMAX,LMIN,MEAN,COUNT,NDTSRT,FAIL,TH,COUNT1,
COUNT2,COUNT3,ACC1,SD);
_LB=J(R,1,0.0);
_LB=LB;
_NC=J(R,1,0.0);
_NC=NC;
_N=J(R,1,0);
_N=N;
_R=R;
_SIGMA=SIGMA;
_C=C;
_LIM=LIM;
_ACC=ACC;

TRACE=J(7,1,0.0);
IFAULT=0;
QF=-1.0;
```

```
PI=3.14159265358979;
LN28=0.0866;
SIGSQ=_SIGMA*_SIGMA;
INTL1=0.0;
INTL2=0.0;
ERSM1=0.0;
ERSM2=0.0;
LMAX=0.0;
LMIN=0.0;
MEAN=0.0;
COUNT=0;
NDTSRT=1;
FAIL=0;
TH=J(_R,1,0);

COUNT1=0;
COUNT2=0;
COUNT3=0;
ACC1=_ACC;
SD=SIGSQ;

START LN1 (X, FIRST);
_X1=X;
_FIRST=FIRST;
IF ABS(_X1) > 0.1 THEN IF _FIRST=1 THEN LN1=LOG(1.0+_X1);
ELSE LN1=LOG(1.0+_X1)-_X1;
ELSE DO;
Y=_X1/(_X1+2.0);
TERM=2.0*Y*Y*Y;
K=3.0;
IF _FIRST=1 THEN S=2.0*Y;
ELSE S=-_X1*Y;
Y=Y*Y;
S1=S+TERM/K;
```

```
DO WHILE (S1^=S);
K=K+2.0;
TERM=TERM*Y;
S=S1;
S1=S+TERM/K;
END;
LN1=S;
END;
RETURN (LN1);
FINISH LN1;

START ORDER;
/* FIND ORDER OF ABSOLUTE VALUES OF _LB;*/
DO J=1 TO _R;
LJ=ABS(_LB[J]);
DO K=J-1 TO 1 BY -1;
IF LJ > ABS(_LB[TH[K]]) THEN TH[K+1]=TH[K];
ELSE GOTO L1;
END;
K=0;
L1:TH[K+1]=J;
END;
NDTSRT=0;
FINISH ORDER;

START ERRBD(U,CX) GLOBAL(_LB,_NC,_N,_R,_LIM,SIGSQ,COUNT1);
/* FIND BOUND ON TAIL PROBABILITY USING MGF. CUTOFF POINT RE-
TURNED TO CX */
_U1=U;
/* RUN COUNTER;*/
COUNT1=COUNT1+1;
IF COUNT1 > _LIM THEN PRINT 'WARNING:COUNT1 > LIM';

CONST=_U1*SIGSQ;
```

```
SUM1=_U1*CONST;
_U1=2.0*_U1;
DO J=_R TO 1 BY -1;
NJ=_N[J];
LJ=_LB[J];
NCJ=_NC[J];
X=_U1*LJ;
Y=1.0-X;
CONST=CONST+(LJ*(NCJ/Y+NJ))/Y;
SUM1=SUM1+NCJ*(X/Y)*(X/Y)+NJ*(((X*X)/Y)+LN1(-X,0));
LN1=LN1(-X,0);
* PRINT ’LN1=’ LN1;
END;
ERRBD=EXP(-0.5*SUM1);
CX=CONST;
RETURN(ERRBD);
FINISH ERRBD;

START CTFF(ACCX,UPN) GLOBAL(LMAX,LMIN,MEAN);
/*FIND CTFF SO THAT P(QF > CTFF) < ACCX IF UPN > 0, P(QF < CTFF) < ACCX
OTHERWISE */
_ACCX1=ACCX;
U2=UPN;
U1=0.0;
C1=MEAN;
C2=0;
CONST=0;
IF U2 > 0.0 THEN RB=2.0*LMAX;
ELSE RB=2.0*LMIN;
U=U2/(1.0+U2*RB);
ERRBD=ERRBD(U, C2);
DO WHILE (ERRBD > _ACCX1);
U1=U2;
C1=C2;
```

```
U2=2.0*U2;
U=U2/(1.0+U2*RB);
ERRBD=ERRBD(U, C2);
END;
U=(C1-MEAN)/(C2-MEAN);
DO WHILE (U < 0.9);
U=(U1+U2)/2.0;
IF (ERRBD(U/(1.0+U*RB),CONST) > _ACCX1) THEN DO;
U1=U;
C1=CONST;
END;
ELSE DO;
U2=U;
C2=CONST;
END;
U=(C1-MEAN)/(C2-MEAN);
END;
CTFF=C2;
UPN=U2;
RETURN (CTFF);
FINISH CTFF;

START TRUNCATION(U, TAUSQ)
GLOBAL(_LB,_NC,_N,_R,_LIM,PI,SIGSQ,COUNT2);
/* BOUND INTEGRATION ERROR DUE TO TRUNCATION AT U*/
_U2=U;
_TAUSQ1=TAUSQ;

COUNT2=COUNT2+1;
IF COUNT2 > _LIM THEN PRINT ’WARNING: COUNT2 > LIM.’;

SUM1=0.0;
PROD2=0.0;
PROD3=0.0;
```

```
S=0;
SUM2=(SIGSQ+_TAUSQ1)*_U2*_U2;
PROD1=2.0*SUM2;
_U2=2.0*_U2;
DO J=1 TO _R;
LJ=_LB[J];
NCJ=_NC[J];
NJ=_N[J];
X=(_U2*LJ)*(_U2*LJ);
SUM1=SUM1+NCJ*X/(1.0+X);
IF X > 1.0 THEN DO;
PROD2=PROD2+NJ*LOG(X);
PROD3=PROD3+NJ*LN1(X,1);
S=S+NJ;
END;
ELSE PROD1=PROD1+NJ*LN1(X,1);
END;
SUM1=0.5*SUM1;
PROD2=PROD1+PROD2;
PROD3=PROD1+PROD3;
X=(EXP(-SUM1-0.25*PROD2))/PI;
Y=(EXP(-SUM1-0.25*PROD3))/PI;
IF S=0 THEN ERR1=1.0;
ELSE ERR1=X*2.0/S;
IF PROD3 > 1.0 THEN ERR2=2.5*Y;
ELSE ERR2=1.0;
IF ERR2 < ERR1 THEN ERR1=ERR2;
X=0.5*SUM2;
IF X < =Y THEN ERR2=1.0;
ELSE ERR2=Y/X;
IF ERR1 < ERR2 THEN TRUNCATION=ERR1;
ELSE TRUNCATION=ERR2;
RETURN (TRUNCATION);
FINISH TRUNCATION;
```

```
START FINDU(UTX, ACCX);
/*FIND U SUCH THAT TRUNCATION(U) < ACCX & TRUNCATION(U/1.2) >
ACCX*/
_ACCX2=ACCX;
UT=UTX;
U=UT/4.0;

IF TRUNCATION(U,0) > _ACCX2 THEN DO;
TRUN=TRUNCATION(U,0);
U=UT;
TRUN=TRUNCATION(U,0);
DO WHILE (TRUN > _ACCX2);
UT=UT*4.0;
U=UT;
TRUN=TRUNCATION(U,0);
END;
END;

ELSE DO;
UT=U;
U=U/4.0;
TRUN=TRUNCATION(U,0);
DO WHILE (TRUN <= _ACCX2);
UT=U;
U=U/4.0;
TRUN=TRUNCATION(U,0);
END;
END;
U=UT/2.0;
IF TRUNCATION(U,0) <= _ACCX2 THEN UT=U;
U=UT/1.4;
IF TRUNCATION(U,0) <= _ACCX2 THEN UT=U;
U=UT/1.2;
```

```
IF TRUNCATION(U,0) <= _ACCX2 THEN UT=U;
U=UT/1.1;
IF TRUNCATION(U,0) <= _ACCX2 THEN UT=U;
UTX=UT;
FINISH FINDU;

START INTEGRATE(NTERM, INTERV, TAUSQ, MAIN)
GLOBAL(_LB,_NC,_N,_R,_C,_ACC,PI,SIGSQ,INTL1,INTL2,ERSM1,ERSM2);
/*CARRY OUT WITH NTERMS, AT STEPWISE INTERV. IF NOT MAIN THEN MUL-
TIPLY INTEGRAND BY 1.0-EXP(-0.5*TAUSQ*U*U)*/
_NTERM=NTERM;
_INTERV=INTERV;
_TAUSQ2=TAUSQ;
_MAIN=MAIN;
INPI=_INTERV/PI;
DO K=_NTERM TO 0 BY -1;
U=(K+0.5)*_INTERV;
SUM1=-2.0*U*_C;
SUM2=ABS(SUM1);
SUM3=-0.5*SIGSQ*U*U;
DO J=_R TO 1 BY -1;
NJ=_N[J];
X=2.0*_LB[J]*U;
*PRINT ’X=’ X;
Y=X*X;
SUM3=SUM3-0.25*NJ*LN1(Y,1);
Y=_NC[J]*X/(1.0+Y);
Z=NJ*ATAN(X)+Y;
SUM1=SUM1+Z;
SUM2=SUM2+ABS(Z);
SUM3=SUM3-0.5*X*Y;
END;
X=INPI*(EXP(SUM3))/U;
IF (_MAIN=0) THEN X=X*(1.0-EXP(-0.5*_TAUSQ2*U*U));
```

```
SUM1=SIN(0.5*SUM1)*X;
SUM2=0.5*SUM2*X;
IF ABS(SUM1) < _ACC THEN DO;
INTL1=INTL1+SUM1;
ERSM1=ERSM1+SUM2;
END;
ELSE DO;
INTL2=INTL2+SUM1;
ERSM2=ERSM2+SUM2;
END;
END;
FINISH INTEGRATE;

START CFE(X)
GLOBAL(_LB,_NC,_N,_R,_LIM,PI,LN28,COUNT3,NDTSRT,FAIL,TH);
/* COEF OF TAUSQ IN ERROR WHEN CONVERGENCE FACTOR OF EXP(-
0.5*TAUSQ*U*U) IS USED WHEN DF IS EVALUATED AT X */
_X2=X;
COUNT3=COUNT3+1;
IF COUNT3 > _LIM THEN PRINT 'WARNING:COUNT3 > LIM';
IF NDTSRT=1 THEN DO;
CALL ORDER;
END;
AXL=ABS(_X2);
IF _X2=0.0 THEN SXL=0.0;
ELSE IF _X2 > 0.0 THEN SXL=1.0;
ELSE SXL=-1.0;
SUM1=0.0;
DO J=_R TO 1 BY -1;
T=TH[J];
IF _LB[T]*SXL > 0.0 THEN DO;
LJ=ABS(_LB[T]);
AXL1=AXL-LJ*(_N[T]+_NC[T]);
AXL2=LJ/LN28;
```

```
IF AXL1 > AXL2 THEN AXL=AXL1;
ELSE DO;
IF AXL > AXL2 THEN AXL=AXL2;
SUM1=(AXL-AXL1)/LJ;
DO K=J-1 TO 1 BY -1;
SUM1=SUM1+(_N[TH[K]]+_NC[TH[K]]);
END;
GOTO L;
END;
END;
END;
L: IF SUM1 > 100.0 THEN DO;
CFE=1.0;
FAIL=1;
END;
ELSE CFE=EXP((LOG(2.0))*(SUM1/4.0))/(PI*AXL*AXL);
RETURN (CFE);
FINISH CFE;

*START QF;
DO J=1 TO _R;
NJ=_N[J];
LJ=_LB[J];
NCJ=_NC[J];
IF (NJ < 0) — (NCJ < 0.0) THEN DO;
IFAULT=3;
GOTO EXIT;
END;
SD=SD+LJ*LJ*(2*NJ+4.0*NCJ);
MEAN=MEAN+LJ*(NJ+NCJ);
IF LMAX < LJ THEN LMAX=LJ;
ELSE IF LMIN > LJ THEN LMIN=LJ;
END;
IF SD=0.0 THEN DO;
IF _C > 0.0 THEN QF=1.0;
ELSE QF=0.0;
GOTO EXIT;
END;
IF (LMIN=0.0) & (LMAX=0.0) & (_SIGMA=0.0) THEN DO;
IFAULT=3;
GOTO EXIT;
END;
SD=SQRT(SD);

IF LMAX < -LMIN THEN ALMX=-LMIN;
ELSE ALMX=LMAX;

UTX=16.0/SD;
UP=4.5/SD;
UN=-UP;

CALL FINDU(UTX, 0.5*ACC1);

IF (_C^=0.0) & (ALMX > 0.07*SD) THEN DO;
CFE=CFE(_C);
TAUSQ=0.25*ACC1/CFE;
IF FAIL=1 THEN FAIL=0;
ELSE IF TRUNCATION(UTX,TAUSQ) < (0.2*ACC1) THEN DO;
SIGSQ=SIGSQ+TAUSQ;
CALL FINDU(UTX,0.25*ACC1);
PRINT UTX;
TRACE[6]=SQRT(TAUSQ);
END;
END;
TRACE[5]=UTX;
ACC1=0.5*ACC1;

L1: D1=CTFF(ACC1,UP)-_C;
```

```
IF D1 < 0.0 THEN DO;
QF=1.0;
GOTO EXIT;
END;
D2=_C-CTFF(ACC1,UN);
IF D2 < 0.0 THEN DO;
QF=0.0;
GOTO EXIT;
END;
IF D1 > D2 THEN INTV=2.0*PI/D1;
ELSE INTV=2.0*PI/D2;

NT=INT(UTX/INTV);
NTM=INT(3.0/SQRT(ACC1));
IF NT > NTM*1.5 THEN DO;
INTV1=UTX/NTM;
X=2.0*PI/INTV1;
IF X <= ABS(_C) THEN GOTO L2;

TAUSQ=0.33*ACC1/(1.1*(CFE(_C-X)+CFE(_C+X)));
IF FAIL=1 THEN GOTO L2;
ACC1=0.67*ACC1;
IF NTM > _LIM THEN DO;
IFAULT=1;
GOTO EXIT;
END;

CALL INTEGRATE(NTM,INTV1,TAUSQ,0);
_LIM=_LIM-NTM;
SIGSQ=SIGSQ+TAUSQ;
TRACE[3]=TRACE[3]+1;
TRACE[2]=TRACE[2]+NTM+1;

CALL FINDU(UTX,0.25*ACC1);
```

```
ACC1=0.75*ACC1;
PRINT UTX;
GOTO L1;
END;

L2: TRACE[4]=INTV;
IF NT > _LIM THEN DO;
IFAULT=1;
GOTO EXIT;
END;
CALL INTEGRATE(NT, INTV, 0, 1);
TRACE[3]=TRACE[3]+1;
TRACE[2]=TRACE[2]+NT+1;
QF=0.5-INTL1-INTL2;
TRACE[1]=ERSM1+ERSM2;
ERSM1=ERSM1+ERSM2;

PRINT 'QF=' QF;
X=ERSM1+_ACC/10.0;
IF X=ERSM1 THEN IFAULT=2;
IF 2*X=2*ERSM1 THEN IFAULT=2;
IF 4*X=4*ERSM1 THEN IFAULT=2;
IF 8*X=8*ERSM1 THEN IFAULT=2;

EXIT: TRACE[7]=COUNT1;

RETURN(QF);
FINISH QF;
```

## C.3 SUBROUTINES FOR CALCULATING THE MAXIMUM VALUE OF RAO'S QUADRATIC ENTROPY

```
/**********************************************************************************/
/* Define the function of divc() to get the Rao's diversity coefficient.*/
/**********************************************************************************/
START DIVC(DF,DIS,SCALE);
IF ANY(DF < 0) THEN DO;
PRINT "NEGATIVE VALUE IN DF";
STOP;
END;
IF DIS=J(NROW(DF),NROW(DF),0) THEN
DIS=J(NROW(DF),NROW(DF),1)-DIAG(REPEAT(1,NROW(DF)))*SQRT(2);
ELSE DO;
IF NROW(DF)^=NROW(DIS) THEN DO;
PRINT "NON CONVENIENT DF" STOP;
END;
END;
DIV=REPEAT(0,NCOL(DF));
DO I=1 TO NCOL(DF);
IF SUM(DF[,I]) < 1E-16 THEN DIV[I,]=0;
ELSE DIV[I,]=(T(DF[,I])*(DIS##2)*DF[,I])/2/(SUM(DF[,I])**2);
END;
IF SCALE=1 THEN DO;
DIVCMAX=DIVCMAX(DIS);
DIV=DIV/DIVCMAX;
END;
RETURN(DIV);
FINISH DIVC;

/**********************************************************************************/
/* Define the function of divcmax() to get the Maximal value of Rao's*******/
/* diversity coefficient ;******************************************************/
/**********************************************************************************/
```

```
START DIVCMAX(DIS, EPSILON,COMMENT) GLOBAL(RESULT);
IF EPSILON <= 0 THEN DO;
PRINT ”EPSILON MUST BE POSITIVE”;
STOP;
END;
D2=DIS##2/2;
N=NROW(D2);
RESULT=J(N,4,0);
MATTRIB RESULT COLNAME=(SIM PRO MET NUM);
RELAX=0;
X0=D2[,+]/SUM(D2);
RESULT[,1]=X0;
OBJECTIVE0=T(X0)*D2*X0;
IF COMMENT=1 THEN PRINT ”EVOLUTION OF THE OBJECTIVE FUNCTION:”;
XK=X0;
DO;
LOOP1A: DO;
LOOP2A: MAXI_TEMP=T(XK)*D2*XK;
IF COMMENT=1 THEN PRINT MAXI_TEMP;
DELTAF=-2#D2*XK;
SATURE=J(NROW(XK),NCOL(XK),1);
DO I=1 TO NROW(XK);
DO J=1 TO NCOL(XK);
IF (ABS(XK[I,J]) < EPSILON) THEN SATURE[I,J]=1;
ELSE SATURE[I,J]=0;
END;
END;
IF RELAX^=0 THEN DO;
SATURE[RELAX]=0;
RELAX=0;
END;
YK=-DELTAF;
DO I=1 TO NROW(YK);
DO J=1 TO NCOL(YK);
```

```
IF SATURE[I,J]=1 THEN YK[I,J]=0;
END;
END;
_COUNT=0;
_SUM=0;
DO I=1 TO NROW(YK);
DO J=1 TO NCOL(YK);
IF SATURE[I,J]=0 THEN DO;
_COUNT=_COUNT+1;
_SUM=_SUM+YK[I,J];
END;
END;
END;
_MEAN=_SUM/_COUNT;
DO I=1 TO NROW(YK);
DO J=1 TO NCOL(YK);
IF SATURE[I,J]=0 THEN YK[I,J]=YK[I,J]-_MEAN;
END;
END;
IF MAX(ABS(YK)) < EPSILON THEN GOTO LOOP2B;
ALPHA_MAX=1;
_RATIO=1;
DO I=1 TO NROW(YK);
DO J=1 TO NCOL(YK);
IF YK[I,J] < 0 THEN DO;
_RATIO=-XK[I,J]/YK[I,J];
IF _RATIO < ALPHA_MAX THEN ALPHA_MAX=_RATIO;
END;
END;
END;
ALPHA_OPT=(-T(XK)*D2*YK)/(T(YK)*D2*YK);
IF (ALPHA_OPT > ALPHA_MAX) | (ALPHA_OPT < 0) THEN
ALPHA=ALPHA_MAX;
ELSE ALPHA=ALPHA_OPT;
```

```
IF(ABS(MAXI_TEMP-T(XK+ALPHA*YK)*D2*(XK+ALPHA*YK))  <  EPSILON)
THEN GOTO LOOP2B;
XK=XK+ALPHA*YK;
GOTO LOOP2A;
LOOP2B: END;
IF SUM(SATURE)=0 THEN DO;
IF COMMENT=1 THEN DO;
PRINT ”KT1” XK;
END;
END;
IF SUM(SATURE)=0 THEN GOTO LOOP1B;

VECTD2=D2*XK;
_COUNT_=0;
_SUM_=0;
DO I=1 TO NROW(VECTD2);
DO J=1 TO NCOL(VECTD2);
IF SATURE[I,J]=0 THEN DO;
_COUNT_=_COUNT_+1;
_SUM_=_SUM_+VECTD2[I,J];
END;
END;
END;
_MEAN_=_SUM_/_COUNT_;
_COUNT2_=0;
DO I=1 TO NROW(VECTD2);
DO J=1 TO NCOL(VECTD2);
IF SATURE[I,J]=1 THEN DO;
_COUNT2_=_COUNT2_+1;
_MAT_=_MAT_//SATURE[I,J];
END;
END;
END;
U=2#(J(_COUNT2_,1,_MEAN_)-_MAT_);
```

```
IF (MIN(U) > =0) THEN DO;
IF COMMENT=1 THEN DO;
PRINT ”KT2” XK;
END;
END;
IF (MIN(U) > =0) THEN GOTO LOOP1B;
ELSE DO;
IF COMMENT=1 THEN DO;
PRINT ”RELAXATION” XK;
END;
DO I=1 TO N;
IF SATURE[I]=1 THEN SATU=SATU//I;
END;
DO I=1 TO NROW(U);
DO J=1 TO NCOL(U);
IF U[I,J]=MIN(U) THEN _RELAX_=_RELAX_//SATU[I,J];
END;
END;
RELAX=_RELAX[1];
END;
GOTO LOOP1A;
LOOP1B: END;
IF COMMENT=1 THEN PRINT OBJECTIVE0 MAXI_TEMP;
RESULT[,4]=XK;
DO I=1 TO NROW(RESULT);
IF RESULT[I,4] < EPSILON THEN RESULT[I,4]=0;
END;
XK=X0/SQRT(SUM(X0#X0));
DO UNTIL (MAX(XK-YK) <= EPSILON);
YK=D2*XK;
YK=YK/SQRT(SUM(YK#YK));
IF MAX(XK-YK) > EPSILON THEN XK=YK;
ELSE DO;
PRINT ”STOP5”;
```

```
END;
END;
X0=YK;
RESULT[,2]=X0/SUM(X0);
RESULT[,3]=X0#X0;
RESTOT=DIVC(RESULT[,4],DIS,0);
PRINT RESULT RESTOT;
*RETURN(RESTOT);
FINISH DIVCMAX;
```

## C.4 SUBROUTINES FOR CALCULATING DISTANCE IN TWO MULTINOMIAL POPULATIONS

```
/******************************************************************************/
/*The following subroutine calculate the Bhattacharyya Distance. */
/******************************************************************************/
START DISTB(XI,XJ);
/*XI,XJ ARE TWO MULTINOMIAL DISTRIBUTED VECTORS WITH TOTAL NUM-
BER IN THE LAST COLUMN;*/
NCOL=NCOL(XI);
SUM=0;
DO I=1 TO NCOL-1;
SUM=SUM+(SQRT(XI[I])-SQRT(XJ[I]))**2;
END;
DISTB=SQRT(SUM);
RETURN(DISTB);
FINISH DISTB;

/******************************************************************************/
/*The following subroutine calculate the Rao's Quadratic Entropy Distance. */
/******************************************************************************/
START DISTQE(XI,XJ,DELTA);
```

```
/*XI,XJ ARE TWO MULTINOMIAL DISTRIBUTED VECTORE WITH TOTAL NUM-
BER IN THE LAST COLUMN;*/
NCOL=NCOL(XI);
P_BAR=(XI[1:NCOL-1]+XJ[1:NCOL-1])/(XI[NCOL]+XJ[NCOL]);
SST=P_BAR‘*DELTA*P_BAR;
SSW1=(XI[1:NCOL-1])‘*DELTA*(XI[1:NCOL-1])/(XI[NCOL]*XI[NCOL]);
SSW2=(XJ[1:NCOL-1])‘*DELTA*(XJ[1:NCOL-1])/(XJ[NCOL]*XJ[NCOL]);
DISTQE=(SST-SSW1*XI[NCOL]/(XI[NCOL]+XJ[NCOL])
-SSW2*XJ[NCOL]/(XI[NCOL]+XJ[NCOL]))/SST;
*PRINT SST SSW1 SSW2 DISTQE;
RETURN(DISTQE);
FINISH DISTQE;
```

## C.5 SUBROUTINES FOR MINIMUM QUADRATIC ENTROPY CLUSTERING ALGORITHM

```
/*************************************************************************/
/*The following subroutine calculates the minimum quadratic entropy */
/*clustering criterion. */
/*************************************************************************/
START MEC_QE(MAT,MAT1,DIST,C,V);
MINQE=0;
N=NROW(MAT);
DO I=1 TO N;
K=J(1,C,0);
TOTAL=J(C,2,0);
DO J=1 TO N;
IF DIST[I,J] <= V THEN DO;
K[MAT[J,2]]=K[MAT[J,2]]+1;
TOTAL[MAT[J,2],]=TOTAL[MAT[J,2],]+MAT1[J,1:2];
END;
END;
DELTA=J(C,C,0);
```

```
DO II=1 TO C;
DO JJ=1 TO C;
IF II=JJ THEN DELTA[II,JJ]=0;
ELSE IF K[II]^=0 & K[JJ]^=0 THEN
DELTA[II,JJ]=(TOTAL[II,]/K[II]-TOTAL[JJ,]/K[JJ])*T(TOTAL[II,]/K[II]-
TOTAL[JJ,]/K[JJ]);
ELSE IF K[II]=0 & K[JJ]^=0 THEN
DELTA[II,JJ]=(TOTAL[JJ,]/K[JJ])*T(TOTAL[JJ,]/K[JJ]);
ELSE IF K[II]^=0 & K[JJ]=0 THEN
DELTA[II,JJ]=(TOTAL[II,]/K[II])*T(TOTAL[II,]/K[II]);
ELSE DELTA[II,JJ]=0;
END;
END;
*DELTA=GETDELTA1(K);
NN=K[+];
IF NN^=0 THEN DO;
QE=K/NN*DELTA*K`/NN;
MINQE=MINQE+QE;
END;
END;
RETURN(MINQE);
FINISH MEC_QE;

/****************************************************************************/
/*The following subroutine recluster data based on MQEC in hill-climbing */
/*iterations. */
/****************************************************************************/

START RC_QE(MAT,MAT1,DIST,C,V);
TEMPMAT=MAT;
N=NROW(TEMPMAT);
DO I=1 TO N;
MINQE1=MEC_QE(TEMPMAT,MAT1,DIST,C,V);
CL=TEMPMAT[I,2];
```

```
K=J(1,C,0);
DO J=1 TO N;
IF DIST[I,J] <= V & (I^=J) THEN K[TEMPMAT[J,2]]=K[TEMPMAT[J,2]]+1;
END;
DO CC=1 TO C;
IF K[CC]=MAX(K) THEN DO;
TEMPMAT[I,2]=CC;
CC=C;
END;
END;
MINQE2=MEC_QE(TEMPMAT,MAT1,DIST,C,V);
IF MINQE2 > =MINQE1 THEN TEMPMAT[I,2]=CL;
ELSE MINQE1=MINQE2;
END;
RETURN(TEMPMAT);
FINISH RC_QE;


/*************************************************************************/
/*The following subroutine calculates Huber and Arabie */
/*Adjusted Rand Index.*/
/*************************************************************************/

START RAND(MAT,C1,C2);
/*MAT HAS TO BE N*2 MATRIX WITH FIRST COLUMN AS TRUE CLUSTER, AND SECOND COLUMN AS NEW CLUSTER*/
/*C1 IS THE NUMBER OF TRUE CLUSTERS, C2 IS THE NUMBER OF NEW CLUSTERS*/
NEWMAT=J(C1,C2,0);
N=NROW(MAT);
SUMSQ_IJ=0;
DO I=1 TO C1;
DO J=1 TO C2;
DO K=1 TO N;
```

```
IF MAT[K,1]=I & MAT[K,2]=J THEN NEWMAT[I,J]=NEWMAT[I,J]+1;
END;
SUMSQ_IJ=SUMSQ_IJ+NEWMAT[I,J]*NEWMAT[I,J];
END;
END;
SUMSQ_I=NEWMAT[+,]*T(NEWMAT[+,]);
SUMSQ_J=T(NEWMAT[,+])*NEWMAT[,+];
A=(SUMSQ_IJ-N)/2;
B=(SUMSQ_J-SUMSQ_IJ)/2;
C=(SUMSQ_I-SUMSQ_IJ)/2;
D=(SUMSQ_IJ+N*N-SUMSQ_I-SUMSQ_J)/2;
ARI=(COMB(N,2)*(A+D)-((A+B)*(A+C)+(C+D)*(B+D)))/((COMB(N,2))**2-
((A+B)*(A+C)+(C+D)*(B+D)));
RETURN(ARI);
FINISH RAND;
```

# VITA

Yueqin Zhao
Department of mathematics and statistics
Old Dominion University
Norfolk, VA 23529

## Education

PhD Old Dominion University, Norfolk, VA, USA (May 2010)
Major: Computational and Applied Mathematics (Biostatistics)

MS Old Dominion University, Norfolk, VA, USA (May 2004)
Major: Computational and Applied Mathematics (Biostatistics)

BS Shanghai University of Finances & Economics, Shanghai, P.R.China. (July 2000)
Major: Statistics

## Experience

Instructor and Biostatistician (Sep 2005–Current)
Eastern Virginia Medical School, Norfolk, VA

Graduate Research Assistant (Jan 2005–Sep 2005)
Eastern Virginia Medical School, Norfolk, VA

Graduate Teaching Assistant (Aug 2001–Dec 2004)
Old Dominion University, Norfolk, VA

## Publications

**Zhao, Y.** and Naik, D. (2010), "Analysis of biodiversity with Rao's quadratic entropy." Under preparation.

Typeset using LaTeX.