Journal of Statistical Research *ISSN 0256 - 422 X*
2005, Vol. 39, No. 2, pp. 1–15
Bangladesh

NONPARAMETRIC TESTS FOR ORDERED DIVERSITY IN A GENOMIC SEQUENCE

Pranab K. Sen

Departments of Biostatistics, and Statistics and Operations Research
University of North Carolina at Chapel Hill, NC27599-7420, USA
Email: pksen@bios.unc.edu

SUMMARY

In genomics (SNP and RNA amino acid studies), typically, we encounter enormously large dimensional qualitative categorical data models without an ordering of the categories, thus preempting the use of conventional measures of dispersion (variation or diversity) as well as other measures which assume some latent trait variable(s). The Gini-Simpson diversity measure, often advocated for diversity analysis in one-dimensional models, has been adapted to formulate measures of diversity and co-diversity based on the Hamming distance in the multidimensional setup. Based on certain (molecular) biologically interpretable monotone diversity perspectives, an ordering of the Gini-Simpson measures across the genome (positions) is formulated in a meaningful way. Motivated by this feature, nonparametric inference for such ordered measures is considered here, and their applications stressed.

Keywords and phrases: Gini-Simpson diversity index; Hamming distance; High-diemensional qualitative data models; U-statistics.

AMS Classification:

1 Introduction

For qualitative categorical data models, conventional measures of variation and covariation are not meaningful. Hence, diversity and co-diversity analyses have been advocated for such models. As an example, consider a SNP (*single neucleotide polymorphism*) model for the DNA nucleotides (A, C, G, T) encompassing a large number (K) of positions (or genes). Without any interpretable ordering of the labels A, C, G, T (and even a plausible latent-effect model), we encounter a (large) K-dimensional categorical data model with 4^K possible response category combinations. For RNA codons, there being 20 amino acids, the number of possible response category combinations jumps to 20^K. There is apparently diversity or qualitative variation of the response in each position, and also possibly co-diversity among the positions, and these can not be interpreted in terms of product moments or even by some

latent variables, accounting for plausible dependence. For simplicity and manageability of statistical alalysis, it is often assumed that the positions exhibit independent responses, and sometimes, it is even assumed that they have the same probability law for the K positions; in reality, neither the independence nor the identical distribution assumption may not be taken for granted. Therefore it is of considerable interest to formulate suitable measures of diversity and co-diversity, and relate them to the molecular biological undercurrents so as to gain some meaningful insight of the overall biodiversity picture.

The celebrated Gini-Simpson diversity index (Gini 1912, Simpson 1949) has found some very useful applications in genetics and bioinformatics, and in particular genomics. For SNP with a large number of positions, there is a practical difficulty in using directly the Gini-Simpson index. In that respect, Hamming distance based measures have been advocated in the literature (Pinheiro et al. 2000, 2005), Tzeng et al. (2003), Schaid et al. (2005), and others. Basically the Hamming distance is an unweighted average (over the K positions) of the marginal Gini-Simpson indexes. As such, it may not be very sensitive to inter-positions covariability or co-diversity, although it takes into account the stochastic dependence among the positions as well as possible heterogeneity of their marginal distributions. Faced with this limitation, we are to examine the diversity-codiversity perspectives in a detailed manner. In that context, based on suitable genomic (and polygenic) interpretation, we propose some monotone diversity-codiversity perspectives, and for that we formulate some nonparametric estimation as well as testing procedures.

Along with the preliminary notion, these measures are introduced in Section 2. The monotone diversity features are then outined in Section 3. Section 4 is devoted to sample counterparts and formulation of suitable testing procedures. The concluding section deals with some illustrations of the methodology developed in earlier sections.

2 Preliminary Notion

A multinomial probability law is completely characterized by its cell probabilities. In high-dimensional models, the number of cells may be so large that incorporating the set of all cell probabilities in modeling and statistical analysis could be an impractical or dreadful task, unless the number of observations (or sample size n) is also very large. A measure of variability in one-dimensional models, usually formulated solely in terms of the cell probabilities, needs considerably scrutiny in the multidimensional case, as such diversity analysis needs to be based on the set of all (joint) probabilities. Further, as these categories differ only qualitatively, the usual measures of variation or dispersion for quantitative data models are not usable. Even *latent variable (effects) models* may not be generally appropriate in such studies. Gini (1912) came up with a very interesting measure of diversity or lack of concentration; almost after 4 decades, Simpson (1949), apparently unaware of Gini's work, considered the same measure for biodiversity in some ecological studies. This *Gini-Simpson index* (GSI), as it is referred to in the literature, has been extensively used in many applied fields (economics, social sciences, psychometry, and genetics, among others) and genomics

is no exception. A significant amount of work on diversity and dissimilarity coefficients is due to C. R. Rao (1982 a,b,c), rich with statistical interpretations and inferential tools. However, in most of these usages, there are multiple response variables (often, too many), and hence, the GSI needs extensions to suit such more complex models. In genomics, as we shall discuss later on, the problem is much more acute due to an abundance of positions, leading to a genuine *curse of dimensionality* problem.

Consider a simple multinomial distribution for a C cell (labelled as $1, \ldots, C$) model, with respective cell probabilities $\pi_1, \ldots, \pi_C$, and let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_C)'$. Note that $\boldsymbol{\pi}'\mathbf{1} = 1$, so that $\boldsymbol{\pi}$ bdlongs to the simplex:

$$\mathcal{S}_{C-1} = \{\mathbf{x} \in [0,1]^C : \mathbf{x}'\mathbf{1} = 1\}. \tag{2.1}$$

Note that the concentration is least when the π_c are all equal, and it is the maximum when one of the C cell probabilities is 1 while the others are all 0; the GSI has been posed as

$$I_{GS}(\boldsymbol{\pi}) = 1 - \boldsymbol{\pi}'\boldsymbol{\pi} = 1 - \sum_{c=1}^{C} \pi_c^2. \tag{2.2}$$

If we draw two independent observations (say, X and Y), each assuming one of the C labels $1, \ldots, C$ with the common probabilities $\pi_1, \ldots, \pi_C$, then

$$P\{X \neq Y\} = \sum_{c=1}^{C} \pi_c(1 - \pi_c) = 1 - \sum_{c=1}^{C} \pi_c^2 = I_{GS}(\boldsymbol{\pi}). \tag{2.3}$$

The $I_{GS}(\boldsymbol{\pi})$ is an estimable parameter (Hoeffding 1948) admitting an unbiased estimator, a U-statistic, which possesses some nonparametric optimality properties. In a K-dimensional case, assume that in each of the K positions, the number of categories is the same which we label as $1, \ldots, C$. Thus, each observation $\mathbf{X} = (X_1, \ldots, X_K)'$ has K coordinates with X_k can taking on each of the labels $1, \ldots, C$, there being a totality of C^K possible realizations. Typically in genomics, C is fixed (viz., 4 for DNA nucleotides and 20 for RNA amino acids) but K is very large, so that C^K may be so large compared to the number of observations (n) that conventional discrete multivariate analysis may be of little assistance. The main difficulty arises from this *curse of dimensionality* (i.e., $K >> n$) problem. For a K-variate normal distribution, the variation - covariation is completely characterized by its (positive semi-definite (p.s.d.)) dispersion matrix having $K(K+1)/2$ unknown elements. For multidimensional categorical data models, though belonging to the exponential family, the diversity-codiversity (of various orders) can not be simply charaterized by marginal and two-factor joint probabilities.

Borrowing analogy with the classical multinormal case, albeit little less emphatically, we may define co-disagreement or codiversity among a pair of positions $(k, q) : 1 \leq k < q \leq K$, as

$$CD_{GS}(k, q) = P\{X_k \neq X_q\} = 1 - \sum_{c=1}^{C} \pi_{kq,cc}, \tag{2.4}$$

where $\pi_{kq,cc} = P\{X_k = X_q = c\},\ c = 1, \ldots, C;\ k \neq q = 1, \ldots, K$ denote the set of all possible two-factor concordance probabilities. Note that for $k = q$, $CD^*_{GS}(k, k) = 0$, as it should be, so that if we consider the $K \times K$ matrix of these codiversity measures, we have all null diagonal elements. Note that these codiversity indexes depend on the two-factor cell probabilities and are not simply based on the marginals. On the other hand, by analogy with the Gini-Simpson indexes, we may define

$$I_{GS}(k, q) = P\{X_k \neq Y_q\} = 1 - \sum_{c=1}^{C} \pi_{kc}\pi_{qc} \tag{2.5}$$

for $k, q = 1, \ldots, K$. However, note that the product of the marginal probabilities fails to capture any information on their joint probability, and further

$$\sum_{c=1}^{C} \pi_{kc}\pi_{qc} \ \leq \ \frac{1}{2}\{\sum_{c=1}^{C} \pi^2_{k,c} \ + \ \sum_{c=1}^{C} \pi^2_{q,c}\}, \tag{2.6}$$

for every pair $k \neq q = 1, \ldots, K$, we obtain that

$$I_{GS}(k, q) \geq \{I_{GS}(k) + I_{GS}(q)\}/2,\ \forall k \neq q, \tag{2.7}$$

where the equality sign holds only when $\boldsymbol{\pi}_k = \boldsymbol{\pi}_q$. Thus, if we want to gather information on the homogeneity of the marginal IGS then these $I_{GS}(k, q)$ provide additional information. Our main interest corners some marginal indexes exhibiting some order relationship in a meaningful way, incorporating additional information from the intersites diversity measures.

The Hamming distance based on the K vector of marginal IGS is a summaritative measure of this marginal diversity, without accounting for the codiversity indexes, although it takes into account, to a certain extent, possible stochastic hetergeneity and dependence among the positions. If two response vectors $\mathbf{X}$ and $\mathbf{Y}$ are from the common multidimensional multinomial law, the Hamming distance is defined as

$$\begin{aligned} HD(\boldsymbol{\Pi}) &= K^{-1} \sum_{k=1}^{K} P\{X_k \neq Y_k\} \\ &= K^{-1} \sum_{k=1}^{K} [\sum_{c=1}^{C} \pi_{kc}(1 - \pi_{kc})] \\ &= K^{-1} \sum_{k=1}^{K} [1 - \sum_{c=1}^{C} \pi^2_{kc}] \\ &= K^{-1} \sum_{k=1}^{K} I_{GS}(\boldsymbol{\pi}_k), \end{aligned} \tag{2.8}$$

where the $K \times C$ matrix $\boldsymbol{\Pi}$ consists of the K marginal probability vectors $\boldsymbol{\pi}_k = (\pi_{k1}, \ldots, \pi_{kC}), k = 1, \ldots, K$. Note that $\boldsymbol{\Pi}$ is not the original C^K probability vector of all possible realizations

of the $\mathbf{X}$ but a projection of that into its K marginal vectors. All these measures would be useful in drawing statistical inference on the diversity pattern across the K positions, as we sometimes encounter in genomic studies. These will be elaborated in the subsequent sections.

3 Monotone Diversity

In genomics, the *Central Dogma* (Crick 1970) states that once information (i.e., the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein) passes into protein, it can not get out again. The transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein, may be possible, but the transfer from protein to protein, or protein to nucleic acid is not possible. The central dogma has been extended in later years. In some genetic system, RNA templates RNA. Also, retroviruses (which have the ability ro reverse the normal flow of genetic information) can copy their RNA genomes into DNA by a mechanism called *reverse transcription* (RT). The genetic variability of HIV is relatively high compared to other retroviruses. Stochastic evolutionary forces act on genomes (*molecular evolution*), and the genes are not simple.

It makes no sense to assume that the K positions have independent and identically distributed (i.i.d.) (categorical) responses, although for the sake of managability and simplicity of statistical modeling and analysis, such i.i.d. clauses are often presumed. It is therefore of interest to explore plausible departures from such i.i.d. clauses, supporting them from molecular biological interpretations, and developing suitable statistical inference tools to judge the feasibility of such models. Realizing that stochastic equilibrium is anticipated following any signal to noise activity, it may be reasonable to consider some models where

$$I_{GS}(\boldsymbol{\pi}_k) \text{ is monotone in } k(\leq K). \tag{3.1}$$

This immediately suggests two statistical problems:

(i) How to estimate the $I_{GS}(\boldsymbol{\pi}_k)$ subject to such isotonic constraints, and

(ii) How to test for such possible isotonic GSI's?

In either case, it may not be reasonable to impose the independence clause.

It is possible to describe the joint probability law for $\mathbf{X}$ as

$$P\{\mathbf{X}=\mathbf{x}\} = P\{X_1 = x_1\} \prod_{k=2}^{K} P\{X_k = x_k | X_j = x_j, 1 \leq j \leq k-1\}, \tag{3.2}$$

where $\mathbf{x} = (x_1, \ldots, x_K)$ with each x_j taking on the C labels $1, \ldots, C$. Thus, there is a possible transition from the state x_{k-1} to the state x_k, at the kth site, for $k = 1, \ldots, C$. In some cases, it might be plausible to assume a first-order *Markov chain*, so that the conditional probabilities can be expressed as

$$P\{X_k = x_k | X_j = x_j, j < k\} = P\{X_k = x_k | X_{k-1} = x_{k-1}\}, k = 2, \ldots, K. \tag{3.3}$$

For a stationary Markov chain, these transition probabilities do not depend on k, so that the independence of the positions can be relaxed to a stationay first-order Markov chain property. In that case, we have $C(C+1)$ parameters consisting of C marginal probabilities $\pi_c, c = 1, \ldots, C$ and C^2 transition probabilities $\pi_{cd},\ c, d = 1, \ldots, C$,there being $C(C-1)$ linearly independent parameters among them. Also, by definition,

$$P\{X_k \neq Y_{k+1}\} = 1 - \sum_{c=1}^{C} \pi_c \pi_{cc}, \tag{3.4}$$

which does not depend on k. In general, for any $m \geq 1$,

$$P\{X_k \neq Y_{k+m}\} = 1 - \sum_{c=1}^{C} \pi_c \pi_{cc}^{(m)}, \tag{3.5}$$

where the $\pi_{cc}^{(m)}$ are the mth order transition probabilities as can be obtained by the power matrix form prevailing for Markov chains. Thus, for stationary Markov chains, the marginal IGS are all the same, while the $I_{GS}(k, q)$ depend only on $|k - q|$.

With a monotone nondecreasing GSI across the positions, under a Markovian setup, the stationarity of the chain may no longer hold, though it might be argued that for each k,

$$m_k^2 = \sum_{c=1}^{C} \pi_{k,c}^2 \leq \sum_{c=1}^{C} \pi_{k-1,c}^2 = m_{k-1}^2. \tag{3.6}$$

As a result, we have for every k,

$$I_{GS}(k-1, k) \geq I_{GS}(k-1, k-1). \tag{3.7}$$

If $\boldsymbol{\pi}_{k-1} = \boldsymbol{\pi}_k$ then, of course, $I_{GS}(k-1, k) = I_{GS}(k, k) = I_{GS}(k-1, k-1)$. On the other hand, if the two points $\boldsymbol{\pi}_{k-1}$ and $\boldsymbol{\pi}_k$ are very close to each other in the sense that

$$m_k^2 \leq \sum_{c=1}^{C} \pi_{k-1,c} \pi_{k,c} \leq m_{k-1}^2, \tag{3.8}$$

then we have

$$I_{GS}(k-1, k-1) \leq I_{GS}(k-1, k) \leq I_{GS}(k, k). \tag{3.9}$$

In a similar way, we may argue that the following inequality

$$\sum_{c=1}^{C} \{\pi_{k-1,k;cc} - \pi_{k-1,c} \pi_{k,c}\} \leq 0 \tag{3.10}$$

holds, then we have

$$CD_{GS}(k-1, k) \geq I_{GS}(k-1, k), \tag{3.11}$$

where the equality sign holds when the events are independent. By chain-rule, we have therefore

$$CD_{GS}(k,q) \geq I_{GS}(\boldsymbol{\pi}_k),\ \forall k < q = 1,\ldots,K. \tag{3.12}$$

However, in genomic sequences, the precise ordering of the positions may often be left to honest statistical guess, and hence, the above inequality restraints may not be always sound in terms of biological interpretations.

If we let $\pi_c^* = (\pi_{k,c} + \pi_{q,c})/2,\ c = 1,\ldots,C$, and note that $\boldsymbol{\pi}^* = (\pi_1^*,\ldots,\pi_C^*)'$ belongs to the simplex $\mathcal{S}_{C-1}$ and the contour on $\mathcal{S}_{C-1}$ marked by the intersection with the sphere $\boldsymbol{\pi}^{*'}\boldsymbol{\pi}^* = m^2$ corresponds to a value of m that lies between the radii of the two spheres formed by the two $\boldsymbol{\pi}_k$ and $\boldsymbol{\pi}_q$. Therefore, under the assumed monotonicity condition on the $I_{GS}(k)$, we obtain that

$$\sum_{c=1}^{C} \pi_{k,c}\pi_{q,c} \leq \sum_{c=1}^{C} (\pi_c^*)^2 \leq \sum_{c=1}^{C} \pi_{k,c}^2, \tag{3.13}$$

for every $k < q$. This, in turn, implies that

$$I_{GS}(\boldsymbol{\pi}_k) \leq I_{GS}(k,q) \leq I_{GS}(\boldsymbol{\pi}_q), \tag{3.14}$$

for every pair $(k,q) : 1 \leq k < q \leq K$. This is the basic set of inequality-restraints on the GSI for each position as well as inter-position GSI's. However, it does not utilize the information contained in the $CD_{GS}(k,q)$. For reasons explained before, we shall not incorporate this additional information; otherwise, the formulation could be quite cumbersome. As a matter of fact, along the sameline, it follows that under the same regularity conditions,

$$CD_{GS}(k,q) \geq CD_{GS}(k',q'),\ \forall k \geq k',\ q \geq q'. \tag{3.15}$$

On the other hand, if positions k and q are far apart, for a first-order Markov chain, $\pi_{k,q}(c,c)$ should behave like the product of the two marginal probabilities, so that in a stationary case, $CD_{GS}(k,q)$ behaves like $I_{GS}(k,q)$. Therefore, we have more reasons to avoid the measures $CD_{GS}(k,q)$ for additional information.

Thus, as a composite measure of the monotone diversity, we could pose the following:

$$\begin{aligned} J(K) &= \sum_{1\leq k<q\leq K} w_{kq}\{I_{GS}(\boldsymbol{\pi}_q) - I_{GS}(\boldsymbol{\pi}_k)\} \\ &+ \sum_{1\leq k<q\leq K} w_{kq}\{I_{GS}(k,q) - \frac{1}{2}(I_{GS}(\boldsymbol{\pi}_k) + I_{GS}(\boldsymbol{\pi}_q))\} \\ &= \sum_{1\leq k<q\leq K} \frac{1}{2} w_{kq}\{I_{GS}(\boldsymbol{\pi}_q) + 2I_{GS}(k,q) - 3I_{GS}(\boldsymbol{\pi}_k)\}, \end{aligned} \tag{3.16}$$

where the w_{kq} are nonnegative weights tuned to the distance between k and q. For example, we could take the genetic distance in some genomic context where Euclidean distance may not work out. Also, based on suitable biological interpretations, such (pseudo-)distance

measures can be formulated in specific cases. As such, we let $d(k,q)$ be a suitable distance measure between the sites k and q, and let

$$w_{kq} = d(k,q)/\{\sum_{1\le k<q\le K} d(k,q)\},\ \forall (k,q), \tag{3.17}$$

then we have weights monotone increasing with the gap between the two positions; as $d(k,k)=0$, we may let (conventionally) that $w_{kk}=0,\ \forall k$. Whenever the Euclidean distance is meaningful, we take $d(k,q)=|q-k|, \forall k,q=1,\ldots,K$, so that the weights simplify to

$$w_{kq} = 6(q-k)/\{K(K^2-1)\},\ 1\le k<q\le K. \tag{3.18}$$

Under the null hypothesis of stochastic independence of the positions and homogeneity of their marginal probability vectors, $I_{GS}(\boldsymbol{\pi}_k) = I_{GS}(\boldsymbol{\pi}_q), \forall k\neq q=1,\ldots,K$ and $I_{GS}(k,q) = I_{GS}(\boldsymbol{\pi}_k),\ \forall k\neq q=1,\ldots,K$, and hence, $J(K)=0$, and it is nonnegative otherwise. For monotone nonincreading GSI's, we need to work with the measure $-J(K)$ which will be zero under the null hypothesis and nonnegative under monotone nonincreasing alternatives.

4 Constrained Statistical Inference

Consider a set of n sequences, each one containing a stochastic K-vector with categorical responses at each position, labelled as $1,\ldots,C$. Thus, we have a set of n i.i.d. stochastic vectors $\mathbf{X}_1,\ldots,\mathbf{X}_n$ where $\mathbf{X}_i=(X_{1i},\ldots,X_{Ki})',\ i=1,\ldots,n$, and

$$\begin{aligned} X_{ki} \ &= \ c, \ \text{if the response of the } \textit{ith} \text{ observation at the} \\ &\quad \textit{kth} \text{ position has the label } c, \ \ c=1,\ldots,C, \end{aligned} \tag{4.1}$$

for $k=1,\ldots,K;\ i=1,\ldots,n$. The probability of $\mathbf{X}_i=\mathbf{c}$ (where $\mathbf{c}=(c_1,\ldots,c_K)'$ with each c_k ranging over the set $\{1,\ldots,C\}$) follows a structured multinomial probability law over the set of C^K possible discrete realizations $\mathcal{S}_K=\{\mathbf{c}: c_k=1,\ldots,C,\ k=1,\ldots,K\}$. In genomic applications, typically, K is much larger than n, so that C^K would be telescopically larger than n.

There is therefore a genuine *curse of dimensionality* problem with statistical resolutions for such high-dimensional low-sample size categorical data models. The first task is therefore to reduce the number of unknown parameters by suitable dimension-reduction tools. Unfortunately, the classical *projection pursuit* tools are of very little use in this context. In classical multivariate normal populations, *multivariate analysis of variance* (MANOVA) models have been used extensively for external analysis or homogeneity of different groups. This problem is also encountered in genomics study (Pinheiro et al. 2000, 2005). For example, if we consider several groups of people in different geographical parts of the World (viz., Africa, South Asia, Europe, North America, Australia), depending on the extent of the HIV (human immunodeficiency virus) invasion and the type, the AIDS retrovirus might have reverse transcription to different extents, resulting in mutation in the genome possibly to different extent, so that classical MANOVA concepts are genuinely appealing in

drawing statistical inference on their possible homogeneity. The success of statistical tools in MANOVA primarily lies with the fundamental sub-group or additive decomposability of measures of dispersions, classically known as the *(M)ANOVA decomposability*. Even in normal theory models, with the increase in the dimension, the effectiveness of statistical tests rapidly breakes down unless the sample sizes increase at a very rapid rate. This is the basic curse of dimensionality problem.

The problem is more acute for purely qualitative categorical data models, as in here, where $K >> n$. Ordered alternatives are not easy to formulate unless suitable measures are used to reduce the dimension drastically. The Hamming distance is a first step in this direction, and has come out with a nice and interpretable statistical way for MANOVA in categorical data models. Decomposability of Hamming distance based sample measures has been extensively studied in the literature (Sen 1999, 2004; Pinheiro et al. 2000, 2005). As has been discussed before, the Hamming distance does not presume that the positions have independent and identically distributed response distribution. Nevertheless, it attaches equal weight to all the positions. For this reason, in the preceding section, we have considered some variants of the Hamming distance which have more natural appleal if there are some aprior information on the overall type of stochastic dependence among the positions.

We consider now two related ordered alternative problems for such high- dimensional genomic data models. First an internal analysis problem in the light of the findings in the preceding section. We want to test the null hypothesis that the different positions have the same Gini-Simpson index, against monotone alternatives in (3.1). For the kth position, an optimal unbiased nonparametric estimator of $I_{GS}(\boldsymbol{\pi}_k)$ is the sample U-statistic (Hoeffding 1948):

$$\begin{aligned} U_{n,k} &= \binom{n}{2}^{-1} \sum_{1\le i<j\le n} I(X_{ki} \neq X_{kj}) \\ &= \sum_{c=1}^{C} \{n_{kc}(n-n_{kc})\}/\{n(n-1)\}, k = 1, \ldots, K, \end{aligned} \tag{4.2}$$

where n_{kc} is the number of observations having the label c in the kth position, for $c = 1, \ldots, C$. These U-statistics are generally dependent and possibly nonidentically distributed. Further, we have noted in (2.5 - (2.7) that by virtue of (2.7) and the monotone nature of the marginal I_{GS}, there is additional information that we need to take into account. The sample counterpart of $I_{GS}(k, q)$ is a U-statistic too; it is given by

$$U_{n;k,q} = \sum_{c=1}^{C} \{n_{kc}(n-n_{qc})\}/\{n(n-1)\}, \tag{4.3}$$

for $1 \le k < q \le K$. These U-statistics are also possibly not independent of each other or of the marginal U-statistics. In passing, we may remark that in a similarl way, an unbiased

estimator of $CD_{GS}(k,q)$ is

$$M_{n;k,q} = 1 - n^{-1}\sum_{c=1}^{C} n_{kq;cc},\ k \neq q = 1,\ldots,K, \tag{4.4}$$

where $n_{kq;cc}$ is the number of observation having the label c in both the kth and qth positions. In fact, $M_{n;k,q}$ is also a U-statistic corresponding to a kernel of degree 1 (as such having independent and identically distributed summands).

To test for the hypothesis of homogeneity of the $I_{GS}(\boldsymbol{\pi}_k),\ k = 1,\ldots,K$, against a monotone nondecreasing pattern, under the assumption (3.14), can be based on the sample counterpart of (3.16). This is given by

$$\begin{aligned}
\hat{J}_n(K) &= \sum_{1\le k<q\le K} \frac{1}{2} w_{kq}\{U_{n,q} + 2U_{n;k,q} - 3U_{n,k}\} \\
&= \sum_{1\le k<q\le K} \frac{w_{kq}}{2n(n-1)}\{n_{qc}(n-n_{qc}) + 2n_{kc}(n-n_{qc}) - 3n_{kc}(n-n_{kc})\} \\
&= \sum_{1\le k<q\le K} w_{kq}\{(n_{qc}-n_{kc})(n-n_{qc}-3n_{kc}(n_{qc}-n_{kc})\}/\{2n(n-1)\} \\
&= \sum_{1\le k<q\le K} w_{kq}\{(n_{qc}-n_{kc})(n-4n_{qc}+3n_{kc})\}/\{2n(n-1)\}. \qquad (4.5)
\end{aligned}$$

It may be noted that $\hat{J}_n(K)$ is itself a U-statistic based on a kernel of degree 2, which we denote by $\phi_w(\mathbf{X}_i, \mathbf{X}_j)$. Therefore the asymptotic normality results follow from the classical results of Hoeffding (1948). There is, however, something more to note in this context.

By definition, the kernel is an weighted average of $K(K+1)/2$ kernels, which for any pair of observations (vectors) may not be all stochastically independent. If they were independent, the variance of kernel would have been $O(K^{-1})$. Even without such a strong assumption of independence of all the K positions, it can be shown that under fairly general inter-position stochastic dependence, especially under a suitable mixing dependence condition, when K is large, the kernel $\phi_w(.)$ has variance $O(K^{-1})$. As such, if K is large, as is generally the case in genomics, we would have an advantage that the standardizing factor in the asymptotic distribution of $\hat{J}_n(K)$ would be $\sqrt{Kn}$ instead of $\sqrt{n}$. This also makes it possible to have smaller values of n eligible for the good asymptotic normality approximation when K is not small. In the contemplated situation where $K >> n$, this not only makes the result applicable in a broader setup but also leads to a faster rate of convergence. On the other hand, the variance of the statistic $\hat{J}_n(K)$ (even under the null hypothesis of homogeneity) depends in an intricate way on the interposition stochastic dependence pattern. Hence there is a genuine need to estimate this variance from the sample data in a suitable way. Although, technically, it can be shown that the variance functional is itself a U-statistic of degree 4, and hence can be unbiasedly estimated from the sample, such an estimator would be in general quite computationally cumbersome. For this reason, we use the classical jackknife method to estimate this variance.

Note that $\hat{J}_n(K)$ being a U-statistic unbiasedly estimates $J(K)$. If we denote the entire set of n sequences by $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ then on taking out $\mathbf{X}_i$, the ith column of $\mathbf{X}$ we denote the resulting set of sequences by $\mathbf{X}^{(-i)}$, for $i = 1, \ldots, n$; the corresponding statistics are denoted by $\hat{J}_{n-1}^{(-i)}, i = 1, \ldots, n$. Each of these statistics is an unbiased estimator of the common $J(K)$. Let then

$$\hat{J}_{n,i}(K) = n\hat{J}_n(K) - (n-1)\hat{J}_{n-1}^{(-i)},\ i = 1, \ldots, n. \tag{4.6}$$

These are the so called pseudo-values, and in the present case, they are all unbiased for $J(K)$. As such, the jackknife vesrion of $\hat{J}_n(K)$, the average of these pseudo-values, agrees with the estimate itself; because of unbiasedness, there is no question of bias reduction. However, these pseudo-values can be incorporated in the variance estimation. We define then the Tukey variance estimator by

$$\hat{V}_n(K) = (n-1)^{-1} \sum_{i=1}^{n} \{\hat{J}_{n,i} - \hat{J}_n(K)\}^2. \tag{4.7}$$

It follows that $\hat{V}_n(K)$ is a consistent estimator of the variance of $\sqrt{n}\{\hat{J}_n(K) - J(K)\}$, in the sense that the ratio of the two converges stochastically (in fact, almost surely) as n increases. Therefore, under quite general regularity conditions and without being restricted to the null hypothesis, we claim that as n increases,

$$Z_n = \sqrt{n}(\hat{J}_n(K) - J(K))/\sqrt{\hat{V}_n(K)} \ \stackrel{\mathcal{L}}{\Rightarrow}\ \mathcal{N}(0,1). \tag{4.8}$$

Now under the null hypothesis, $J(K) = 0$ while under the contemplated alternatives, $J(K)$ is positive. Hence, a one-sided test appears to be the natural case. With that in mind, we consider the test statistic as

$$T_n = \sqrt{n}\hat{J}_n(K)/\sqrt{\hat{V}_n(K)} \tag{4.9}$$

and reject the null hypothesis when T_n exceeds a critical level $\tau_n(\alpha)$ corresponding to a level of significance $\alpha (0 < \alpha < 1)$. We can approximate well $\tau_n(\alpha)$ by the upper α-quantile of the standard normal distribution.

It might be interesting to note that in genomic studies, usually K is large while n, albeit large, is usually much smaller than K, i.e., we have the situation where $K >> n$. Although individually each indicator function is zero-one valued, the statistic $\hat{J}_n(K)$ involves a kernel which is an weigted average of $K(K+1)/2$ such indicator function. As such, under a suitable weak dependence condition, the kernel when properly standardized (by the scale factor $K^{1/2}$ and centering $J(K)$) is asymptotically (in K) normal with zero mean and a finite variance. It is also worth noting that we are not assuming the homogeneity of the marginal multinomial laws (nor their independence). Hence, though it might be tempting to use a double-jackknife method that incorporates jackknifing on the K positions in addition to the jackknifing on the individual sequences, a theoretical justification is still mot established. If, however, the positions in a sequence exhibit a stationary process, then such a double-jackknife method would work out to our advantage. We have not discussed the bootstrap

methodology. That is also routinely applicable on the sequences. On the other hand, as the weighted kernel is asymptotically (in K) normal, small sample adjustments for bootstrap methods (to improve the approximation) can also be made effectively. This feature enables us to use the jackknife variance estimator $\hat{V}_n(K)$ even for moderate values of n when K is large. This interesting feature is unique with this Hamming-type distance measures as contrasted with conventional measures that run into the curse of dimensionality problem to a much greater extent.

5 Some Illustrations

While the curse of dimensionality is prevalent in many fields of application, here, we confine ourselves to genomic studies where such a formulation of weighted Hamming distance has a lot of useful applications. As illustrative examples, we consider the following.

Consider a SNP (single nucleotide polymorphism) data model where there are K (usually very large) positions, and at each position the response is one of the 4 nucleotides A, C, G and T. Typically, we have a number (n) of such sequences where $n << K$. The responses at these positions can not be generally taken to be independent nor identically distributed. In the context of AIDS and HIV (human immunodeficiency virus) studies, the scientific focus is on the genetic variability of SNP's. We may note that retrovirus, like HIV, has the ability to reverse the normal flow of genetic information from genomic DNA, and that the genetic variability of HIV is relatively high compared to other retroviruses (Coffin 1986). This way, we encounter a typically high-dimensional purely qualitative multivariate stochastic process, and the covariability aspects are of considerable study-importance. The HIV retrovirus also distorts plausible stationarity of the responses over the positions. For the special case of a pair of positions, Karnoub et al. (1999) considered some conditional tests of independence of mutations, and studied their large sample perspectives. It is not uncommon to have smaller sequence sizes and a large number of positions. It remains to see how their proposed methodology provides satisfactory resolutions in such a large K and small n case. Conventional internal (multivariate) analysis tools (such as the canonical correlation, principal component model, factor analysis) are of limited utility in this high-dimensional discrete set-up. Also, the modern data miniting tools, although very much used in genomics studies, need more statistical foundations to facilitate at least the statistical modeling and analysis perspectives (Durbin et al. 1998, Ewens and Grant 2001, Waterman 1995).

Motivated by the immense need of identification of disease genes in human diseases, there is a challenging problem of mapping the genetic basis which is characterized by high-heterogeneity accompanied by multiple causative loci with probably multiple alleles at these causative loci. Association of multiple genes with specific human disease(s) is therefore very much in the mind of genomic researchers. Use of multiple markers has been an important genomic approach to this mapping of disease genes. However, from statistical modeling and analysis perspectives of such complex association (and causative relations), there remains

a lot to accomplish. Schaid et al. (2005) have incorporated the Hamming distance type measures, and this approach needs more appraisal in the light of monotone dependence in some meaningful ways. There is some incompleteness in dealing with the complex statistical distribution theory of such compound measures, as typically, $K >> n$, and even n could be small. In the same vein, Tzeng et al. (2003) virtually used the Gini-Simpson index and compounded them over the loci or positions. The pertinent fact is that these positions may not have independent responses, and in combining the locus-specific measures into one is therefore subject to the same question of very high-dimension and low sample size. The present study, done independently of their work, provides some complementary methodology. It is hoped that it would bridge the gap to a certain extent.

As of now, researchers confronted with this curse of dimensionality problems in genomic studies, mostly assume that the positions have independent and identically distributed responses. At the rampage of HIV, it is expected that a monotone genetic variability (diversity) across the positions is more likely to be the case (whenever the positions are ordered in an accessible manner. As such, the methodology discussed in the last two sections should be of immense help in statistical spatial-analysis of SNP's in the presence of retroviruses. In genetics, multi-factorial genetics is coming up as a natural contender of the classical Mendelian genetics in many polygenic models. Pharmacogenomics and the drug-discovery ventures associated with disease genes have raised the necessity of looking into genetic variation in a much broader setup. It is our hope that the proposed methodology would find its way into statistical analysis and modeling of such enormously large dimensional categorical data models.

Acknowledgements

The author is grateful to the reviewers for their most helpful comments which clarified some of the ambiguities of the earlier version. This research was supported by the Cary C. Boshamer endowment fund at the University of North Carolina, Chapel Hill.

References

[1] Chakraborty, R. and Rao, C. R. (1991). Measurement of genetic variation for evolutionary studies. In *Handbook of Statistics, Volume 8: Statistical Methods in Biological and Medical Sciences* (eds: C. R. Rao and R. Chakraborty), Elsevier, Amsterdam, pp. 271-316.

[2] Coffin, J. M. (1986). Genetic variation in AIDS virus. *Cell 46*, 1-4.

[3] Crick, F. H. C. (1970). Central dogma of molecular biology. *Nature, 227*, 561 - 563.

[4] Durbin R., Eddy, S. Krogh, A. and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models for Proteins and Nucleic Acids*, Cambridge University Press, England.

[5] Ewens, W. J. and Grant, G. R. (2001). *Statistical Methods in Bioinformatics: An Introduction*, Springer-Verlag, New York.

[6] Gini, C. W. (1912). Variability e mutabilita. *Studi Economico-Giuridici della R. University de Calgiary, (2)*, 3 - 159.

[7] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning,* Springler-Verlag, New York.

[8] Hoeffding, W. (1948). On a class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics 19*, 293 - 325.

[9] Karnoub, M. C., Seillier-Moiseiwitsch, F. and Sen, P. K. (1999). A conditional approach to the detection of correlated mutations. *Institute of Mathematical Statistics Lecture Notes - Monograph Series 33*, 221 - 234.

[10] Pinheiro, H. P., Seillier-Moiseiwitsch, F., Sen, P. K. and Eron, J. (2000), Genomic sequence analysis and quasi-multivariate CATANOVA. *Handbook of Statistics, Volume 18 : Bioenvironmental and Public Health Statistics* (eds: P. K. Sen and C. R. Rao), Elsevier, Amsterdam, pp. 713 - 746.

[11] Pinheiro, H. P. , Pinheiro, A. S. and Sen, P. K. (2005). Comparison of genomic sequences using the Hamming distance. *Journal of Statistical Planning and Inference, 130*, 323-339.

[12] Rao, C. R. (1982 a). Diversity and dissimilarity coefficients: A unified approach. *heoretical Population Biology, 21*, 24 - 43.

[13] Rao, C. R. (1982 b). Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhya, A 44*, 1 - 21.

[14] Rao, C. R. (1982 c). Gini-Simpson index of diversity : A characterization, generalization and applications. *Utilitus Mathematica 21*, 273 - 282.

[15] Schaid, D.J., McDonnell, S. K., Hebbring, S. J., Cunningham, J. M. and Thibodeau, S. N. (2005). Nonparametric tests of association of multiple genes with human disease. *Amer. J. Human Genet. 76*, 780 - 793.

[16] Sen, P. K. (1999). Utility-oriented Simpson-type indexes and inequality measures. *Calcutta Statistical Association Bulletin 49*, 1 - 22.

[17] Sen, P. K. (2004). *Excursions in Biostochastics: Biometry to Biostatistics to Bioinformatics*, Invited Lecture Series, No. 5, May 2001, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan.

[18] Silvapulle, M. J. and Sen, P. K. (2004). *Constrained Statistical Inference: Inequality, Order, and Shape Restraints*, Wiley and Sons, New York.

[19] Simpson, E. H. (1949). The measurement of diversity. *Nature 163*, 168.

[20] Tzeng J.-Y., Byerley W., DevlinB., Roeder, K. and Wasserman, L. (2003). Outlier detection and false discovery rate for whole genome DNA matching. *Jour. Amer. Statist. Assoc. 98*, 236 - 246.

[21] Waterman, M. (1995).*Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman-Hall, London, UK.