GINI DIVERSITY INDEX, HAMMING DISTANCE AND CURSE OF DIMENSIONALITY

PRANAB K. SEN
Departments of Biostatistics, and Statistics and Operations Research,
University of North Carolina at Chapel Hill, NC27599-7420, USA

Abstract

The celebrated Gini(-Simpson) biodiversity index has found very useful applications in ecology, bio-environmetrics, econometry, psychometry, genetics, and lately in bioinformatics as well. In such applications, mostly, categorical data models, without possibly an ordering of the categories, crop up, which may preempt routine use of conventional measures of quantitative diversity analysis. Further, in real life problems, mostly, genuine multidimensional data models are encountered. The Hamming distance incorporates the idea of Gini-Simpson diversity index in a variety of multidimensional setups, without making very stringent structural regularity assumptions. In bioinformatics as well as many other large biological system analysis studies, the curse of dimensionality (arising in multidimensional purely qualitative categorical data models) is a geneuine concern. The role of Hamming distance based analysis is appraised in this context. Subgroup or MANOVA decomposability aspects are specially appraised in this setup.

1. Introduction

Variation (or diversity) abounds in various statistical models for quantitative as well as qualitative response variables. For continuous or discrete random variables, the well known measures of central tendency and variation (or dispersion) are respectively the conventional mean and standard deviation. The popularity of these measures stems from the fact that for a normal distribution, they characterise the location and scale parameters of the distribution. For non-normal distributions, the mean and standard deviation may not characterise the location and scale (even if we confine ourselves to the so called location-scale family of densitites), and as such, these measures may not have natural appeal. This problem is more acute with the standard deviation than the mean. Their estimation may also encounter lack of robustness and optimality properties (to a lesser or greater extent depending on

the departure from normality). Consideration of a location-scale family of distributions provide a basis for robust measaures, and some of these have been formulated in a greater generality in a nonparametric fashion too. As such for conventional quantitative data models, often, robustness and efficiency considerations prompt some alternative measures which are less sensitive to outliers, heavy tails etc.. The median, inter-quartile range, and various $M-, L-$ and $R-$estimators of location and scale parameters have found their utility in statistical appraisal of central tentency and dispersion of quantitative data models (viz., Jurečková and Sen 1996). There are some quasi-quantitative models where a set of categorical responses with at least a partial ordering of the categories may be observed. For such binary or polytomous categorical data models arising in many fields of application (viz., social networks, psychometry (item analysis) and biological (quantal) assays), often a continuous underlying trait is conceived so that measures of location and variation can be formulated in terms of such latent trait variables. This approach has been generalized to a much broader set-up under suitable generalized linear models (GLM) (viz., McCullagh and Nelder 1989). In an intermediate scenario there are data models relating to a set of ordered class-intervals, albeit the underlying variable may well be continuous. Specific parametric form of the underlying continuous distribution may often be difficult to comprehend, and for simplicity of modeling and analysis, often it is assumed to be normal. In such setups, nonparametrics has flared up in natural ways to supplement conventional measures with more robust and meaningful alternatives. Sans quantitative labels, the concept of an underlying latent trait variable may not be very appealing, and hence, such conventional measures of central tendency and dispersion are generally not meaningful.

The situtation is quite different for purely qualitative categorical data models where the categories differ qualitatively and there may not be any implicit ordering of these categories. For example, in a simple categorical model with the response categories being the color of eye-ball, the variation is purely qualitative, and no ordering of them may be perceived. Similarly, in a simple (genetic) linkage model, there are 4 genotypic categories, AA, Aa, aA and aa with dominant A and recessive a. In genomics where the genes show up in very large numbers, for DNA (a double hellic model) with the nucleotides A, C, G and T, at each site or position, there are the 4 categories without any implicit ordering, albeit A pairing with T and G with C. For RNA codons there are some 20 amino acids without any ordering in a quantitative sense. In such a case, a measure of location is not that meaningful and therefore of much interest, albeit, there could be some distinct variation or diversity (of purely qualitative type) in the data models that needs to be statistically addressed properly.

Gini (1912) came up with a very interesting measure of *lack of concentration* or *diversity* that opened the avenue for further fruitful research in diversity analysis of qualitative categorical data models. Simpson (1949), apparently being unaware of the Gini fundamental contribution (measure), proposed the same measure for *bio-diversity* in an ecological context. Our main interest centers in this Gini-Simpson index of diversity and its impact in a large class of fields of application in many interdisciplinary fields; in this context, quite often an enormously large dimensional data set

crops up, often with relatively smaller number of observations. In this respect, the Gini-Simpson index has been generalized to what is known as the Hamming distance. We intend to probe into the MANOVA (multivariate analysis of variance-) or subgroup-decomposability of such measures, and their underlying statistical perspectives. The Shannon (1948) entropy measure, and its variants, such as the Rao quadratic entropy measure and the utility-oriented Gini-Simpson Index also deserve attention.

The Gini-Simpson diversity index and its ramifications are reconciled in Section 2. Along with some natural multi-dimensional ramifications, a general form of the Hamming distance is presented in Section 3. ANOVA or subgroup decomposability of these indexes is appraised in Section 4. Keeping in mind the very high-dimensional models, typically, arising in bioinformatics and large biological systems studies, the curse of dimensionality problem is critically assessed in Section 5. The concluding section deals with some general asymptotics relevant to the present set-up.

2. The Gini-Simpson Index

Keeping in mind that a measure of diversity or concentration should have a meaningful physical interpretation, we may gather that for purely qualitative (categorical) data models, characterized by the absence of any interpretable quantitative variation, the usual measures of variation (advoated for quantitative data models) may not be at all suitable. Since in such qualitative data models, the individual cell (or category) frequencies (or probabilities) convey all statistical information, a suitable measure of lack of concentration or diversity is to be based on these entities only, and they should have good physical interpretation. There are some alternative thinking in this respect based on three very novel ideas: (i) the Gini-Simpson index (GSI), introduced by Gini (1912), quite sometimes before the others, (ii) Wiener's (1948) cybernetics theory, and (iii) Shannon entropy measure (SEM). Simpson (1949) had a distinct emphasis on bio-diversity while Shannon (1948) laid down the foundation of the information theory where the role of the entropy measure has been explored to a maximum extent. A more statistical treatise of information theory is contained in Kullback (1959). Cybernetics seems to have a very dominant role in statistical chaos theory, and it is also getting more attention in information theory, artificial intelligence etc.. The Gini-Simpson index, as will be formulated now, seems to have a very natural interpretation as would be detailed below.

Consider a categorical data model with C qualitative (and possibly unordered) categories, labelled as $1,\ldots,C$ having respective cell probabilities $\pi_1,\ldots,\pi_C$. We may note then if all the cells are equally probable, then there is no concentration so that the diversity is a maximum. In the other extreme case where one cell has the unit probability and the others have all null probability, then there is a maximum concentration in one of the categories (or minimum diversity). Thus, a measure of diversity (or lack of concentration) should satisfy these two restraints, namely, that it is equal to zero if we have a degenerate probability law with only one cell having the entire probability mass, and it is equal to

one (if so normalized) if all the cells are equally probable. Of course, there should be some notion of ordering of diversity in this setup, and that aspect would be elaborated as we proceed. The GSI (Gini 1912) capitalizes this idea and is expressed as

$$\begin{aligned} I_{GS}(\pi) &= 1-\pi'\pi \\ &= 1-\sum_{c=1}^{C}\pi_c^2. \end{aligned} \tag{2.1}$$

Note that $\pi \in S_{C-1}$, where S_{C-1} is the $(C-1)$-simplex $(=\{\mathbf{x}:\mathbf{x}\in[0,1]^C, \sum_{c=1}^{C}x_c=1\}$. The C vertices of this simplex correspond to the degenerate cases $\pi_j=1, \pi_k=0, \forall k\neq j;\ j=1,\ldots,C$. At each vertex, $I_{GS}(\pi)=0$, so that $I_{GS}(\pi)=0$ on the vertices of the simplex. Similarly, the centroid of the simplex corresponds to $\pi=C^{-1}\mathbf{1}$ where I_{GS} attains a maximum value $1-C^{-1}$. For this reason, a normalized form of the I_{GS} is defined as

$$I^*_{GS}(\pi)=\frac{C}{C-1}I_{GS}(\pi). \tag{2.2}$$

Keeping this picture in mind, we express the simplex S_{C-1} as the set-theoretic union of the contours $S_{C-1}(p):\ p\in[0,1]$, where

$$S_{C-1}(p)=\{\mathbf{x}\in S_{C-1}:\mathbf{x}'\mathbf{x}=1-p\},\ 0\le p\le 1. \tag{2.3}$$

Note that by virtue of (2.2), the upper limit for p is $(C-1)/C$ (and not one). This set-theoretic union also defines an ordering in terms of diversity, and this will be termed the Gini-ordering of the probability on a simplex. In passing we may note that if we consider a sphere of radius $1-p$ and center at the origin, then the sphere intersects the Simplex S_{C-1} in a contour which is $S_{C-1}(p)$. If $p=0$ this contour relates to the discrete set of C vertices of the simplex, and if $p=1-C^{-1}$ then this contour reduces to the single point which is the centroid of the simplex S_{C-1}. For $1-C^{-1}<p\le 1$, the sphere of radius $1-p$ fails to intersect the simplex, while for $0<p<1-C^{-1}$, the contour $S_{C-1}(p)$ is a $C-1$ dimensional spherical surface on S_{C-1}, either contained in the simplex or having segments that belong to the simplex, and the ordering of the contours $S_{C-1}(p)$ is visibly based on these annular spherical contours which are solely characterized by their radius. The following figure depicts the ordering in the particular case of $C=3$, where S_2 is an equilateral triangle with the centroid with coordinates (1/3, 1/3, 1/3) and 3 vertices (1,0,0), (0,1,0) and (0,0,1), all located on the original 3-dimensional unit cube. A similar picture holds for any $C\ge 3$,

Figure 1 : A display of the annular contours

This Gini-ordering will be of genuine importance when in a later section we consider some statistical inference problems based on the GSI and its ramifications.

With the same notations as before, for the C-category qualitative data model, the Shannon (1948) entropy measure is defined as

$$\mathcal{E}(\pi) = -\sum_{c=1}^{C} \pi_c \log \pi_c, \tag{2.4}$$

and it is also defined on S_{C-1}. Further note that $x \log x$ is equal to zero when $x = 0$ or 1, and for every $0 < x < 1$,

$$-\log x = -log(1-(1-x)) = \sum_{j\geq 1} j^{-1}(1-x)^j, \tag{2.5}$$

so that

$$\mathcal{E}(\pi) = \sum_{j\geq 1} j^{-1} \sum_{c=1}^{C} \pi_c (1-\pi_c)^j \geq I_{GS}(\pi). \tag{2.6}$$

As a result (Sen 1999), the Entropy may have some inflationary tendency (relative to the GSI). In fact, Rao (1982a,b,c) examined the role of the entropy measure in some genetic variation in evolutionary studies (see also Chakraborty and rao 1992), and Rao has clearly pointed out the limitations of $\mathcal{E}(\pi)$ in the context of measuring biological diversity. Rao suggested some modifications of the entropy measure that we shall only discuss briefly.

The Gini-Simpson index has been widely used in various interdisciplinary fields. In some of these usages, there could be additional information associated with the C categories, such as some (implicit) ordering (not necessarily linear) or some utilty-levels which can be incorporated in a meaningful way. Sen (1999) and Chatterjee and Sen (2000) elaborated some of these measures and their characteristics in the context of poverty, income inequality and also in quality of life studies. We shall discuss them briefly later on. Other developments relate to similar measures based on more sophisticated distance function. In this respect, Rao's quadratic entropy function and its generalizations by Nayak (1986a, b) and Nayak and Gastwirth (1989) are especially noteworthy. Basically, they considered a $C \times C$ matrix $\Delta = ((d_{cc'}))$ where the $d_{cc'}$ stand for suitable distance between the category c and c^l (so that by definition, $d_{cc} = 0, \forall c = 1, \ldots, C$). Then the Rao quadratic entropy measure is expressed as

$$\mathcal{H}(\pi) = \pi' \Delta \pi = 2 \sum_{1\leq c<c'\leq C} d_{cc'} \pi_c \pi_{c'}. \tag{2.7}$$

If we let all the $d_{cc'}$ to be equal to 1 then the quadratic entropy measure reduces to the Gini-Simpson index. In passing, we may remark that for purely qualitative categorical data models, a meaningful choice of the $d_{cc'}$ (other than all being equal to 1) may not be always feasible and hence the Gini-Simpson index has a more natural appeal in this context. Even for some of these purely qualitative data models, some utility scores $u(c), c = 1, \ldots, C$ can be attached in a meaningful way. In such a

case, a utility-oriented Gini-Simpson index (Sen 1999) can be defined as

$$I_{UGS}(\pi) = \sum_{c=1}^{C} u(c)\pi_c(1-\pi_c) \tag{2.8}$$

Typically the utility scores are so standardized that they are nonnegative and lie in the interval $[0.1]$. Again, if all these scores are equal to 1, then the I_{UGS} reduces to the I_{GS}. We shall find it more convenient to work with such a utility-oriented Gini-Simpson inndex, and as such, our findings apply to the Gini-Simpson index as well. If there is an implicit ordering (though not necessarily linear) of the labels $1,\ldots,C$, so that $u(c)$ is monotone in $c(=1,\ldots,C)$, then it might be reasonable to let (Chatterjee and Sen 2000) $d_{cc'} = u(c \cap c')$, for $c,c' = 1,\ldots,C$. In that case the utility-oriented Gini-Simpson index resembles a quadratic entropy measure. However, in general, neither is a particular case of the other. For such monotone utilities, some utility-oriended Gini-Simpson Indexes have some 'tiltedness' properties which are useful in more detailed analysis of diversity with emphasis on the underlying ordering; we refer to Chatterjee and Sen (2000) for a detailed treatise of these properties with emphasis on income inequality, poverty, and quality of life studies. It has been shown there that for ordered categorical data models, the Lorenz-ordering of distributions and their'tiltedness' are not necessarily isomorphic. For such quasi-quantitative models, we have an implicit ordering in terms of the 'tiltedness' of the probability laws, providing a bit more information than the simple Gini-ordering. However, sans such an implicit ordering, utility-scores need to be assessed with special reference to the specific problem at hand, and may not be universally advocated.

3. The Hamming distance

Keeping in mind some genomic studies, we consider here a typical scenario where we have a number (K) of positions or sites, where at each position or site, we have a number (C) of qualitative (and usually unordered) categories. An object, in each position, can take on one of the C possible labels $1,\ldots,C$, there being thus a totality of C^K possible realisations for each object. Thus, if we consider a two-way table with K columns and C rows, then within each column we would a (random) row containing the number 1 while the other $C-1$ rows have the number 0. It is therefore possible that two or more positions may have the manifestation of the same level c for some $c = 1,\ldots,C$. Further, no column is empty, albeit, some of the rows may be empty (for a particular object).

To describe this model, we consider a set of n vectors $\mathbf{X}_i = (X_{i1},\ldots,X_{iK})',\ i = 1,\ldots,n$ where X_{ik} stands for the particular label $(1,\ldots,C)$ for the ith observation in the kth position, for $k = 1,\ldots,K$. Thus, each $\mathbf{X}_i$ has a totality of C^K possible realizations; we denote the corresponding probability law by a multi-dimensional multinomial law with the probability elements

$$\Pi = ((\pi(\mathbf{c}))),\ \mathbf{c} \in C, \tag{3.1}$$

where $\mathbf{c} = (c_1, \ldots, c_K)'$ with each c_j taking on the labels $1, \ldots, C$, so that

$$\mathcal{C} = \{\mathbf{c} : c_j = 1, \ldots, C; j = 1, \ldots, K\}. \tag{3.2}$$

The cardinality of $\mathcal{C}$ is C^K. In some applications in genomics, it is tacitly assumed that the K positions are stochastically independent (with respect to the manifestation of the labels $1, \ldots, C$), so that we may then take

$$\pi(\mathbf{c}) = \prod_{k=1}^{K} \pi_k(c_k), \ \forall \mathbf{c} \in \mathcal{C}, \tag{3.3}$$

where the $\pi_k(.)$ are the marginal probability elements for the kth position. In some cases, it is even assumed that these marginals are the same, so that the positions are then assumed to have independent and identically distributed manifestations. In real life applications, however, neither the assumption of independence nor homogeneity of the marginal multinomial laws may turn out to be reasonable, and hence, we would like to deal with the general model in (3.1) allowing possible dependence as well as heterogeneity.

One possible way to describe the probability law for $\mathbf{X}_i$ is to write the joint probability as

$$P\{\ \mathbf{X}_i = \mathbf{c}\ \} = P\{X_{i1} = c_1\} \prod_{k=2}^{K} P\{X_{ik} = c_k | X_{ij} = c_j, j \leq k-1\}, \ \mathbf{c} \in \mathcal{C}. \tag{3.4}$$

It is easily conceivable that we have a transition from the label c_{k-1} to c_k, at the kth position, for $k = 1, \ldots, K$, where c_0 is treated as the trivial (or unconditional) element. As such, if we could assume a Markov chain model, then we could simplify the model (3.1) in terms of the transition probabilities and the marginal one for the first position. In that case, the number of unknown probabilities in the model reduces (from $C^K - 1$) to $C - 1 + (K-1)C(C-1)$, while in the stationary case, it reduces further to $C^2 - 1$. Again such a Markov chain assumption may not be generally tenable in genomics and many other applications. Sans the Markov chain assumption, the number of parameters in the model can become unmanageably large as K or C becomes large - as is typically the case in real applications.

For diversity analysis for such high-dimensional qualitative categorical data models, we take recourse to the Gini-Simpson index and extend it in a natural way to accomodate possible dependence as well as heterogeneity of the marginals. For two vectors $\mathbf{X}_i, \mathbf{X}_j, \ i \neq j$, each following the probability law in (3.1), the Hamming distance is defined as

$$d_{ij} = K^{-1} \sum_{k=1}^{K} I(X_{ik} \neq X_{jk}), \tag{3.5}$$

Thus, for the entire sample of n vectors, we have the Hamming distance

$$D_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} d_{ij} \tag{3.6}$$

Note that as in the case of the GSI, the Hamming distance based measure D_n is a U-statistic corresponding to a symmetric kernel of degree 2. As such, D_n is an optimal nonparametric estimator of its population counterpart (which is an estimable parameter in the sense of Hoeffding (1948)):

$$\begin{aligned}\Delta &= E[D_n] = K^{-1}\sum_{k=1}^{K} P\{X_{ik} \neq X_{jk}\} \\ &= K^{-1}\sum_{k=1}^{K}\{1-\sum_{c=1}^{C}\pi_k^2(c)\} \\ &= K^{-1}\sum_{k=1}^{K} I_{GS}(\pi_k), \end{aligned} \tag{3.7}$$

which is the average of the marginal Gini-Simpson indexes. Recall that for each $k(=1,\ldots,K)$, the sample counterpart of I_{GS} is

$$\begin{aligned}U_{nk} &= \binom{n}{2}^{-1}\sum_{1\le i<j\le n} I(X_{ik}\neq X_{jk}) \\ &= \sum_{c=1}^{C}\frac{n_{kc}(n-n_{kc})}{n(n-1)},\ k=1,\ldots,K, \end{aligned} \tag{3.8}$$

where n_{kc} is the number of observations in the kth position having the label c, for $c=1,\ldots,C$, and $n=\sum_{c=1}^{C} n_{kc}$, for all $k=1,\ldots,K$. Thus,

$$D_n = K^{-1}\sum_{k=1}^{K}\frac{\sum_{c=1}^{C} n_{kc}(n-n_{kc})}{n(n-1)}. \tag{3.9}$$

We have observed in the previous section that there is a nice ordering of contours on the simplex S_{C-1} based on the ordered values of the GSI. In the present case, the Hamming distance D_n or its population counterpart Δ is an average over K GSI; though each of them has an ordering similar to that in Section 2, their average may not have this ordering simply in terms of constant Hamming distance values. This is not surprising as even for the continuous (multivariate) case such an ordering demands for delicate structure on the parameters. For example, for the multi-sample multinormal dispersion problem, instead of the variance ordering, we need a matrix-version resulting in an ordering of the eigenvalues of the dispersion matrix, i.e., the difference of two such positive definite matrices being positive semi-definite. In the present case, consider two probability matrices Π_1 and Π_2 (pertaining to model (3.1)) with the respective marginal probability vectors π_{1k} and π_{2k}, for $k=1,\ldots,K$. Then we define a more stringent ordering (very similar to the layer alternatives in the multivariate location model) as : Π_1 has more (layer) diversity than Π_2 if

$$I_{GS}(\pi_{1k}) \ \geq\ I_{GS}(\pi_{2k}),\ \forall k=1,\ldots,K, \tag{3.10}$$

with the strict inequality for at least some $k(=1,\ldots,K)$. Although the Δ-constant contours are no longer spherical, this partial ordering can be used to order the Hamming distances in a manageable way. Of course, if all the marginal probability laws are the same, then the Hamming distance is the same as the common Gini-Simpson index, and hence, the Gini-ordering considered in Section 2 would remain in tact.

In applications in bioinformatics (and genomics), often the different positions may have different weights, so that we could consider an immediate generalisation of D_n to accomodate their relative importance. We conceive of a set of nonnegative weights $w_k, k = 1,\ldots,K$ and standardise them by letting $\sum_{k=1}^{K} w_k = 1$. Then, a weighted version of the Hamming distance can easily be conceived as a convex (linear) combination of the U_{nk} as

$$\begin{aligned} D_n^* &= \sum_{k=1}^{K} w_k U_{nk} \\ &= \sum_{k=1}^{K} w_k \sum_{c=1}^{C} \frac{n_{kc}(n-n_{kc})}{n(n-1)}. \end{aligned} \tag{3.11}$$

It is also possible (whenever utility-score could be validly attached to the different labels) to cosider a utility-oriented weighted Hamming distance based measure:

$$\begin{aligned} D_n^*(\mathbf{w},\mathbf{u}) &= \sum_{k=1}^{K} w_k U_{nk}(\mathbf{u}) \\ &= \sum_{k=1}^{K} w_k \sum_{c=1}^{C} \frac{u_c n_{kc}(n-n_{kc})}{n(n-1)}. \end{aligned} \tag{3.12}$$

In some applications, a natural measure of distance between two positions can be conceived; the genetic distance in some multifactorial genetic models is a classical example of such a distance measure. In such a case, for every pair $(k,q) : 1 \le k < q \le K$ of positions, we conceive of a suitable distance δ_{kq}. By definition, the δ_{kq} are all nonnegative. In that case, we may consider a Gini-Simpson index for the pair (k,q) of positions. We define these between position GSI as U_{nkq}, where

$$U_{nkq} = n^{-1} \sum_{i=1}^{n} I(X_{ik} \ne X_{iq}),\ k \ne q = 1,\ldots,K. \tag{3.13}$$

Note that the pair (X_{ik}, X_{iq}) may not have stochastically independent coordinates, but still the U_{nkq} qualify for a U-statistic based on a kernel of degree 1. Further, $U_{nkk} = 1,\ \forall k = 1,\ldots,K$. We define then

$$D_n^o(\delta) = \sum_{1 \le k < q \le K} \delta_{kq} U_{nkq}. \tag{3.14}$$

Such a composite measure would be more meaningful for co-variability or co-diversity, although it might not reflect the diversity as a whole.

4. Decomposability Perspectives

In statistics, the classical analysis of variance (ANOVA) model provides a simple interpretation of decomposability. In its most simple form, for a one-way layout model, if we have G samples from G different populations, all assumed to have a common dispersion, but possibly different means, then the total sum of squares can be decomposed into two components: the pooled within group sum of squares and the between group sum of squres; the former is independent of any possible inter-group differences in means while the later is sensitive to such differences. In the context of poverty and income inequality measurement, there being some qualitative factors in addition to some quantitative ones, it has been argued (Shorrocks 1980, Rao 1982, Sen 1997) that a measure should have the subgroup or additively decomposability property in the sense that a combined group (nonnegative) measure should be decomposable into two additive components representing the within group and between group measures. This is very much in line with the ANOVA-decomposability mentioned above. In the present context, we are primarily confronted with qualitative data models, and hence, we would like to appraise the subgroup or additively decompsability perspectives in a meaningful sense. The Gini-Simpson index being a measure of diversity is nonnegative and it should satisfy a similar decomposability axiom. Indeed this has been studied in detail by Pinheiro et al. (2000, 2005) and Sen (1999,2004), among others. We are confronted here with a multi-dimensional model. Therefore, it might be better to bring the analogy with the MANOVA- decomposability, albeit in our purely qualitartive catergorical data models. In multi-sample multivariate models, the total sum of product matrix can be similarly decomposed into 'within group' and 'between group' components, each being a matrix of the seme order as the total. The rather discouraging aspect of this MANOVA-decomposability is that whenever the sample size is smaller than the dimension of the above matrices, they cease to be positive definite (p.d.) and that creates considerable difficulties for statistical analysis (even under the conventional assumption of underlying multinormal distributions). The situation becomes worse for nonnormal distributions, not to speak of high-dimensional categorical data models.

Motivated by the above remarks, let us first discuss the decomposability prospects for the Gini-Simpson index, and then append a general discussion on the Hamming distance. Consider G independent groups of observations where in the gth group, there are n_g independent observations $X_{g1},\ldots,X_{gn_g}$ with each X_{gi} taking on a categorical response labelled as $1,\ldots,C$ with a probability vector $\pi_g = (\pi_{g1},\ldots,\pi_{gC})'$, for $g = 1,\ldots,G$. As in (2.2), for the gth group, we define the sample counterpart of the Gini-Simpson index by

$$U_n^{(g)} = \sum_{c=1}^{C} n_{gc}(n_g - n_{gc})/\{n_g(n_g-1)\},\ g = 1,\ldots,G, \tag{4.1}$$

where the n_{gc} stand for the number of observations in the gth group belonging to the cth category, for $c = 1,\ldots,C;\ g = 1,\ldots,G$, and $n = n_1 + \ldots, n_G$ the total sample size. Side by side, we define the

sample measure of the Gini-Simpson index for the pair (g, g') of groups as

$$U_n^{(gg')} = \sum_{c=1}^{C} n_{gc}(n_{g'} - n_{g'c})/\{n_g n_{g'}\},\ g \neq g' = 1, \ldots, G. \tag{4.2}$$

Note that by definition

$$\delta_{gg} = EU_n^{(g)} = 1 - \sum_{c=1}^{C} \pi_{gc}^2,\ g = 1, \ldots, G, \tag{4.3}$$

and similarly

$$\delta_{gg'} = EU_n^{(gg')} = 1 - \sum_{c=1}^{C} \pi_{gc}\pi_{g'c},\ g \neq g' = 1, \ldots, G. \tag{4.4}$$

Let us denote the pooled sample Gini-Simpson index by

$$U_n = \sum_{c=1}^{C} n_{.c}(n - n_{.c})/\{n(n-1)\}, \tag{4.5}$$

where $n_{.c} = \sum_{g=1}^{G} n_{gc},\ c = 1, \ldots, C$.

Note that by construction

$$U_n = \sum_{g=1}^{G} \frac{n_g(n_g - 1)}{n(n-1)} U_n^{(g)} + \sum_{g \neq g'} \frac{n_g n_{g'}}{n(n-1)} U_n^{(gg')}, \tag{4.6}$$

which are the some sort of within and between gropus components. However, we consider a more refined decomposition based on the following considerations:

$$\begin{aligned} \sum_{g=1}^{G} n_g(n_g - 1) &= \sum_{g=1} G n_g^2 - n, \\ \sum_{g \neq g'} n_g n_{g'} &= n^2 - \sum_{g=1}^{G} n_g^2, \end{aligned} \tag{4.7}$$

and the right hand sides do not match. Further, note that by the AM-GM inequality,

$$\sum_{c=1}^{C} \pi_{gc}\pi_{g'c} \leq \sum_{c=1}^{C} (\pi_{gc}^2 + \pi_{g'c}^2)/2,\ \forall g \neq g' = 1, \ldots, G, \tag{4.8}$$

where the equality sign holds only when $\pi_g = \pi_{g'}$. Therefore,

$$\delta_{gg'} \geq \frac{1}{2}\{\delta_g + \delta_{g'}\},\ \forall g \neq g' = 1, \ldots, G. \tag{4.9}$$

Further, we write

$$\frac{n_g(n-n_g)}{n(n-1)} = \frac{n_g}{n} - \frac{n_g(n-n_g)}{n(n-1)},\ g = 1,\ldots,G, \tag{4.10}$$

so that we can rewite the pooled sample measure as

$$\begin{aligned} n(n-1)U_n &= (n-1)\sum_{g=1}^{G} n_g U_n^{(g)} + \\ &\quad \sum_{g\neq g'} n_g n_{g'}\{U_n^{(gg')} - \frac{1}{2}\{U_n^{(g)} + U_n^{(g')}\}. \end{aligned} \tag{4.11}$$

Note that the denominator of $U_n^{(gg')}$ and $U_n^{(g)}$ are not the same, and hence, even if $\delta_{gg'} \geq (\delta_g + \delta_{g'})/2)$, their sample counterparts may not satisfy the same inequality. To see this, consider the null hypothesis where the π_g are all the same, so that $EU_n^{(gg')} = \delta = \delta_g, \forall g, g' = 1,\ldots,G$, and hence, under this null hypothesis,

$$E\{U_n^{(gg')} - (U_n^{(g)} + U_n^{(g')})/2\} = 0, \forall g \neq g' = 1,\ldots,G, \tag{4.12}$$

so that each of these $U_n^{(gg')} - (U_n^{(g)} + U_n^{(g')})/2$ will assume both positive and negative values under the null hypothesis. Led by this stronger motivation, we consider the following subgroup or ANOVA decomposability of the Gini-Simpson index:

$$n(n-1)U_n = (n-1)\sum_{g=1}^{G} n_g U_n^{(g)} + \sum_{1\leq g<g'\leq G} n_g n_{g'}\{2U_n^{(gg')} - U_n^{(g)} - U_n^{(g')}\}, \tag{4.13}$$

and designate the two terms on the right hand side as the 'within group' and 'between group' components. This decomposition has been stressed in Sen(1999, 2004) and Pinheiro et al. (2005).

Based on the above decomposition and the fact that the 'between group' component has zero expectation only under the null hypothesis, while the 'within group' component is an unbiased estimator of a positive quantity (a weighted average of the δ_g which are all nonnegative) it seems quite natural to consider the following ANOVA-type test statistic

$$\mathcal{L}_n = \frac{\sum_{1\leq g<g'\leq G} n_g n_{g'}(2U_n^{(gg')} - U_n^{(g)} - U_n^{(g')})}{\sum_{g=1}^{G} n_g U_n^{(g)}}. \tag{4.14}$$

As under alternatives, the numertor has a positive expectation, we are to use a one-sided critical region rejecting the null nypothesis for large positive values of $\mathcal{L}_n$. The crus of the problem is therefore to find a critical value of $\mathcal{L}_n$.

Under the null hypothesis, all the π_g being the same, all the n observations in the pooled sample have a common multinomial law with an unknown parameter vector of rank $C-1$. This makes it naturally appealing to use the permutational distribution of the test statistic generated by all possible

$n!/(n_1!\cdots n_G!)$ partitioning into G groups. This procedure yields a permutationally (conditionally) distribution-free test for the null hypothesis. However, if the n_g are not small, the enumeration of the exact permutation distribution may become prohibitively laborious. Hence, there is a need to provide asymptotic distributional results that would provide good approximations for moderate sample sizes s well. The basic difficulty stems from the fact that the numerator of $\mathcal{L}_n$ is not a linear statistic, and its denominator is not permutationally invariant. Hence, there is a need for showing that the denominator has a nice convergence property while for the numerator, suitable permutational central limit theorems can be used to yield the desired asymptotic results. We shall discuss this problem in a general context in the next section.

5. The Curse of Dimensionality

To motivate the setup, let us refer back to Section 3 and the general models described in (3.1) through (3.4). Typically, in such models, K, the number of positions or sites, is large, often, even much larger than n, the number of observations. The curse of dimensionality problem is particularly perceptible in this 'large dimension, small(er) sample size' context, and the present problem is much more acute due to the pure qualitative categorical nature of the response variables. In convensional multinormal distributional models, granted that variation- covariation features are solely characterized by their dispersion matrices, various attempts have been made to strengthen standard results on Wishart matrices under such high-dimensional, smaller sample size setups. Use of various generalized inverses, pertinent subset of variables selection, canonical analysis, and other related statistical tools all rest on certain linear structures which pertains to such multinormal systems. Even for quantitative but nonnormal systems, such linear structures may not generally hold. The situation is worse in the present context of purely qualitative response variables.

In the same context of subgroup or ANOVA decomposability as treated in the previous section, we may consider G groups and use the same notations, as before, with one extension that here $\mathbf{X}_{gi}$ stands for a K-vector of responses $X_{gk,i},\ k = 1,\ldots,K$, where $X_{gk,i}$ refers to the categorical label for the response of the ith observation in the gth group at the kth position, and it can take on the labels $1,\ldots,C$. Sans any spanning linear subspace in such categorical data models, it seems logical to consider the entire $G \times K$ matrix of Gini-Simpson indexes for each group and each position. These Gini-Simpson indexes are denoted by $U_{n,gk},\ k = 1,\ldots,K, g = 1,\ldots,G$, and their population counterparts (or expectations in this case) are denoted by $\delta_{gk}, k = 1,\ldots,K, g = 1,\ldots,G$. We let

$$\Delta_g = (\delta_{g1},\ldots,\delta_{gK})',\ g = 1,\ldots,G, \tag{5.1}$$

and basically we want to test for their homogeneity, i.e.,

$$H_0 : \Delta_1 = \cdots = \Delta_G = \Delta \text{ unknown}, \tag{5.2}$$

against the composite alternatives that they are not all equal. The CATANOVA (categorical ANOVA) tools (Anderson and Landis 1980, 1982) tools can be easily conceived to test this hypothesis. However, in view of the basic problem that when K is large (compared to n), the number of parameters apprearing in the dispersion matrices of the $U_{n,gk}$ becomes so large that standard multinormality based asymptotics (running parallel to the multinormal MANOVA case) may not be appropriate. Notwithstanding that the $U_{n,gk}, k = 1,\ldots,K$ are neither independent nor identically distributed, it would not be wise to treat them as i.i.d. copies in attempting to resolve the curse of dimensionality problem, although it is often done in bioinformatics! We therefore propose to proceed along the lines of Pinheiro et al. (2000, 2005), and incorporate the Hamming distance based ANOVA tools as a direct extension of the analysis considered in the previous two sections.

Whereas Pinheiro et al. (2000, 2005) and others (mainly to reduce the burden of nuisance parameters) assumed independence of the K positions, and some others used the classical CATANOVA tools, we like to incorporate the subgroup-decomposabality directly prospect in the formulation of our test procedure. We note that for each $k(=1,\ldots,K)$, the $U_{n,gk},\ g = 1,\ldots,G$, and the pooled sample $U_{n,\cdot}$ satisfy the same subgroup-decomposability property studied in detail in the preceding section. Therefore, we could define the K-vectors of individual as well as pooled sample Gini-Simpson indexes and adapt this subgroup-decomposability in a natural way. However, if K is large, working with a multivariate statistic would generally involve a large dispersion matrix full of nuisance parameters ($K(K+1)/2$ in number) which are all functions of the underlying probability elements in (3.1)-(3.2). Therefore, unless n is large compared to $K(K+1)/2$, estimation of these nuisance parameters may entail a loss of sample information, and as a result, multinormal approximations may not be very plausible.

We work with a general weighted version of the Hamming distance introduced the pooled sample utility-oriented weighted Hamming distance $D_n^*(\mathbf{w},\mathbf{u})$ in (3.12). Similarly, for the gth group, we define

$$D_{n,g}^*(\mathbf{w},\mathbf{u}) = \sum_{k=1}^{K} w_k \sum_{c=1}^{C} \frac{u_c n_{gkc}(n_g - n_{gkc})}{n_g(n_g-1)},\ g = 1,\ldots,G, \tag{5.3}$$

where n_{gkc} stands for the number of observations in the gth sample (of size n_g) which show the label c at the position k, for $c = 1,\ldots,C; k = 1,\ldots,K$. Similarly, using the notation in (4.3), as extended to this vector case, we define

$$D_{n,gg'}^*(\mathbf{w},\mathbf{u}) = \sum_{k=1}^{K} w_k \sum_{c=1}^{C} \frac{n_{gkc}(n_{g'} - n_{g'kc})}{n_g n_{g'}}, \tag{5.4}$$

for every pair $g \neq g'(=1,\ldots,G)$. Then, as a direct extension of (4.13), we consider the following subgroup-decomposition of the pooled same utility oriented weighted Hamming distance :

$$n(n-1)D_n^*(\mathbf{w},\mathbf{u}) \quad = \quad (n-1)\sum_{g=1}^{G} n_g D_{n,g}^*(\mathbf{w},\mathbf{u})$$

$$+ \sum_{1\le g<g'\le G} n_g n_{g'} \{2D^*_{n,gg'}(\mathbf{w},\mathbf{u}) - D^*_{n,g}(\mathbf{w},\mathbf{u}) - D^*_{n,g'}(\mathbf{w},\mathbf{u})\}. \tag{5.5}$$

We denote the left hand by SS_T and the two terms on the right hand side by SS_W and SS_B, representing the 'within group' and 'between group' components respectively. Here also under the hypothesis (of homogeneity of all the G groups), SS_B has null expectation, while under any alternative resulting in any deviation from this homogeneity, $E[SS_B]$ is positive. Thus, the subgroup (or ANOVA)-decomposability observed for the simple Gini-Simpson index holds in a much more general context of utility-oriented weighted Hamming distance. Having observed this characteristic decomposability property, it seems natural to consider a test statistic of the form:

$$\mathcal{L}^*_n = SS_B/SS_W, \tag{5.6}$$

and rejecting the null hypypotesis for large positive values of this statistic, i.e., using a one-sided test.

The crux of the problem is therefore to find out suitable critical levels for $\mathcal{L}^*_n$ (based on its distribution under the null hypothesis) that would lead to a prespecified level of significance $\alpha : 0 < \alpha < 1$, at least in a suitable asymptotic setup. We may note in passing that the Δ_g all involve the marginal (multinomial) distributional parameters while the covariance terms of their sample counterparts involve, in addition, second-order joint (multinomial) distributional parameters. Even under the null hypothesis, these large number of parameters are unspecified. Hence, the distribution of $\mathcal{L}^*_n$ (even under the null hypothesis) involves a large number of nuisance parameters. In this respect, the situation may well be comparable to the classical Neyman-Scott problem with many suisance parameters, and this becomes even more acute when K is large compared to the $n_g,\ g = 1,\ldots,G$ (even when G is not large). This is particularly the situation marred by the curse of dimensionality problem. As such, prospects for an exact parametric test or even a conventional likelihood ratio type test (including pseudo-, penalized-, or quasi- likelihood tests) which allows for elimination and estimation of nuisance parameters in their formulation) are rather bleak, even in an asymptotic setup. However, under the null hypothesis of homogeneity, the joint distribution of all the n observations in the pooled sample remains invariant under any permutation of them, and this permutational invariance structure (same as in the case of the classical Gini-Simpson index) renders manageable testing procedures. Therefore, at least for small to moderate values of the sample sizes, $n_1,\ldots,n_G$, this permutation distribution can be evaluated by considering all possible $n!$ (equally likely) permutations of the combined sample observations among themselves) or equivalently all possible equally likely partitioning of the n observations among the G groups of (sizes $n_1,\ldots,n_G$). Thus, conditionally (permutationally) distribution-free tests for the null hypothesis of homogeneity against possible heterogeneity of the utility-oriented weighted Hamming distances for the g populations can be constructed by an appeal to this permutational invariance structure. This task becomes prohibitively laborious as the group sample sizes increase. For this reason, even if we take recourse to permutational invariance, asymptotics are necessary to provide suitable methodological justification to suitable approximations for critical values of the test statistic. We provide a brief outline of some of these developments in the concluding section.

6. General Asymptotics

We may note that the within group statistics $D^*_{n,g}(\mathbf{w},\mathbf{u})$ are all U-statistics (Hoeffding 1948) based on a (symmteric) kernel of degree 2, and similarly, the between group $D^*_{n,gg'}(\mathbf{w},\mathbf{u})$ are two-sample (generalized) U-statistics based on a kernel of degree (1,1). Therefore, SS_B being a linear combination of generalized U-statistics, is itself a generalized U-statistics. This enables us to make use of the general asymptotics for (generalized) U-statistics for our study. Asymptotic normality results hold for U-statistics and related functionals under very general regularity conditions; one of them being that the kernel is stationary of order 0 (Hoeffding 1948). While this approach works out well under alternative hypotheses, under the null hypothesis of homogeneity, we encounter a degenerate case where the kernel is stationary of order 1. We therefore need to appraise the null hypothesis case in more detail. We incorporate some asymptotics for degenerate pseudo-U-statistics in this context.

The basic idea is to use the Hoeffding decomposition of U-statistic in a slightly extended form, and then to force a permutational invariance principle to yield the desired asymptotic normality results. For a kerenel $\phi(\mathbf{X},\mathbf{Y})$ of degree 2, we define

$$\phi_1(\mathbf{x}) = E\{\phi(\mathbf{x},\mathbf{Y})|\mathbf{X}=\mathbf{x}\}, \tag{6.1}$$

where the expectation is under the null hypothesis. Also, we denote by $\theta = E[\phi(\mathbf{X},\mathbf{Y})]$, and let

$$\begin{aligned} \psi_1(\mathbf{X}) &= \phi_1(\mathbf{X}) - \theta, \\ \psi_2(\mathbf{X},\mathbf{Y}) &= \phi(\mathbf{X}.\mathbf{Y}) - \phi_1(\mathbf{X}) - \phi_1(\mathbf{Y}) + \theta. \end{aligned} \tag{6.2}$$

Then, we have

$$\phi(\mathbf{X},\mathbf{Y}) = \theta + \psi_1(\mathbf{X}) + \psi_1(\mathbf{Y}) + \psi_2(\mathbf{X},\mathbf{Y}), \tag{6.3}$$

and ψ_2 is the second-order term in this orthogonal representation. In our notation, SS_B, under the null hypothesis, can be expressed as

$$T_n = \sum_{1\le r<s\le n} \eta_{rs}\psi_2(\mathbf{X}_r,\mathbf{X}_s), \tag{6.4}$$

where the η_{rs} are nonstochastic elements, satisfying the two conditions that

$$\sum_{1\le r<s\le n} \eta_{rs} = 0, \quad \sum_{1\le r<s\le n} \eta_{rs}^2 = \binom{n}{2}; \tag{6.5}$$

and where the $\mathbf{X}_r$ are i.i,d,r.v.'s. First, invoking the permutational invariance structure of the joint distribution of the $\mathbf{X}_r$, moments of T_n under the permutation measure (of order one and two) are computed to estimate the variance of T_n in a natural way. Secondly, a martingale (array) construction

enables us to use a dependent central limit theorem (Dvoretzky 1972) to derive the asymptotic normality of T_n. Finally, the permutation variance estimator is used to studentize the statistic T_n and then use the Slutzky theorem to obtain the asymptotic normality of this studentized form. This can then be used to derive good approximations to the critical level of $\mathcal{L}_n^*$. The details are to be communicated in a subsequent methodologic paper.

References

Anderson, R. J. and Landis, J. R. (1980). CATANOVA for multi-dimensional contingency tables: Nominal-scale response. *Commun. Statist. Theor. Meth. 9*, 1191 - 1206.

Anderson, R.J. and Landis, J. R. (1882). CATANOVA for multi-dimensional contingency tables: Ordinal-scale response. *Commun. Statist. Theor. Meth. 11*, 257-270.

Chakraborty, R. and Rao, C. R. (1991). Measurement of genetic variation in 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.

Chatterjee, S. K. and Sen, P. K. (2000). On stochastic ordering and a general class of poverty indexes. *Calcutta Statist. Assoc. Bull. 50*, 137-156.

Dvoretzky, A. (1972). Central limit theorem for dependent random variables. *Proc. 6th Berkeley Symp. Math. Statist. Probab.* (eds. L. LeCam et al.) Univ. Calif. Press, Los Angeles, CA. Vol. 2, pp. 513-555.

Foster, J. and Shorrocks, A.F. (1988). Inequality and poverty ordering. *Euro. Econ. Rev. 32,* 654-662.

Gini, C. W. (1912). Variabilita e mutabilita. *Studi Economico-Giuridici della R. Universita di Cagliari 3*, 3 - 159.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist. 19*, 293-325.

Kullback, S. (1959). *Information Theory and Statistics*, John Wiley, New York.

Light, R.J. and Margolin, B, H. (1971). An analysis of variance for categorical data. *J. Amer. Statist. Assoc. 66*, 534-544.

Light, R. J. and Margolin, B. H. (1974). An analysis of variance for categorical data II : Small sample comparisons with chi square and other competitors. *J. Amer. Statist. Assoc. 69*, 755-764.

Lorenz, M. O. (1905). Method of measuring concentration of wealth. *J. Amer. Statist. Assoc. 9*, 209 - 219.

Liang, K.-Y. (2002). *Generalized Linear Models, Estimating Functions and Multivariate Extensions* Special Invited Lecture Series in Statistical Science (No. 2), Academia Sinica, Taipei, Taiwan.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Second Ed., Chapman Hall, London, U.K.

Nayak, T. K. (1986). An analysis of diversity using Rao's quadratic entropy. *Sankhyā B 48*. 1-9.

Nayak, T. K. (1986). Sampling distribution in analysis of diversity. *Sankhyā, 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. *J. Bus. Econ. Statist. 7*, 453-460.

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

Pinheiro, H. P., Pinheiro, A. S. and Sen, P. K. (2005). Analysis of genomic sequence using Hamming distance. *J. Statist. Plan. Infer. xxx*, in press.

Pinheiro, A.S., Sen, P. K. and Pinheiro, H. P. (2005). Asymptotic normality of a class of pseudo U-statistics for first-order stationary kernels. (to be submitted).

Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theor. Popln. Biol. 21*, 24 - 43.

Rao, C. R. (1982). Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhyā A 44*, 1-21.

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

Rao, C. R. (1984). Convexity properties of entropy functions and analysis of diversity. In *Inequalities in Statistics and Probability* (ed, Y, L, Tong Inst. Math. Statist., Hayword, Calif. pp. 68-77.

Sen, A. K. (1973). *On Economic Inequality*, Clarendon Press, Oxford, U.K.

Sen, A. K. (1976). Measurement of poverty: An axiomatic approach. *Econometrica 44*, 219 - 232.

Sen, A. K. (1997). *On economic Inequality: Extended Edition*, Clarendon Press, Oxford, U.K.

Sen, P. K. (1986). The Gini coefficient and poverty indexes: Some reconciliation. *J. Amer. Statist. Assoc. 81*, 1050 - 1057.

Sen, P. K. (1999). Ytility-oriented Simpson-type indexes and inequality measures. *Calcutta Statist. Assoc. Bull. 49*, 1-21.

Sen, P. K. (2004). *Excursions in Biostochastics: Biometry to Biostatistics to Bioinformatics*, Invited Lecture Series in Statistical Science, Academia Sinica, Taipei, Taiwan, 205pp.

Shannon, C. E. (1948). *Bell Syst. Tech. J. 27*, 379-423, 623-656.

Shannon, C. E. and Weaver, W. W. (1949). *The Mathematical Theory of Communication*, Univ. Illinois Press, Champaign, Ill.

Shorrocks, A. F. (1980). The class of additively decomposable inequality measures. *Econometrica 48*, 613-625.

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

Wiener, N. (1948). *Cybernetics*, John Wiley, New York.