Journal of Multivariate Analysis 101 (2010) 1559–1573

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Entropy based constrained inference for some HDLSS genomic models: UI tests in a Chen–Stein perspective

Ming-Tien Tsai [a,*], Pranab Kumar Sen [b,c]

[a] Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, ROC

[b] Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599-7420, USA

[c] Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599-7420, USA

ARTICLE INFO

Article history:
Received 31 January 2009
Available online 18 March 2010

AMS 2000 subject classification:
62F30
62H15

Keywords:
Chen–Stein Theorem
Hamming–Shannon pooled measure
Lorenz ordering
Ordered alternatives
Permutation jackknife
Subgroup decomposability
Union–intersection principle

ABSTRACT

For qualitative data models, Gini–Simpson index and Shannon entropy are commonly used for statistical analysis. In the context of high-dimensional low-sample size (HDLSS) categorical models, abundant in genomics and bioinformatics, the Gini–Simpson index, as extended to Hamming distance in a pseudo-marginal setup, facilitates drawing suitable statistical conclusions. Under Lorenz ordering it is shown that Shannon entropy and its multivariate analogues proposed here appear to be more informative than the Gini–Simpson index. The nested subset monotonicity prospect along with subgroup decomposability of some proposed measures are exploited. The usual jackknifing (or bootstrapping) methods may not work out well for HDLSS constrained models. Hence, we consider a permutation method incorporating the union–intersection (UI) principle and Chen–Stein Theorem to formulate suitable statistical hypothesis testing procedures for gene classification. Some applications are included as illustration.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

In genomic studies, as will be illustrated later with the SARSCoV data model [1], one encounters very high-dimensional purely qualitative categorical data models resting on complex multi-dimensional multinomial models without any ordering of categories. Although a genuine discrete multivariate analysis approach may appear to be tempting, there are some basic hurdles. The primary impasse arises due to an enormous number of parameters and a significantly smaller sample size. On top of that, it is very unlikely that the coordinate responses (attributes), in such high dimensions, are stochastically independent, even approximately. On the same count, the homogeneity of the marginal multinomial laws may not be generally tenable. Further, underlying restraints on the parameters (or probability laws) are persistent and often difficult to put in a simple form wherein standard constrained statistical inference [2] procedures can be readily incorporated. In such a *curse of dimensionality under constrained environment*, conventional statistical modeling and analysis tools may be of very little help.

For simple (one-dimensional) multinomial models, the Gini–Simpson (GS) index [3,4] is a useful measure of (qualitative) diversity, and this has been used in biodiversity, genetic variation and in other contexts too. The Shannon [5] entropy is also very appropriate in this setup. To use either of these measures in the multi-dimensional case, the complexities of joint probabilities (parameters) may create genuine hurdles for simpler models or statistical inference. The complexity accelerates

* Corresponding author.
E-mail address: mttsai@stat.sinica.edu.tw (M.-T. Tsai).

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.03.004

fast as the dimension increases, typically the case in *high-dimension low-sample size* (HDLSS) genomic models. Primarily due to this awkward feature, several researchers have used the Hamming distance, a natural extension of the GS index in a pseudo-marginal setup, for HDLSS genomic problems [6–8,1]. In Section 2, it is shown that the Shannon entropy is more informative than the GS index in the sense of the Lorenz ordering. As such, parallel to the GS index, the extension of the Shannon entropy in a Hamming type pseudo-marginal setup is explored.

For HDLSS genomic models, we suspect that the information might not be fully captured in a pseudo-marginal setup. To capture greater information, some new genuine multivariate analogues of Shannon entropy are proposed in Section 3. The nested subset monotonicity prospect along with subgroup decomposability of the proposed new measures are also exploited in the same section. Section 4 is mainly devoted to large sample size models, albeit in a possibly high-dimensional setup. These results may provide useful tools for genes classification in the HDLSS genomic models. Section 5 explores the role of the Chen–Stein Theorem in the HDLSS setup. Section 6 is devoted to statistical inference problems under possibly constrained setups. For the HDLSS genomic models, the usual jackknife variance estimators (of these measures) are not stable [1]. Hence, we consider a modified method to construct more appropriate jackknife procedures. By the property of nested subset monotonicity and subgroup decomposability, it is easy to see that the proposed new Hamming–Shannon pooled measures are more informative than the pseudo-marginal type Hamming–Shannon measures. As such, the testing procedure proposed by Sen et al. [1] for statistical comparison of different groups is further improved. Some new testing procedures for gene classification are proposed in this section. The difficulties of HDLSS asymptotics in this HDLSS genomic context are assessed and suitable permutation procedures are appraised along with. Specifically, the relative performance of UI tests and conventional global alternative tests in HDLSS setups are highlighted, and the Chen–Stein methodology is thoroughly exploited. In the final section, the disease genes for the SARSCoV dataset are identified by the methods proposed in Sections 4–6 with the help of Chen–Stein Theorem, respectively.

2. Preliminary notion and the Lorenz ordering

For motivation, we start with a simple multinomial model relating to C ($\geq$2) (unordered) qualitative categories, labeled as $1, \ldots, C$. Let $\pi_1, \ldots, \pi_C$ denote the respective cell probabilities, and we denote by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_C)^t$. Note that $\boldsymbol{\pi}$ belongs to the $(C-1)$-simplex $S_{C-1} = \{\mathbf{x} \in [0,1]^C : \mathbf{x}^t\mathbf{1} = 1\}$. Thus, we have a constrained parameter space. Unfortunately, S_{C-1} is not very regular in the sense that it does not have rotation or other invariance properties, nor it is a positively homogeneous cone. Usual measures of dispersion are therefore not appropriate in this context.

In the context of measuring biodiversity, Simpson [4], apparently unaware of the work of Gini [3], defined an index $\mathrm{I}_{\mathrm{GS}}(\boldsymbol{\pi})$ that is known in the literature as the GS index. It is defined as

$$\mathrm{I}_{\mathrm{GS}}(\boldsymbol{\pi}) = \sum_{c=1}^{C} \pi_c(1-\pi_c) = 1 - \boldsymbol{\pi}^t\boldsymbol{\pi} \tag{1}$$

so that $0 \leq \mathrm{I}_{\mathrm{GS}}(\boldsymbol{\pi}) \leq (C-1)/C, \forall \boldsymbol{\pi} \in S_{C-1}$, where the lower point is attained when $\boldsymbol{\pi}$ is on one of the C vertices (with no diversity) and the upper bound is attained when $\boldsymbol{\pi} = C^{-1}\mathbf{1}$, i.e., the diversity is a maximum. A standardized GS index is defined as $\mathrm{I}^*_{\mathrm{GS}}(\boldsymbol{\pi}) = C(C-1)^{-1}\mathrm{I}_{\mathrm{GS}}(\boldsymbol{\pi})$ with the natural range (0, 1), thus qualifying it as a measure of diversity.

For the same multinomial law, the Shannon entropy measure is defined as

$$\mathrm{I}_E(\boldsymbol{\pi}) = -\sum_{c=1}^{C} \pi_c \log \pi_c, \quad \boldsymbol{\pi} \in S_{C-1}. \tag{2}$$

It is easy to express $\mathrm{I}_E(\boldsymbol{\pi})$ as $\sum_{r\geq 1} r^{-1} \sum_{c=1}^{C} \pi_c(1-\pi_c)^r = \sum_{r\geq 1} r^{-1} H_r(\boldsymbol{\pi})$, where the $H_r(\boldsymbol{\pi})$ are all nonnegative (over $\boldsymbol{\pi} \in S_{C-1}$), $H_1(\boldsymbol{\pi}) \geq H_2(\boldsymbol{\pi}) \geq \cdots \geq H_r(\boldsymbol{\pi}) \geq \cdots$, and $H_1(\boldsymbol{\pi}) = \mathrm{I}_{\mathrm{GS}}(\boldsymbol{\pi})$. Thus, $\mathrm{I}_E(\boldsymbol{\pi})$ attains the minimum value 0 when $\boldsymbol{\pi}$ is on one of the C vertices (no diversity) and maximum value $\log C$ at the centroid $C^{-1}\mathbf{1}$. Thus, a standardized entropy measure is $\mathrm{I}^*_E(\boldsymbol{\pi}) = \mathrm{I}_E(\boldsymbol{\pi})/\log C = (-1/\log C)\sum_{c=1}^{C} \pi_c \log \pi_c$ with the natural range (0, 1). Let $a = (C-1)/C$. Noting that for $y \in (0,1)$, $-\log(1-ay)$ is convex in y (for every $0 < a \leq 1$), and the fact that $H_1^r(\boldsymbol{\pi}) = (\sum_{c=1}^{C} \pi_c(1-\pi_c))^r \leq H_r(\boldsymbol{\pi}) = \sum_{c=1}^{C} \pi_c(1-\pi_c)^r, \forall r \geq 1$, it is easy to show that

$$\mathrm{I}^*_E(\boldsymbol{\pi}) \geq \log(1 - a\mathrm{I}^*_{\mathrm{GS}}(\boldsymbol{\pi}))/\log(1-a), \quad \forall \boldsymbol{\pi} \in S_{C-1},$$

where the right hand side is a convex function of $\mathrm{I}^*_{\mathrm{GS}}(\boldsymbol{\pi})$, and it assumes the value 0 and 1 according as $\mathrm{I}^*_{\mathrm{GS}}(\boldsymbol{\pi}) = 0$ and 1. Note further that

$$H_r(\boldsymbol{\pi})/H_1(\boldsymbol{\pi}) = \sum_{c=1}^{C} \omega_c(\boldsymbol{\pi})(1-\pi_c)^{r-1}, \tag{3}$$

where $\omega_c(\boldsymbol{\pi}) = \pi_c(1-\pi_c)/H_1(\boldsymbol{\pi})$; $\sum_{c=1}^{C} \omega_c(\boldsymbol{\pi}) = 1$, we obtain that $H_r(\boldsymbol{\pi}) \leq H_1(\boldsymbol{\pi})(1-C^{-1})^{r-1}, \forall r \geq 1$, so that

$$\mathrm{I}^*_E(\boldsymbol{\pi}) \leq \frac{1}{-\log(1-a)} \sum_{r\geq 1} \frac{1}{r} H_1(\boldsymbol{\pi}) a^{r-1}$$

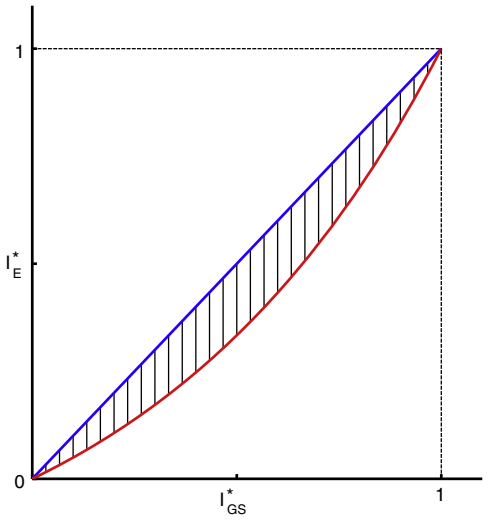

Fig. 1. The Lorenz curve of $I_E^*(\boldsymbol{\pi})$ with respect to $I_{GS}^*(\boldsymbol{\pi})$.

$$= \frac{H_1(\boldsymbol{\pi})}{a}\left\{\frac{1}{-\log(1-a)}\sum_{r\geq 1}\frac{1}{r}a^r\right\}$$
$$= H_1(\boldsymbol{\pi})/a = I_{GS}^*(\boldsymbol{\pi}). \tag{4}$$

This leads us to Fig. 1 depicting a Lorenz ordering of $I_E^*(\boldsymbol{\pi})$ with respect to $I_{GS}^*(\boldsymbol{\pi})$.

Looking at the nested monotonicity property established in Sections 3.1–3.3 [viz., (11)–(31)], we notice that a finite projection will also be a convex function relative to the GS index. However, that curve will lie above the I_E^* curve and below the diagonal line, the difference of the two curves depict the extent of loss of the degree of convexity due to a finite-term representation.

Side by side, we also consider the Kullback and Leibler [9] entropy measure which is more general in the spirit of the log-likelihood ratio measure. Let $\boldsymbol{\pi}^o = C^{-1}\mathbf{1}$, the centroid of S_{C-1}, be the maximum diversity point (as the entropy $I_E^*(\boldsymbol{\pi}^o) = 1$). Then the Kullback–Leibler information $I(\boldsymbol{\pi}, \boldsymbol{\pi}^o)$ is defined as

$$\sum_{c=1}^{C}\pi_c\log(\pi_c/\pi_c^o) = \log C + \sum_{c=1}^{C}\pi_c\log\pi_c = \log C - I_E(\boldsymbol{\pi}) = \log C\{1 - I_E^*(\boldsymbol{\pi})\}, \tag{5}$$

and thus the standardized Kullback–Leibler information is directly related (and complementary) to the standardized entropy measure $I_E^*(\boldsymbol{\pi})$. Thus, the Lorenz ordering also applies to the Kullback–Leibler information.

3. Nested subset monotonicity and subgroup decomposability

One of the basic properties of the GS index is its subgroup decomposability. Basically, if we have G (≥ 2) probability vectors $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_G$ and we define a pooled vector $\boldsymbol{\pi}^* = \sum_{g=1}^{G} w_g\boldsymbol{\pi}_g$, where the w_g are nonnegative weights adding up to 1, then

$$I_{GS}(\boldsymbol{\pi}^*) = \sum_{g=1}^{G} w_g I_{GS}(\boldsymbol{\pi}_g) + \sum_{c=1}^{C}\sum_{g=1}^{G} w_g(\pi_{gc} - \pi_c^*)^2, \tag{6}$$

where the first term on the right hand side of (6) represents the within population component while the second one represent the between population component. Both are nonnegative and resembles the classical ANOVA decomposition. In the same setup, let us examine the entropy measure. Note that

$$I_E(\boldsymbol{\pi}^*) = -\sum_{c=1}^{C}\pi_c^*\log\pi_c^* = \sum_{g=1}^{G} w_g I_E(\boldsymbol{\pi}_g) + \sum_{g=1}^{G} w_g\sum_{c=1}^{C}\pi_{gc}\log(\pi_{gc}/\pi_c^*)$$
$$= \sum_{g=1}^{G} w_g I_E(\boldsymbol{\pi}_g) + \sum_{g=1}^{G} w_g I(\boldsymbol{\pi}_g, \boldsymbol{\pi}^*). \tag{7}$$

Next note that for $x \in (0, 1)$, $x \log x$ is convex, so that by the Jensen inequality, for every $c : 1 \le c \le C$,

$$\sum_{g=1}^{G} w_g \pi_{gc} \log(\pi_{gc}/\pi_c^*) \ge 0, \tag{8}$$

where the equality sign holds only when all the π_{gc} are equal to π_c^*. Thus, the entropy measure also satisfies the subgroup decomposability property. Using the expansion that $-x \log x = -x \log(1 - (1 - x)) = \sum_{r \ge 1} r^{-1} x(1 - x)^r$, it follows from routine computations that the between group component in (7) can be expressed as

$$\sum_{g=1}^{G} w_g \sum_{c=1}^{C} (\pi_{gc} - \pi_c^*)^2 + \sum_{r \ge 2} \sum_{g=1}^{G} w_g [H_r(\boldsymbol{\pi}_g) - H_r(\boldsymbol{\pi}^*)]. \tag{9}$$

Again, provoking convexity, it can be shown that $\sum_{g=1}^{G} w_g H_r(\boldsymbol{\pi}_g) \ge H_r(\boldsymbol{\pi}^*)$, $\forall r \ge 2$, where the equality sign holds when all the $\boldsymbol{\pi}_g$ are the same. This extra nonnegative set of terms explains why the entropy measure may have advantages over the GS index.

Let us now examine the situation for multi-dimensional models. With G ($\ge$2) groups, each having K ($\gg$1) positions, and at each position there is a categorical response with possibly C ($\ge$2) unordered categories, we may denote the $K \times C$ matrix of the probabilities by $\boldsymbol{\Pi}_g = ((\pi_{gkc}))$, for $g = 1, \ldots, G$. We construct a vec form of these matrices, and denote them by $\boldsymbol{\pi}_g, g = 1, \ldots, G$, all being then CK-vectors. As such, we may define the entropy in the same way as before, and consider the subgroup decomposability as in there. Thus, letting $\mathbf{c} = (c_1, \ldots, c_K)^t$, with each c_k assuming the labels $1, \ldots, C$, we have a set $\mathcal{C}_K$ of C^K possible realization of $\mathbf{c}$. Thus, we may let

$$\mathrm{I}_E(\boldsymbol{\pi}_g) = -\sum_{\mathbf{c} \in \mathcal{C}_K} \pi_g(\mathbf{c}) \log \pi_g(\mathbf{c}), \quad \boldsymbol{\pi}_g \in S_{C^K - 1},\ g = 1, \ldots, G. \tag{10}$$

Note that the $\boldsymbol{\pi}_g$ are C^K-vectors, defined on S_{C^K-1}, with centroid $C^{-K}\mathbf{1}$, so that $0 \le \mathrm{I}_E(\boldsymbol{\pi}_g) \le K \log C$, where the upper bound is attained at the centroid. Thus, as an index of diversity, a standardized version would be $\mathrm{I}_E^*(\boldsymbol{\pi}_g) = (K \log C)^{-1} \mathrm{I}_E(\boldsymbol{\pi}_g)$, where range would be [0, 1]. Even so, for large K, there is a need to incorporate underlying structural complexities for dimensional reduction in order to have simpler statistical resolutions. Nested subset monotonicity prospect along with subgroup decomposability are therefore exploited in this context.

3.1. Nested subgroup decomposability

Let $\mathcal{G}$ be the class of 2^G possible subsets of $\{1, 2, \ldots, G\}$, and consider any sequence $\{\mathcal{G}_r, r \ge 0\}$ of nested subsets of $\mathcal{G}$ so that

$$\emptyset = \mathcal{G}_0 \subseteq \mathcal{G}_1 \subseteq \cdots \subseteq \mathcal{G}_r \subseteq \cdots \subseteq \mathcal{G}. \tag{11}$$

Define $\mathrm{I}_E(\boldsymbol{\pi}_g)$ as in (10) and then as in (6), let $\mathrm{I}_E(\mathcal{G}_r)$ be the within $\mathcal{G}_r$ pooled entropy, for $r \ge 0$. Since $\mathrm{I}_E(\mathcal{G}_0) = 0$, then provoking the convexity of $x \log x$ for $\forall x \ge 0$, it follows that for a nested sequence $\{\mathcal{G}_r, r \ge 0\}$

$$0 = \mathrm{I}_E(\mathcal{G}_0) \le \mathrm{I}_E(\mathcal{G}_1) \le \cdots \le \mathrm{I}_E(\mathcal{G}_r) \le \cdots \le \mathrm{I}_E(\mathcal{G}), \tag{12}$$

where

$$\mathrm{I}_E(\mathcal{G}_r) = -\sum_{\mathbf{c} \in \mathcal{C}_K} \pi_{\mathcal{G}_r}(\mathbf{c}) \log \pi_{\mathcal{G}_r}(\mathbf{c}), \quad \boldsymbol{\pi}_{\mathcal{G}_r} \in S^a_{C^K - 1},\ a = 1, \ldots, G, \tag{13}$$

where a is the cardinality of the set $\mathcal{G}_r$. Further, as in (7), we have for every $r \ge 0$,

$$\mathrm{I}_E(\mathcal{G}_r) = \mathrm{I}_{EW}(\mathcal{G}_r) + \mathrm{I}_{EB}(\mathcal{G}_r), \tag{14}$$

representing the 'within' and 'between' group components. Using (6)–(9) and (12), it follows then

$$\mathrm{I}_{EB}(\mathcal{G}_r) = 0 \Longrightarrow \mathrm{I}_{EB}(\mathcal{G}_s) = 0, \quad \forall 0 \le s \le r. \tag{15}$$

3.2. Nested subset decomposability

A more pertinent nested subset decomposability property related to the K positions. Let $\mathcal{K}$ be the set of 2^K subsets of $\{1, 2, \ldots, K\}$, so that

$$\mathcal{K} = \{\mathcal{K}_q = \{k_1, \ldots, k_q\} : 1 \le k_1 < \cdots < k_q \le K; q \le K\}. \tag{16}$$

Partition $\mathbf{c} = (c_1, \ldots, c_K)^t$ as $(\mathbf{c}^t_{\mathcal{K}_q}, \mathbf{c}^t_{\mathcal{K}_q^c})$, where $\mathbf{c}^t_{\mathcal{K}_q} = (c_{k_1}, \ldots, c_{k_q})$ and $\mathbf{c}^t_{\mathcal{K}_q^c}$ is the complementary set, for all $\mathbf{c} \in \mathcal{C}_K$. Let then

$$\mathcal{C}_{\mathcal{K}_q} = \{\mathbf{c}_{\mathcal{K}_q} : \mathcal{K}_q \subseteq \mathcal{K}\}, \quad 0 \le q \le K. \tag{17}$$

For a given $\mathcal{K}_q \subseteq \mathcal{K}$, we let $\boldsymbol{\Pi}^0_{g\mathcal{K}_q} = ((\pi^0_g(\mathbf{c}_{\mathcal{K}_q})))$ and $\boldsymbol{\pi}^0_g(\mathbf{c}_{\mathcal{K}_q})$ be the corresponding vec form, where

$$\pi^0_g(\mathbf{c}_{\mathcal{K}_q}) = \sum_{\mathbf{c}_{\mathcal{K}_q^c} \in \mathcal{C}_{\mathcal{K}_q^c}} \pi_g(\mathbf{c}), \quad \forall \mathbf{c}_{\mathcal{K}_q} \in \mathcal{C}_{\mathcal{K}_q}. \tag{18}$$

Then by definition,

$$\pi^0_g(\mathbf{c}_{\mathcal{K}_q}) \geq \pi_g(\mathbf{c}), \quad \forall \mathbf{c} \in \mathcal{C}_{\mathcal{K}}, \tag{19}$$

and by (18) and (19),

$$\begin{aligned} \mathrm{I}_E(\boldsymbol{\pi}^0_g(\mathbf{c}_{\mathcal{K}_q})) &= -\sum_{\mathbf{c}_{\mathcal{K}_q} \in \mathcal{C}_{\mathcal{K}_q}} \pi^0_g(\mathbf{c}_{\mathcal{K}_q}) \log \pi^0_g(\mathbf{c}_{\mathcal{K}_q}) \\ &= -\sum_{\mathbf{c}_{\mathcal{K}} \in \mathcal{C}_{\mathcal{K}}} \pi_g(\mathbf{c}) \log \pi^0_g(\mathbf{c}_{\mathcal{K}_q}), \end{aligned} \tag{20}$$

and as a result,

$$\mathrm{I}_E(\boldsymbol{\pi}_g) - \mathrm{I}_E(\boldsymbol{\pi}^0_g(\mathbf{c}_{\mathcal{K}_q})) = \sum_{\mathbf{c}_{\mathcal{K}} \in \mathcal{C}_{\mathcal{K}}} \pi_g(\mathbf{c}) \log\{\pi^0_g(\mathbf{c}_{\mathcal{K}_q})/\pi_g(\mathbf{c})\} \geq 0. \tag{21}$$

Thus, for any nested sequence $\{\mathcal{K}_s, s \geq 0\}$ for which $\emptyset = \mathcal{K}_0 \subseteq \mathcal{K}_1 \subseteq \cdots \subseteq \mathcal{K}_s \subseteq \cdots \subseteq \mathcal{K}$, we have

$$\mathrm{I}_E(\boldsymbol{\pi}^0_g(\mathbf{c}_{\mathcal{K}_{s'}})) \leq \mathrm{I}_E(\boldsymbol{\pi}^0_g(\mathbf{c}_{\mathcal{K}_s})), \quad \forall s' \leq s. \tag{22}$$

3.3. Compound nested subset-subgroup decomposability

If we pool the $\boldsymbol{\pi}_g$ over the subset $g \in \mathcal{G}_r (\subseteq \mathcal{G})$, and consider a subset $\mathcal{K}_s (\subseteq \mathcal{K})$, we denote the pooled entropy by $\mathrm{I}_E(\mathcal{G}_r, \mathcal{K}_s)$, for $r, s \geq 0$. Then proceeding as in before we write

$$\mathrm{I}_E(\mathcal{G}_r, \mathcal{K}_s) = \mathrm{I}_{EW}(\mathcal{G}_r, \mathcal{K}_s) + \mathrm{I}_{EB}(\mathcal{G}_r, \mathcal{K}_s), \quad \forall r, s \geq 0, \tag{23}$$

where both the 'within' and 'between' components are nonnegative. Further using (22) and (23), we have for all $r \geq r', s \geq s'$

$$\mathrm{I}_{EW}(\mathcal{G}_{r'}, \mathcal{K}_{s'}) \leq \mathrm{I}_{EW}(\mathcal{G}_r, \mathcal{K}_s), \tag{24}$$

$$\mathrm{I}_{EB}(\mathcal{G}_{r'}, \mathcal{K}_{s'}) \leq \mathrm{I}_{EB}(\mathcal{G}_r, \mathcal{K}_s). \tag{25}$$

Let $\mathrm{I}_E(\mathcal{G}_r, k)$ and $\mathrm{I}_{EB}(\mathcal{G}_r, k)$ be respectively the pooled and between group measures for the kth marginal laws $\boldsymbol{\pi}^0_{gk}, g \in \mathcal{G}_r, k = 1, \ldots, K$. Then by (24) and (25),

$$\begin{aligned} \mathrm{I}_E(\mathcal{G}_r, \mathcal{K}_K) &\geq \max_{1 \leq k \leq K} \mathrm{I}_E(\mathcal{G}_r, k) \\ &\geq \frac{1}{K} \sum_{k=1}^{K} \mathrm{I}_E(\mathcal{G}_r, k) = \mathrm{U}^{(1)}_E(\mathcal{G}_r), \quad \text{say}, \end{aligned} \tag{26}$$

$$\begin{aligned} \mathrm{I}_{EB}(\mathcal{G}_r, \mathcal{K}_K) &\geq \max_{1 \leq k \leq K} \mathrm{I}_{EB}(\mathcal{G}_r, k) \\ &\geq \frac{1}{K} \sum_{k=1}^{K} \mathrm{I}_{EB}(\mathcal{G}_r, k) = \mathrm{U}^{(1)}_{EB}(\mathcal{G}_r), \quad \text{say}. \end{aligned} \tag{27}$$

Similarly, letting $\mathrm{I}_E(\mathcal{G}_r, \mathcal{K}_q)$ and $\mathrm{I}_{EB}(\mathcal{G}_r, \mathcal{K}_q)$ be the pooled and between group entropy measures for the q-dimensional multinomials with $k_1, \ldots, k_q$ positions,

$$\begin{aligned} \mathrm{I}_E(\mathcal{G}_r, \mathcal{K}_K) &\geq \max_{1 \leq k_1 \leq \cdots \leq k_q \leq K} \mathrm{I}_E(\mathcal{G}_r, \mathcal{K}_q) \\ &\geq \binom{K}{q}^{-1} \sum_{1 \leq k_1 \leq \cdots \leq k_q \leq K} \mathrm{I}_E(\mathcal{G}_r, \mathcal{K}_q) \\ &= \mathrm{U}^{(q)}_E(\mathcal{G}_r), \quad \text{say}, \end{aligned} \tag{28}$$

$$\begin{aligned} \mathrm{I}_{EB}(\mathcal{G}_r, \mathcal{K}_K) &\geq \max_{1 \leq k_1 \leq \cdots \leq k_q \leq K} \mathrm{I}_{EB}(\mathcal{G}_r, \mathcal{K}_q) \\ &\geq \binom{K}{q}^{-1} \sum_{1 \leq k_1 \leq \cdots \leq k_q \leq K} \mathrm{I}_{EB}(\mathcal{G}_r, \mathcal{K}_q) \\ &= \mathrm{U}^{(q)}_{EB}(\mathcal{G}_r), \quad \text{say}; \ \forall r \geq 0; q \geq 0. \end{aligned} \tag{29}$$

It follows by similar arguments that

$$\mathrm{U}_E^{(1)}(\mathcal{G}_r) \le \mathrm{U}_E^{(2)}(\mathcal{G}_r) \le \cdots \le \mathrm{U}_E^{(K)}(\mathcal{G}_r), \tag{30}$$

$$\mathrm{U}_{EB}^{(1)}(\mathcal{G}_r) \le \mathrm{U}_{EB}^{(2)}(\mathcal{G}_r) \le \cdots \le \mathrm{U}_{EB}^{(K)}(\mathcal{G}_r), \tag{31}$$

for all $r \ge 0$. These are termed *nested subset monotonicity and subgroup decomposability property*, which provides rationality of a stepdown procedure, useful in multiple hypotheses testing problems.

Let

$$\mathrm{HE}_B^{(1)}(\mathcal{G}_r) = \mathrm{U}_{EB}^{(1)}(\mathcal{G}_r), \quad r \ge 0; \tag{32}$$

$$\mathrm{HE}_B^{(q)}(\mathcal{G}_r) = \mathrm{U}_{EB}^{(q)}(\mathcal{G}_r) - \mathrm{U}_{EB}^{(q-1)}(\mathcal{G}_r), \quad q \ge 2, r \ge 0. \tag{33}$$

Then note that $\mathrm{U}_E^{(1)}(\mathcal{G}_1) = \mathrm{U}_{EB}^{(1)}(\mathcal{G}_1)$ and $\mathrm{HE}_B^{(1)}(\mathcal{G}_1)$ is the average over the K marginal Shannon entropy measures. Similarly, $\mathrm{HE}_B^{(q)}(\mathcal{G}_r)$ is termed the qth order Hamming entropy measure for $\mathcal{G}_r$, $q \ge 1$. In many high-dimensional models, for dimension reduction, often, it is tacitly assumed that

$$\mathrm{HE}_B^{(q)}(\mathcal{G}_r) = 0, \quad q \ge 3, r \ge 0, \tag{34}$$

thus effectively using $\mathrm{HE}_B^{(1)}(\mathcal{G}_r)$ and $\mathrm{HE}_B^{(2)}(\mathcal{G}_r)$ for statistical modeling and analysis, retaining only the marginal and pairwise bivariate distributions.

Finally, we may remark that the same hierarchy of decomposition applies to the GS index in the C^K-contingency table setup, where the first order term yields the classical Hamming distance [10,1]. This interpretation and link of GS index and Hamming distance, albeit quite intuitive, apparently has not been presented in the literature. The simplicity prevailing in the GS index case in somewhat lost in the present case, and it will be illustrated in the next section.

4. Sample measures and gene classification

There is a basic difference between the sample counterparts of GS index and Shannon entropy; the former admits (optimal) unbiased nonparametric estimators while the latter matches the same goal only for large n. Though we are primarily interested in (large) multi-dimensional models, it will be illustrative to discuss this salient point first for the uni-dimensional case.

Let $(n_1, \ldots, n_C)$ have a multinomial $(n, \boldsymbol{\pi})$ law with $\boldsymbol{\pi} \in S_{C-1}$, where $n = \sum_{c=1}^C n_c$. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be n independent and identically distributed (i.i.d.) random vectors where $\mathbf{X}_i = (X_{i1}, \ldots, X_{iC})^t$ with $X_{ij} = 1$ or 0, according as the ith observation belongs to the jth cell or not, for $j = 1, \ldots, C$, $i \ge 1$. Then $\mathbf{X}_i^t \mathbf{1} = 1, \forall\, i \ge 1$ and $\sum_{i=1}^n \mathbf{X}_i = (n_1, \ldots, n_C)^t$. Let $\phi(\mathbf{X}_i, \mathbf{X}_j) = \sum_{c=1}^C 1(X_{ic} \ne X_{jc})$ be a (symmetric) kernel of degree 2. It follows that

$$\theta = \mathcal{E}\phi(\mathbf{X}_i, \mathbf{X}_j) = \sum_{c=1}^C \pi_c(1 - \pi_c) = 1 - \boldsymbol{\pi}^t \boldsymbol{\pi}, \tag{35}$$

the GS index. Thus the U-statistic

$$\begin{aligned} U_n &= \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \phi(\mathbf{X}_i, \mathbf{X}_j) \\ &= \sum_{c=1}^C n_c(n - n_c)/[n(n-1)] \end{aligned} \tag{36}$$

is an optimal (unbiased) nonparametric estimator of θ, the GS index [11]. On the other hand, the Shannon entropy $-\sum_{c=1}^C \pi_c \log \pi_c$ does not have a kernel of finite degree, for which the corresponding U-statistic would be optimal. A related (biased) estimator (von Mises functional)

$$V_n = -\sum_{c=1}^C (n_c/n) \log(n_c/n) \tag{37}$$

is, of course, asymptotically (as $n \to \infty$) optimal, albeit for small n, the optimality property is not generally true. Further, for n not adequately large, estimator of the sampling error of V_n poses additional complications.

Using the notation in (2)–(3), we write for an $m \ge 1$,

$$\mathrm{I}_E(\boldsymbol{\pi}) = \sum_{1 \le r \le m} \frac{1}{r} H_r(\boldsymbol{\pi}) + R_m(\boldsymbol{\pi}), \quad \text{say}, \tag{38}$$

where $H_r(\boldsymbol{\pi}) = \sum_{c=1}^{C} \pi_c(1-\pi_c)^r$ for every $\boldsymbol{\pi} \in S_{C-1}$,

$$0 \le R_m(\boldsymbol{\pi}) \le \frac{C}{m}\left(1-\frac{1}{C}\right)^m. \tag{39}$$

Thus $R_m(\boldsymbol{\pi})$ converges to 0, exponentially in m ($\to \infty$) and uniformly in $\boldsymbol{\pi} \in S_{C-1}$. Consequently, let $m = m_n \sim \log n$, we can make $R_{m_n}(\boldsymbol{\pi}) = o(n^{-2})$ as $n \to \infty$. Further, each $H_r(\boldsymbol{\pi})$ admits a U-statistic, which is unbiased and optimal,

$$U_{nr} = \sum_{c=1}^{C} n_c(n-n_c)^{[r]}/n^{[r+1]}, \quad r \ge 1, \tag{40}$$

where $n^{[m]} = n(n-1)\cdots(n-m+1)$ for $m \le n$, and for $m > n$, $n^{[m]} = 0$ with probability one. However, in this way, we end up with m_n such U-statistics, where $m_n \sim 2C\log n$. Thus when n is large, we may have to deal with a set of U-statistics and that would increase the variability of the estimator of $\mathrm{I}_E(\boldsymbol{\pi})$ based on the first term on the right hand side of (38).

This asymptotic representation along with the fact that all the U_{nr} are reversed martingale with respect to a common sigma subfield imply that not only jackknifing can be used to reduce the bias of the estimator V_n but also the pseudovalues generated by jackknifing provides a (strongly) consistent variance estimator [12]. This feature enables us to make use of standard jackknife procedures to draw statistical conclusions (whenever n is large) based on V_n in (37) and its pseudovalues.

Based on the subgroup/subset decomposability perspective studied in Section 3, we proceed now to the general case of K-dimensional tables relating to G groups, thus putting more emphasis on the $\mathrm{HE}_B^{(1)}(\mathcal{G}_r)$ and $\mathrm{HE}_B^{(2)}(\mathcal{G}_r)$. We may note that the estimates of these measures have possibly unequal variability (i.e., heteroscedasticity) and non-normal distributions. Even asymptotic normality does not preclude heteroscedasticity. As such, even in an asymptotic setup (when $n \to \infty$), we need to estimate the mean squared errors of the estimates of $\mathrm{HE}_B^{(1)}(\mathcal{G}_r)$, $\mathrm{HE}_B^{(2)}(\mathcal{G}_r)$. Following that standard inference tools can be used to draw statistical conclusions. The situation is quite different when the n_g are not all large, and $K \gg n$. This environment is commonly encountered in genomic studies. In the rest of this study we confine ourselves to this $K \gg n$ environment.

Let π_{gkc} denote the cth cell probability for the kth marginal law $\boldsymbol{\pi}_{gk}$ of group g ($1 \le c \le C, 1 \le g \le G, 1 \le k \le K$), and let n_{gkc} be the cell frequencies for the kth marginal table corresponding to the gth group, so that the MLE of π_{gkc} is $\widehat{\pi}_{gkc} = n_{gkc}/n_g$, $1 \le c \le C$, where $n_g = \sum_{c=1}^{C} n_{gkc}$, the same for every k (=1, …, K). Let $\widehat{\boldsymbol{\pi}}_{gk} = (\widehat{\pi}_{gk1}, \ldots, \widehat{\pi}_{gkC})^t$ for $k = 1, \ldots, K$, $1 \le g \le G$. Note that the $\widehat{\boldsymbol{\pi}}_{gk}$ (for a given g) for $k = 1, \ldots, K$ are not necessarily stochastically independent nor they are identically distributed. Suppose that $X_{gi,k}$ takes on the label $1, \ldots, C$ and we denote the (random) label associated with $X_{gi,k}$ by c_{gki}, $1 \le i \le n_g$; $g = 1, \ldots, G$, $k = 1, \ldots, K$. Thus, $\mathbf{X}_{gi}$ corresponds to the vector $\mathbf{c}_{gi} = (c_{g1i}, \ldots, c_{gKi})^t$, and if $\delta_{a,b} = 1$ or 0, according as $a = b$ or not, then

$$\sum_{i=1}^{n_g} \delta_{c,c_{gki}} = n_{gkc}, \quad c = 1, \ldots, C, k = 1, \ldots, K, g = 1, \ldots, G.$$

In the following, we study the jackknife estimator of Shannon entropy $\mathrm{I}_E(\boldsymbol{\pi}_{gk})$. The jackknife estimator is less biased than that of based on the U-statistics, U_{nr}. To proceed it, first consider the plug-in estimator based on the MLE of $\boldsymbol{\pi}_{gk}$

$$\begin{aligned}\mathrm{I}_E(\widehat{\boldsymbol{\pi}}_{gk}) &= -\sum_{c=1}^{C} \frac{n_{gkc}}{n_g}\log\frac{n_{gkc}}{n_g},\\ &= n_g^{-1}\left\{n_g\log n_g - \sum_{c=1}^{C} n_{gkc}\log n_{gkc}\right\}, \quad g = 1, \ldots, G;\ k = 1, \ldots, K,\end{aligned} \tag{41}$$

where the natural assumption that $x\log x = 0$ if $x = 0$ is adopted in this article, and then find out the jackknife estimator of $\mathrm{I}_E(\boldsymbol{\pi}_{gk})$ based on $\mathrm{I}_E(\widehat{\boldsymbol{\pi}}_{gk})$. To proceed it, deleting the ith observation in the gth group and kth position, we then get

$$\mathrm{I}_E^{(-i)}(\widehat{\boldsymbol{\pi}}_{gk}) = (n_g-1)^{-1}\left\{(n_g-1)\log(n_g-1) - \sum_{c=1}^{C}(n_{gkc}-\delta_{c,c_{gki}})\log(n_{gkc}-\delta_{c,c_{gki}})\right\} \tag{42}$$

for $i = 1, \ldots, n_g$. Thus, the corresponding pseudovalues are

$$\begin{aligned}\mathrm{I}_{E,i}(\widehat{\boldsymbol{\pi}}_{gk}) &= n_g\mathrm{I}_E(\widehat{\boldsymbol{\pi}}_{gk}) - (n_g-1)\mathrm{I}_E^{(-i)}(\widehat{\boldsymbol{\pi}}_{gk})\\ &= n_g\log n_g - (n_g-1)\log(n_g-1) - \sum_{c=1}^{C} n_{gkc}\log n_{gkc}\\ &\quad + \sum_{c=1}^{C}(n_{gkc}-\delta_{c,c_{gki}})\log(n_{gkc}-\delta_{c,c_{gki}}), \quad 1 \le i \le n_g.\end{aligned} \tag{43}$$

Therefore, the jackknife estimator of $\mathrm{I}_E(\boldsymbol{\pi}_{gk})$ based on the plug-in estimator $\mathrm{I}_E(\widehat{\boldsymbol{\pi}}_{gk})$ is of the form

$$\begin{aligned}\mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}_{gk}) &= \frac{1}{n_g}\sum_{i=1}^{n_g}\mathrm{I}_{E,i}(\widehat{\boldsymbol{\pi}}_{gk})\\ &= n_g\log n_g-(n_g-1)\log(n_g-1)-\frac{1}{n_g}\sum_{c=1}^{C}n_{gkc}\{(n_g-n_{gkc})\log n_{gkc}\\ &\quad-(n_{gkc}-1)\log(n_{gkc}-1)\},\quad g=1,\ldots,G;\ k=1,\ldots,K.\end{aligned}\tag{44}$$

And hence, its corresponding jackknife variance estimator is

$$\begin{aligned}\widehat{\sigma}^2_{Jgk} &= \frac{1}{n_g-1}\left\{\sum_{i=1}^{C}\frac{n_{gkc}}{n_g}[n_{gkc}\log n_{gkc}-(n_{gkc}-1)\log(n_{gkc}-1)]^2\right.\\ &\quad\left.-\left(\sum_{i=1}^{C}\frac{n_{gkc}}{n_g}[n_{gkc}\log n_{gkc}-(n_{gkc}-1)\log(n_{gkc}-1)]\right)^2\right\},\quad g=1,\ldots,G;\ k=1,\ldots,K.\end{aligned}\tag{45}$$

With $n=\sum_{g=1}^{G}n_g$ and $\sum_{g=1}^{G}n_{gkc}$ replacing n_g and n_{gkc} in (41)–(44), respectively, then we can also find the jackknife estimator $\mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}^*_k)$ of $\mathrm{I}_E(\boldsymbol{\pi}^*_k)$ and its corresponding jackknife variance estimator. For the between group entropy defined in (7) by $\mathrm{I}_{EB}(\boldsymbol{\pi}^*_k)=\sum_{g=1}^{G}w_g\sum_{c=1}^{C}\pi_{gkc}\log(\pi_{gkc}/\pi^*_{kc})$, the plug-in estimator of $\mathrm{I}_{EB}(\boldsymbol{\pi}^*_k)$ will be biased. As such, we consider the jackknife estimator of the between group entropy

$$T_{kB}=\mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}^*_k)-\sum_{g=1}^{G}w_g\mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}_{gk}),\quad k=1,\ldots,K,\tag{46}$$

where $\widehat{\boldsymbol{\pi}}^*_k=\sum_{g=1}^{G}w_g\widehat{\boldsymbol{\pi}}_{gk}$ and $w_g=n_g/n$, $g=1,\ldots,G$.

One of the scientific foci is to classify the K genes into two subsets of disease genes and non-disease genes. For each gene, we set a hypothesis testing problem H_{0k} vs. H_{1k}, $k=1,\ldots,K$. In this marginal formulation, we have a set of K hypotheses corresponding to K genes. The entropy T_{kB} is a real valued statistic and is bounded between 0 and $G\log C$. We use test statistic T_{kB} for testing H_{0k} vs. H_{1k}. Under the alternatives, its distribution tilts towards the upper endpoint, equivalently the right-hand sided p-values. But the distribution of T_{kB}, even under the null hypothesis, may not be the same for each k. These distributions are discrete ones, and hence there are a set of discrete mass points, ties among the T_{kB} cannot be neglected with probability one. Hence the assumption that the p-values have uniform (0, 1) distribution under null hypothesis may not be appropriate. On the other hand, using a level of significance for each marginal hypothesis testing problem, no matter how small it is chosen, when K is indefinitely large, the family wise error rate (FWER) could be large. Thus, controlling the FWER when K is very large may generally entail unduly conservativeness of multiple hypotheses testing schemes. As such, we formulate some alternative procedures.

5. Chen–Stein Theorem in a UI perspective

A multiple hypotheses testing problem, possibly in a constrained inference setup, where the component hypotheses test statistics are unlikely to be independent is confronted here. We let

$$H_0=\bigcap_{k=1}^{K}H_{0k}\quad\text{and}\quad H_1=\bigcup_{k=1}^{K}H_{1k},\tag{47}$$

where H_{0k} and H_{1k} refer to the kth gene and H_{1k} may as well be a restricted alternative hypothesis, for $k=1,\ldots,K$; the test statistics for H_{0k} vs. H_{1k}, $1\le k\le K$ are denoted by T_{nk} (say), $1\le k\le K$, and these are generally not stochastically independent, even under H_0. Further, under H_0, the distribution of T_{nk} may not be the same for all k, while the non-null distributions may be even more heterogeneous. Our contention is to exploit the union–intersection (UI) principle of Roy [13] in this complex setup, and in this respect, we incorporate a version of Chen–Stein [14] Theorem in a more general dependence pattern [15] which suits HDLSS problems better. In the next section, the use of Hamming distance type construction [6] will be further explored.

Arratia et al. [16] provided an updated version of the Chen–Stein Theorem. When K is large, by the results of bivariate extreme statistics [15,17], it can be further simplified as in the following.

Theorem (*Chen–Stein*). *Let $\mathcal{I}$ be an index set with cardinality K. For each $i\in\mathcal{I}$, let ξ_i be an indicator (i.e., zero-one valued) random variable such that $P\{\xi_i=1\}=1-P\{\xi_i=0\}=p_{Ki}$, $1\le i\le K$. For each $i\in\mathcal{I}$, let $\mathcal{J}_i$ be a subset of $\mathcal{I}$ such that ξ_i and*

$\xi_j, j \in \mathcal{J}_i$ *are possibly dependent, and let* $\mathcal{J}_i^c$ *be the complementary subset (of* $\mathcal{J}_i$*) so that* ξ_i *and* $\xi_j, j \in \mathcal{J}_i^c$ *are independent. Let* $W_K = \sum_{k=1}^K \xi_k$ *and* $\lambda_K = \sum_{k=1}^K p_{Kk} = \mathcal{E}(W_K)$*. Further, let*

$$b_{1K} = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}_i} p_{Ki} p_{Kj}, \tag{48}$$

and

$$b_{2K} = \sum_{i \in \mathcal{I}} \sum_{j(\neq i) \in \mathcal{J}_i} \mathcal{E}(\xi_i \xi_j). \tag{49}$$

Finally, let Z_K *be a random variable having Poisson distribution with parameter* λ_K*. Then*

$$\sup_{x \geq 0} |P\{W_K \leq x\} - P\{Z_K \leq x\}| \leq 2(b_{1K} + b_{2K}) \frac{1 - \mathrm{e}^{-\lambda_K}}{\lambda_K}. \tag{50}$$

We may also remark that $(1 - \mathrm{e}^{-\lambda})/\lambda, \lambda \geq 0$ is $\searrow$ in λ with values 1 at $\lambda = 0$ and 0 as $\lambda \to \infty$. Further, $(1 - \mathrm{e}^{-\lambda})/\lambda \leq \min(1, \lambda^{-1}), \lambda \geq 0$. Thus, the Chen–Stein Poisson distributional approximation holds whenever

$$\lim_{K \to \infty} 2(b_{1K} + b_{2K})(1 - \mathrm{e}^{-\lambda_K})/\lambda_K = 0, \tag{51}$$

where λ_K need not be small for large K.

In order to incorporate the Chen–Stein Theorem in the contemplated multiple hypotheses testing problem, let us consider the indicator variables

$$\xi_{Kk} = 1(T_{nk} > c_{Kk,\alpha^*}), \quad 1 \leq k \leq K, \tag{52}$$

where the critical level c_{Kk,α^*} is so chosen that $P_{H_{0k}}\{T_{nk} > c_{Kk,\alpha^*}\} = \alpha^*$, i.e., $p_{Kk} = \mathrm{E}\xi_{Kk} = \alpha^*$ and $\lambda_K = \sum_{k=1}^K p_{Kk} = K\alpha^*$. In the most simple setup, we choose α^* such that

$$P_{H_0}\{Z_K \geq 1\} = 1 - \mathrm{e}^{-\lambda_K} = 1 - \mathrm{e}^{-K\alpha^*} = \alpha, \tag{53}$$

where α $(0 < \alpha < 1)$ is the overall significance level. Thus, if K is large, typically, α^* will be much smaller compared to α (as $\alpha^* = \frac{-1}{K} \log(1 - \alpha) = \frac{1}{K} \sum_{l=1}^{\infty} \alpha^l / l$). This may still result in a less powerful multiple hypotheses testing procedures. Thus, following Sen [15], we conceive of a positive integer, say r_K, such that

$$P_{H_0}\{Z_K \geq r_K\} = 1 - \mathrm{e}^{-\lambda_K} - \cdots - \mathrm{e}^{-\lambda_K} \lambda_K^{(r_K - 1)} / (r_K - 1)! = \alpha. \tag{54}$$

This will result in a larger value of λ_K, i.e., α^*. The interplay of r_K and power (as well as false discovery rate (FDR)) will be illustrated later on.

The multiple hypotheses testing procedure may be formulated as follows. Consider the observed values of the test statistics T_{nk}, $1 \leq k \leq K$ and compute the ξ_{Kk} and W_K as in (52). If

$$W_K \leq r_K - 1, \quad \text{accept } H_0 \text{ (i.e., all the } H_{0k}), \tag{55}$$

and if

$$W_K \geq r_K, \quad \text{reject } H_0, \tag{56}$$

in favor of those H_{1k} for which the ξ_{Kk} are equal to one. Thus, when H_0 is rejected, there will be at least r_K genes for which H_{1k} is accepted, and the number of rejection is random $(\geq r_K)$. When H_0 is accepted, though up to $r_K - 1$ (≥ 0) ξ_{Kk} may be equal to 1, those H_{1k} are not accepted.

The crux of the problem is therefore the choice of the T_{nk}, $1 \leq k \leq K$ and their critical values c_{Kk,α^*}. Excepting when the distribution of T_{nk} under H_0 is specified, the choice of the c_{Kk,α^*} may be analytically harder. We shall explore suitable permutation procedures to prescribe alternative statistical approaches. Along with the case of the Hamming distance type of statistics, this is discussed in the next section.

6. Hamming–Shannon pooled measure and UI test

For SARSCoV sequences, observed several different demographic strata (countries), Sen et al. [1] based on the Hamming distance statistics to study the scientific focus: The statistical comparison of different strata with a view toward coordinating plausible differences to pertinent environmental factors. In this section, we further improve their method for the above mentioned scientific focus. In the meanwhile, we develop new testing procedures for another scientific focus: Gene classification.

6.1. The comparison of different groups

In passing, note that $\mathbf{\Pi}_g = ((\pi_g(\mathbf{c}), \mathbf{c} \in \mathcal{C}_{\mathcal{K}}))$, $1 \le g \le G$, the Hamming distance based on the measures incorporating the $\mathrm{I}_{\mathrm{GS}}(\boldsymbol{\pi}_{gk})$ is defined as

$$\mathrm{I}_{\mathrm{GS}}(\mathbf{\Pi}_g) = K^{-1} \sum_{k=1}^{K} \mathrm{I}_{\mathrm{GS}}(\boldsymbol{\pi}_{gk}), \quad g = 1, \ldots, G, \tag{57}$$

where $\boldsymbol{\pi}_{gk}$ is the kth marginal probability vector corresponding to $\mathbf{\Pi}_g$, $k = 1, \ldots, K$, $g = 1, \ldots, G$. In this formulation, we allow the marginals $\boldsymbol{\pi}_{gk}$ to be possibly different (for different k), and in that way, $\mathbf{\Pi}_g$ to be adaptable to heterogeneity as well as dependence of the marginal measures. The idea of using the Hamming distance is to address the curse of dimensionality problem through a marginal approach. A full K-variate approach when $K \gg n_g$, $1 \le g \le G$ is infeasible. Moreover, the dependence pattern of the K responses is neither well structured nor can be totally ignored. Further, the heterogeneity of the responses from one position to another cannot be ruled out. The Hamming distance provides an average measure that does not ignore dependence or possible heterogeneity. Any single measure for a K-variate response model involves loss of information. But, realizing that the total information is not extractable statistically, this is a natural way of using pseudo-measures that are more sensitive to group divergence. For the consistency of terminologies, we may refer to the Hamming distance as the Hamming–GS measure in this article.

As we have shown that there is a Lorenz ordering of the two standardized measures in Section 2, and this makes it more appealing to consider the Shannon entropy. Similar aggregations of the entropies are considered in this study, we define a Hamming–Shannon measure as

$$\mathrm{I}_E(\mathbf{\Pi}_g) = K^{-1} \sum_{k=1}^{K} \mathrm{I}_E(\boldsymbol{\pi}_{gk}), \quad g = 1, \ldots, G. \tag{58}$$

These are the average of the marginal Shannon [5] entropies, which is equivalent to $\mathrm{U}_E^{(1)}(\mathcal{G}_r)$ defined in expression (26) when $r = 1$. We may refer to $\mathrm{U}_E^{(1)}(\mathcal{G}_r)$, $r \ge 2$ as the Hamming–Shannon pooled measure. Similarly, the Hamming–GS pooled measure is defined with $\mathrm{I}_E(\boldsymbol{\pi}_{gk})$ being replaced by $\mathrm{I}_{\mathrm{GS}}(\boldsymbol{\pi}_{gk})$, $\forall 1 \le g \le G$; $1 \le k \le K$.

In genomic studies, it has been (at least empirically) observed that the Hamming–GS measure may vary according to the HIV positivity status of the sequences with positivity level increasing Hamming–GS measure may also increase, through remaining bounded by $(C-1)/C$ from above [6,7]. Studies made with SARSCoV genome suggest a similar pattern [18]. Parallel to the case of Hamming–GS measure, we frame the null hypothesis

$$\begin{aligned} &H_0 : \mathrm{I}_E(\mathbf{\Pi}_1) = \cdots = \mathrm{I}_E(\mathbf{\Pi}_G) \\ &\text{against} \\ &H_1 : \mathrm{I}_E(\mathbf{\Pi}_1) \le \cdots \le \mathrm{I}_E(\mathbf{\Pi}_G), \end{aligned} \tag{59}$$

with at least one strict inequality sign being true. The parameter space $(S_{C^K-1})^G$ (for $\mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_G$) is denoted by Θ. The parameter spaces, under H_0 as well as H_1, in this formulation, are even more nonregular, complex compared to the one based on the Hamming–GS measure, and are not positively homogeneous cones (subspaces) of Θ. Therefore standard constrained statistical inference (CSI) based on the likelihood approach prospects are bleak. In later section, we illustrate how Roy's UI principle based CSI methodology, developed in [19], can be more conveniently incorporated in this highly nonstandard CSI problem. That approach does not presume that $n_g \gg K$ and the findings remain applicable in genomics as long as n_g is not small, irrespective of $K \gg n_g$, $1 \le g \le G$ or not.

The CSI problem in (59) requires efficient estimators of the Hamming–Shannon measure related to functionals on the simplex S_{C^K-1} (not a single point on it), and hence, first we incorporate the jackknife methodology to obtain the nonparametric estimators and their standard errors. Recall that we do not restrict ourselves to independent positions, nor necessarily to the case of large n_g. In the current SARSCoV data, all the n_g are small, and hence, we intend to emphasize also on the case where $K \gg n_g$ with small n_g, $1 \le g \le G$.

First, we consider the jackknife estimator of Hamming–Shannon measure $\mathrm{I}_E(\mathbf{\Pi}_g)$ based on the plug-in estimators $\mathrm{I}_E(\widehat{\boldsymbol{\pi}}_{gk})$. As we have seen that the usual jackknifing variance estimator may not work out well in HDLSS setups [1], we propose a modified method of jackknifing. Following the results of (41)–(44), the corresponding jackknife estimator of $\mathrm{I}_E(\mathbf{\Pi}_g)$ based on the plug-in estimators $\mathrm{I}_E(\widehat{\boldsymbol{\pi}}_{gk})$ is of the form

$$\begin{aligned} \mathrm{I}_{EJ}(\widehat{\mathbf{\Pi}}_g) &= \frac{1}{\mathrm{K}} \sum_{k=1}^{K} \mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}_{gk}) \\ &= n_g \log n_g - (n_g - 1)\log(n_g - 1) + \frac{1}{Kn_g} \sum_{k=1}^{K} \sum_{c=1}^{C} \{n_{gkc}(n_{gkc} - 1)\log(n_{gkc} - 1) \\ &\quad - (n_g - n_{gkc}) n_{gkc} \log n_{gkc}\}, \quad g = 1, \ldots, G. \end{aligned} \tag{60}$$

Table 1
The stability of modified jackknife variance estimator $\widehat{\Delta}_K$.

	$Q=6$	$Q=12$	$Q=24$	$Q=48$	$Q=96$	$Q=144$	$Q=192$
Taiwan	0.0109	0.0111	0.0111	0.0111	0.0112	0.0112	0.0112
Singapore	0.0169	0.0175	0.0178	0.0179	0.0181	0.0181	0.0181
Hong Kong	0.0560	0.0574	0.0586	0.0589	0.0591	0.0593	0.0592
Beijing	0.0976	0.0976	0.0977	0.0978	0.0975	0.0976	0.0976

Denote the frequency of the cell (c, d) corresponding to the genes (k, q) by $n_{gkq,cd}$ and denote the corresponding indicator variables by $\delta_{c,c_{gki}}$ and $\delta_{d,d_{gqi}}$ respectively. Moreover, for $k \neq q$, $(\delta_{c,c_{gki}}, \delta_{d,d_{gqi}})$ can assume the values (1, 1), (1, 0), (0, 1) and (0, 0) with respective frequencies $n_{gkq,cd}$, $n_{gkc} - n_{gkq,cd}$, $n_{gqd} - n_{gkq,cd}$ and $n_g - n_{gkc} - n_{gqd} + n_{gkq,cd}$. Therefore, for $k \neq q$, the covariance term is

$$\frac{1}{n_g(n_g-1)} \sum_{c=1}^{C}\sum_{d=1}^{C} \left\{ n_{gkq,cd} - \frac{n_{gkc}n_{gqd}}{n_g} \right\} u_{gkc}u_{gqd}.$$

Thus, the jackknife variance estimator of $\mathrm{I}_E(\mathbf{\Pi}_g)$ is

$$\widehat{\sigma}^2_{Jg} = \frac{1}{n_g(n_g-1)K^2} \sum_{k=1}^{K}\sum_{q=1}^{K}\sum_{c=1}^{C}\sum_{d=1}^{C} \left\{ n_{gkq,cd} - \frac{n_{gkc}n_{gqd}}{n_g} \right\} u_{gkc}u_{gqd}, \quad g = 1, \ldots, G, \tag{61}$$

where $u_{gkc} = n_{gkc}\log n_{gkc} - (n_{gkc}-1)\log(n_{gkc}-1)$, $1 \le g \le G$; $1 \le k \le K$; $1 \le c \le C$. In passing we may remark that $x \log x = 0$ for $x = 0$, or 1, and hence, whenever n_{gkc} is 0 or 1, the corresponding term does not show up in the above expression.

Note that while the genes are not necessarily independent, we assume that the G groups are independent. As such, we may construct the UI test for hypothesis testing problem (59) based on the jackknife estimators in (60) and their corresponding jackknife variance estimators in (61). Since the Shannon entropy is more informative than the GS index, the UI test based on Hamming–Shannon measure should perform better than the corresponding one based on Hamming–GS measure for the statistical comparison of different groups.

The jackknife variance estimator in (61) is based on the conventional jackknifing, eliminating one observation at a time, from the gth sample $(1 \le g \le G)$. If the K genes (positions) were independent and identically distributed, one would have employed jackknifing across the K positions, resulting in a more precise variance estimation. However, such an independent and identically distributed clause may not be tenable in the genomic context. As such, we do not consider the details of such double jackknifing. However, in the independent case, the order of the variance of $\mathrm{I}_{EJ}(\widehat{\mathbf{\Pi}}_g)$ would have been K^{-1}, so that $\widehat{\Delta}_{Kg} = K\widehat{\sigma}^2_{Jg}$ would behave steadily for large K. We perform the following Monte Carlo simulation study to appraise the stability of the variance estimator for random sets (Q out of K genes), for $Q = 6, 12, 24, 48, 96, 192$. The jackknife variance estimator (in (61)) for each subset, based on 10,000 replications, for each group, is presented in Table 1.

The jackknife variance estimator for each group seems to be fairly stable, thus suggesting the adaptability of double jackknifing across the genes. However, we may not need this additional stability assumption, particularly when we incorporate the Chen–Stein methodology.

We also intend to incorporate the underlying structural complexities for dimensional reduction. Let $\mathcal{G}^*_g = \{1, \ldots, g\}$, $g = 1, \ldots, G$. By the property of nested subset monotonicity and subgroup decomposability discussed in Section 3, it is easy to note that the Hamming–Shannon pooled measure $\mathrm{I}_E(\mathbf{\Pi}_{\mathcal{G}^*_g})$, which is a special case of $\mathrm{U}^{(1)}_E(\mathcal{G}_r)$, $r \ge 2$ defined in (26), is more informative than the Hamming–Shannon measure $\mathrm{I}_E(\mathbf{\Pi}_g)$. As such, we may reformulate the problem (59) in terms of $\mathrm{I}_E(\mathbf{\Pi}_{\mathcal{G}^*_g})$, $g = 1, \ldots, G$. For these Hamming–Shannon pooled measures $\mathrm{I}_E(\mathbf{\Pi}_{\mathcal{G}^*_g})$, $g = 1, \ldots, G$, it automatically forms the simple ordering relationship: $\mathrm{I}_E(\mathbf{\Pi}_{\mathcal{G}^*_1}) \le \cdots \le \mathrm{I}_E(\mathbf{\Pi}_{\mathcal{G}^*_G})$. Thus, we have no more interest in the problem of testing against the global alternative. Instead, we may consider the hypothesis problem of testing

$$\begin{aligned} &H^S_0 : \mathrm{I}_E(\mathbf{\Pi}_{\mathcal{G}^*_1}) = \cdots = \mathrm{I}_E(\mathbf{\Pi}_{\mathcal{G}^*_G}) \\ &\text{against} \\ &H^S_1 : \mathrm{I}_E(\mathbf{\Pi}_{\mathcal{G}^*_1}) \le \cdots \le \mathrm{I}_E(\mathbf{\Pi}_{\mathcal{G}^*_G}), \end{aligned} \tag{62}$$

with at least one strict inequality sign being true. To find out the test statistic for it, let

$$\begin{aligned} T_{g:g-1} &= \mathrm{I}_{EJ}(\widehat{\mathbf{\Pi}}_{\mathcal{G}^*_g}) - \mathrm{I}_{EJ}(\widehat{\mathbf{\Pi}}_{\mathcal{G}^*_{g-1}}) \\ &= \frac{1}{\mathrm{K}}\sum_{k=1}^{K}[\mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}_{\mathcal{G}^*_{gk}}) - \mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}_{\mathcal{G}^*_{(g-1)k}})], \quad g = 2, \ldots, G, \end{aligned} \tag{63}$$

where $\mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}_{\mathcal{G}^*_{gk}})$ is defined the same as in (44) with the pooled samples of groups $1, \ldots, g$ instead. Note that the right hand side of expression (63) is the jackknife estimator of $\mathrm{I}_{EJ}(\mathbf{\Pi}_{\mathcal{G}^*_g}) - \mathrm{I}_{EJ}(\mathbf{\Pi}_{\mathcal{G}^*_{g-1}})$. However, due to the different sample sizes of pooled

groups $\mathcal{G}_g^*$ and $\mathcal{G}_{g-1}^*$, the corresponding jackknife covariance estimators will be more complex to write in a compact form. As the null hypothesis relates to the homogeneity of the G groups, we take advantage of the resulting permutation invariance structure. Therefore, we proceed with this extended permutation-jackknife methodology. Let $\mathbf{Y}_1 = (T_{2:1}, \ldots, T_{G:G-1})^t$ and $\mathbf{Y}_i$ be the $(i-1)$th corresponding permutation of $\mathbf{Y}_1$, $i = 2, \ldots, N_1$. Consider the corresponding covariance matrix $\mathbf{S}_{N_1} = (N_1-1)^{-1}\sum_{i=1}^{N_1}(\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^t$, where $\bar{\mathbf{Y}} = {N_1}^{-1}\sum_{i=1}^{N_1}\mathbf{Y}_i$. Let $\mathcal{G} = \{1, \ldots, G-1\}$, and for every a: $\emptyset \subseteq a \subseteq \mathcal{G}$, let a' be its complement and $|a|$ its cardinality, there being 2^{G-1} subsets for which $0 \le |a| \le G-1$. For each a: $\emptyset \subseteq a \subseteq \mathcal{G}$, we partition (following possible rearrangement) $\mathbf{Y}_i$ and $\mathbf{S}_{N_1}$ as

$$\mathbf{Y}_i = \begin{pmatrix} \mathbf{Y}_{ia} \\ \mathbf{Y}_{ia'} \end{pmatrix} \quad \text{and} \quad \mathbf{S}_{N_1} = \begin{pmatrix} \mathbf{S}_{N_1 aa} & \mathbf{S}_{N_1 aa'} \\ \mathbf{S}_{N_1 a'a} & \mathbf{S}_{N_1 a'a'} \end{pmatrix}, \tag{64}$$

and write

$$\mathbf{Y}_{ia:a'} = \mathbf{Y}_{ia} - \mathbf{S}_{N_1 aa'}\mathbf{S}_{N_1 a'a'}^{-1}\mathbf{Y}_{ia'}, \tag{65}$$

$$\mathbf{S}_{N_1 aa:a'} = \mathbf{S}_{N_1 aa} - \mathbf{S}_{N_1 aa'}\mathbf{S}_{N_1 a'a'}^{-1}\mathbf{S}_{N_1 a'a}. \tag{66}$$

Proceeding as in [19], we obtain that the UI test for testing H_0 vs. H_1 in (62) based on the test statistic

$$Q_i^2 = \sum_{\emptyset \subseteq a \subseteq \mathcal{G}} \{\mathbf{Y}_{ia:a'}^t \mathbf{S}_{N_1 aa:a'}^{-1}\mathbf{Y}_{ia:a'}\} 1\{\mathbf{Y}_{ia:a'} > \mathbf{0}, \mathbf{S}_{N_1 a'a'}^{-1}\mathbf{Y}_{ia'} \le \mathbf{0}\} \tag{67}$$

$i = 1, \ldots, N_1$, where $1(\cdot)$ denotes the indicator function. Based on the results of Q_i^2, $i = 1, \ldots, N_1$, we can then find out the corresponding permutation p-value for the CSI problem (62).

Similarly, let $\mathcal{G}_g^T = \{g, G\}$, $g = 1, \ldots, G-1$. We may be interested in the hypothesis problem of testing

$$\begin{aligned} &H_0^T : \mathrm{I}_E(\mathbf{\Pi}_g) = \mathrm{I}_E(\mathbf{\Pi}_{\mathcal{G}_g^T}) \\ &\text{against} \\ &H_1^T : \mathrm{I}_E(\mathbf{\Pi}_g) \le \mathrm{I}_E(\mathbf{\Pi}_{\mathcal{G}_g^T}), \quad g = 1, \ldots, G-1, \end{aligned} \tag{68}$$

with at least one strict inequality sign being true. Let $\Gamma_E^S = \{(\mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_G)^t | \mathrm{I}_E(\mathbf{\Pi}_{\mathcal{G}_1^*}) \le \cdots \le \mathrm{I}_E(\mathbf{\Pi}_{\mathcal{G}_G^*})\}$ and $\Gamma_E^T = \{(\mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_G)^t | \mathrm{I}_E(\mathbf{\Pi}_g) \le \mathrm{I}_E(\mathbf{\Pi}_{\mathcal{G}_g^T}), g = 1, \ldots, G-1\}$. We may still refer to Γ_E^S and Γ_E^T as the simple ordering set and the simple tree ordering set, respectively. However, we may also note that Γ_E^S is no longer a proper subset of Γ_E^T. To consider the test statistic for problem (68), let

$$\begin{aligned} T_{G:g} &= \mathrm{I}_{EJ}(\widehat{\mathbf{\Pi}}_{\mathcal{G}_g^T}) - \mathrm{I}_{EJ}(\widehat{\mathbf{\Pi}}_g) \\ &= \frac{1}{K}\sum_{k=1}^{K}[\mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}_{\mathcal{G}_{gk}^T}) - \mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}_{gk})], \quad g = 1, \ldots, G-1. \end{aligned} \tag{69}$$

Also let $\mathbf{Z}_1 = (T_{G:1}, \ldots, T_{G:G-1})^t$, and perform the procedures mentioned above with $\mathbf{Z}_1$ replacing $\mathbf{Y}_1$. Then, we can find the permutation p-value for hypothesis testing problem (68).

6.2. Gene classification (revisited)

The global test based on T_{kB} defined in (46) for gene classification is introduced in Section 4. Here, we would like to incorporate the structural information into the construction of new testing procedures for gene classification. We refer to $\mathrm{I}_E(\pi_{\mathcal{G}_{gk}^T})$ (or $\mathrm{I}_E(\pi_{\mathcal{G}_{gk}^*})$) as the Shannon pooled measure for each k, $1 \le k \le K$. And similarly, refer to $\mathrm{I}_{GS}(\pi_{\mathcal{G}_{gk}^T})$ (or $\mathrm{I}_{GS}(\pi_{\mathcal{G}_{gk}^*})$) as the GS pooled measure. To proceed it, we explain the procedure based on $\mathrm{I}_E(\pi_{\mathcal{G}_g^T})$ in the following, and the others can be similarly performed. Note that $\mathrm{I}_{EJ}(\pi_{\mathcal{G}_{gk}^T})$ is more informative than $\mathrm{I}_{EJ}(\pi_{gk})$. For each k (=1, …, K), we consider the statistics

$$T_{G:g,k} = \mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}_{\mathcal{G}_{gk}^T}) - \mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}_{gk}), \quad g = 1, \ldots, G-1, \tag{70}$$

and let $\mathbf{Z}_{k1} = (T_{G:1,k}, \ldots, T_{G:G-1,k})^t$. Then we proceed the UI-test statistic mentioned above specifically for the kth gene, and denote it by L_{nk}, for $k = 1, \ldots, K$. As such, we need to find a way to compute the right hand side tail probabilities of L_{nk} under H_{0k} for each k. For this, we consider all possible equally likely permutations of the observations for each k, each having the same conditional probability $\frac{1}{N}$, where $N = n!/\prod_{g=1}^{G} n_g!$. This enable us to find a value, say c_{Kk,α^*} such that the proportion of permuted values of L_{nk} above c_{Kk,α^*} is just less than $\alpha^* = K^{-1}\alpha$, namely $P_{H_{0k}}\{L_{nk} > c_{Kk,\alpha^*}\} = \alpha^*$, $k = 1, \ldots, K$. In practice, to overcome the difficulty that N is too large we may choose N_1, which is sufficiently large but $N_1 \ll N$, instead. Next, generate a set of $(N_1 - 1)$ permutations and let $\mathbf{Z}_{ki}$be the $(i-1)$th corresponding permutation of $\mathbf{Z}_{k1}$, $i = 2, \ldots, N_1$. For

this construction, we use the permutation distribution generated by the set of all possible permutations among themselves. For each k (=1, ..., K), we denote the observed ordered values by $l_{k[1]} \le l_{k[2]} \le \cdots \le l_{k[N_1(1-\alpha^*)]} \le \cdots \le l_{k[N_1]}$. Also let $c^0_{k,\alpha^*} = l_{k[N_1(1-\alpha^*)]}$, $k = 1, \ldots, K$. Then, with L_{nk} and its corresponding critical value c^0_{Kk,α^*}, the multiple hypotheses testing problem H_{0k} vs. H_{1K}, $k = 1, \ldots, K$ considered can then be performed. If the observed value L_{nk} is greater than c^0_{Kk,α^*}, we then reject the null hypothesis H_{0K} and classify the kth gene as the disease gene, $k = 1 \le k \le K$. For gene classification, some simulation studies between the UI test and the global test proposed in Section 4 are presented in the last section.

In passing, we may also note that the choice of α^* is crucial. Generally, we may take $\alpha^* = \alpha/K$ for a given level of significance α. This is referred to as the one obtained by Bonferroni type method in the literature. As shown in Section 5, the α^* obtained by Bonferroni type method is less than that of by the more structured Chen–Stein method with $r_K = 1$. We also expect that the procedure based on α^* obtained by the Chen–Stein method with $r_K = 2$ is more powerful than that of with $r_K = 1$. Some numerical studies for SARSCoV RNA genomic dataset will be presented in the last section.

7. Data analysis

After multiple sequence alignment of SARS genome sequence, Sen et al. [1] found many of the deposited SARS genomes were incomplete or large deletion. To set all the SARS genome sequences in equal position for comparison, they ended up using 25 sequences, 12 (=n_1) from Taiwan, 4 (=n_2) from Singapore, 3 (=n_3) from Hong Kong and 6 (=n_4) from Beijing for data analysis, with four groups: $g = 1, 2, 3, 4$ which represent four different geographic regions "Taiwan, Singapore, Hong Kong and Beijing" respectively. Moreover, we have $C = 4$ and $K = 192$. Let $\boldsymbol{\theta}_{\mathrm{GS}} = (\mathrm{I}_{\mathrm{GS}}(\boldsymbol{\Pi}_1), \ldots, \mathrm{I}_{\mathrm{GS}}(\boldsymbol{\Pi}_G))^t$, and denoted by $\Gamma^S_{\mathrm{GS}} = \{\boldsymbol{\theta}_{\mathrm{GS}} | \mathrm{I}_{\mathrm{GS}}(\boldsymbol{\Pi}_1) \le \cdots \le \mathrm{I}_{\mathrm{GS}}(\boldsymbol{\Pi}_4)\}$, which is referred to as the simple ordering set in the literature. Also denoted by $\Gamma^T_{\mathrm{GS}} = \{\boldsymbol{\theta}_{\mathrm{GS}} | \mathrm{I}_{\mathrm{GS}}(\boldsymbol{\Pi}_g) \le \mathrm{I}_{\mathrm{GS}}(\boldsymbol{\Pi}_4), g = 1, 2, 3\}$ which is called the simple tree ordering set in the literature, and the global alternative set $\Gamma^G_{\mathrm{GS}} = \{\boldsymbol{\theta}_{\mathrm{GS}} | \mathrm{I}_{\mathrm{GS}}(\boldsymbol{\Pi}_g), g = 1, 2, 3, 4$, are all not equal$\}$. Note that the set Γ^S_{GS} is a proper subset of the set Γ^T_{GS}. Sen et al. [1] concluded that the most likely alternative is the one of the set Γ^S_{GS} among all 24 simple ordered alternatives, 4 different simple tree ordered alternatives and the global alternative. Numerical value of $\mathrm{I}_E(\widehat{\boldsymbol{\Pi}}_g)$ is much larger than that of the corresponding $\mathrm{I}_{\mathrm{GS}}(\widehat{\boldsymbol{\Pi}}_g)$, $1 \le g \le G$, the details are omitted here. Although the Hamming–Shannon measure $\mathrm{I}_E(\boldsymbol{\Pi}_g)$ is more informative than the Hamming–GS measure $\mathrm{I}_{\mathrm{GS}}(\boldsymbol{\Pi}_g)$, the same conclusion can be made based on the Hamming–Shannon measure $\mathrm{I}_E(\boldsymbol{\Pi}_g)$ for this SARSCoV dataset.

In passing, we may note that the new Hamming–Shannon pooled measures both $\mathrm{I}_E(\boldsymbol{\Pi}_{\vartheta^*_g})$ and $\mathrm{I}_E(\boldsymbol{\Pi}_{\vartheta^T_g})$ are more informative than the Hamming–Shannon measure $\mathrm{I}_E(\boldsymbol{\Pi}_g)$. For the new Hamming–Shannon pooled measure $\mathrm{I}_E(\boldsymbol{\Pi}_{\vartheta^*_g})$ developed in this paper, the most basic hypothesis testing problem is the one of testing H^S_0 against H^S_1 in (62). We appeal to the same permutation subset of 99,999 without replacement data sets. For each permutation data set, we construct the new table of n_{gkc}, and then calculate the corresponding Q^2_i. Arrange these 100,000 numerical values of Q^2_i, we find that the one (Q^2_1) calculated from the original data set is the 8th largest one among these 100,000 values. The obtained permutation p-value is 0.00008. Thus, for the SARSCoV dataset we may conclude that the new Hamming–Shannon pooled measure of four different geographic regions rejects the null hypothesis H^S_0. Next, consider the problem of testing H^T_0 against H^T_1 in (68). Similar arguments as in the simple ordered alternative case, arranging those 100,000 numerical values of Q^2_i, we find that the one (Q^2_1) calculated from the original data set is the largest one among these 100,000 values. Thus, the obtained permutation p-value is 0.00001, which is less than that of for the problem of testing H^T_0 against H^T_1. And hence, we may conclude that the dataset supports this simple tree ordered alternative $\Gamma^T_E = \{(\boldsymbol{\Pi}_1, \ldots, \boldsymbol{\Pi}_4)^t | \mathrm{I}_E(\boldsymbol{\Pi}_g) \le \mathrm{I}_E(\boldsymbol{\Pi}_{\vartheta^T_g}), g = 1, 2, 3\}$.

Furthermore, for this SARSCoV dataset, we would also like to identify the disease genes by the global test and the UI test, respectively. As mentioned in the previous section, the choice of α^* is important for a given level of significance α. Hence, we study the performance of the global tests and UI tests for different kinds of α^* obtained via the Bonferroni type method and the Chen–Stein method, respectively. By applying the Bonferroni type method, we have $\alpha^* = \alpha/K$ for a given level of significance α. First, by applying the global test statistics T_{kB}, $k = 1, \ldots, K$ in (46), we calculate the value of T_{kB} obtained from real data set and the other 99,999 additional permutation values for each k by adopting the Shannon pooled measure and the GS pooled measure (i.e., let $\mathrm{I}_{\mathrm{GS}}(\widehat{\boldsymbol{\pi}}_{gk}) = \left\{1 - \sum_{c=1}^{C} \frac{n_{gkc}(n_{gkc}-1)}{n_g(n_g-1)}\right\}$, $k = 1, \ldots, K$; $g = 1, \ldots, G$, and with $\mathrm{I}_{\mathrm{GS}}(\widehat{\boldsymbol{\pi}}_{gk})$ and $\mathrm{I}_{\mathrm{GS}}(\widehat{\boldsymbol{\pi}}^*_k)$ replacing $\mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}_{gk})$ and $\mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}^*_k)$ in T_{kB}), respectively. The values of test statistics based on Shannon pooled measure are larger than those of corresponding test statistics based on GS pooled measure. Note that for $\alpha = 0.05$, we have $\alpha^* = 0.000260$ and for $\alpha = 0.1$, $\alpha^* = 0.000521$. Table 2 relates to corresponding gene selection outcomes. Basically, for the global tests as well as UI tests, the GS and Shannon pooled measures yield similar results.

Using the Chen–Stein Theorem: (i) $r_K = 1$, i.e., $\alpha^* = \frac{-1}{K}\log(1-\alpha)$ and (ii) $r_K = 2$, i.e., $1 - \alpha = (1 + K\alpha^*)\mathrm{e}^{-k\alpha^*}$, with $\alpha = 0.05$. For (i), the results are the same as those in Table 2. For (ii), for the global alternative as well as restricted alternative case, the Shannon pooled measure based procedure performs better than the GS pooled method. The results are presented in Table 3. The same differential picture holds for $\alpha = 0.10$.

For this SARSCoV dataset, $K = 192$ is only moderate large. With the help of Chen–Stein Theorem, the UI test based on GS pooled measure identifies the disease genes, and the UI test based on Shannon pooled measure makes no more further improvement. However, for the other datasets, with larger K, the UI test based on Shannon pooled measure may perform better than the GS pooled measure.

Table 2
The relative performance of conventional global tests and UI tests.

Tests	Detected genes when $\alpha = 0.05$	Detected genes when $\alpha = 0.1$
Global test: GS	None	$k = 130, 183$
Global test: Shannon	None	$k = 130, 183$
UI test: GS	$k = 130, 151, 183, 185$	$k = 130, 151, 183, 185$
UI test: Shannon	$k = 130, 151, 183, 185$	$k = 130, 151, 183, 185$

Table 3
Exploitation of the Chen–Stein Theorem.

Tests	Detected genes when $\alpha = 0.05$ with $r_K = 2$
Global test: GS	$k = 130, 178, 183$
Global test: Shannon	$k = 130, 151, 178, 183, 185$
UI test: GS	$k = 65, 83, 87, 120, 121, 130, 147, 151, 178, 183, 185$
UI test: Shannon	$k = 65, 83, 87, 120, 121, 130, 147, 151, 178, 183, 185$

Acknowledgments

The authors thank the two reviewers for their helpful comments. The authors also thank Chien-Chih Huang for carrying out the numerical works reported in this paper. This work was partly supported by the Grants from National Science Council of Republic of China under Contract No. NSC 96-2628-M-001-024 and the Cary C. Boshamer Foundation at the University of North Carolina.

Appendix

We derive the Eqs. (45) and (61) in the followings. Let then

$$\begin{aligned} Z_{gk,i} &= \mathrm{I}_{E,i}(\widehat{\boldsymbol{\pi}}_{gk}) - \mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}_{gk}) \\ &= \sum_{c=1}^{C}\left\{(n_{gkc} - \delta_{c,c_{gki}})\log(n_{gkc} - \delta_{c,c_{gki}}) - \frac{1}{n_g}[n_{gkc}(n_{gkc}-1)\log(n_{gkc}-1) - (n_g - n_{gkc})n_{gkc}\log n_{gkc}]\right\} \end{aligned}$$

for $i = 1, \ldots, n_g$. Then the corresponding jackknife variance estimator of $\mathrm{I}_E(\boldsymbol{\pi}_{gk})$ is

$$\begin{aligned} \widehat{\sigma}^2_{Jgk} &= \frac{1}{n_g(n_g-1)}\sum_{i=1}^{n_g}[\mathrm{I}_{E,i}(\widehat{\boldsymbol{\pi}}_{gk}) - \mathrm{I}_{EJ}(\widehat{\boldsymbol{\pi}}_{gk})]^2, \\ &= \frac{1}{n_g(n_g-1)}\sum_{i=1}^{n_g}\left\{\sum_{c=1}^{C} Z^2_{gk,ic} + \sum_{1\le c\ne d\le C} Z_{gk,ic}Z_{gk,id}\right\}, \quad g = 1, \ldots, G;\ k = 1, \ldots, K \end{aligned} \tag{A.1}$$

where $Z_{gk,ic}$ is the cth term in $Z_{gk,i}$. Recall that n_{gkc} of the $\delta_{c,c_{gki}}$ are equal to 1 while the remaining $n_g - n_{gkc}$ are zeros. Also simultaneously, $\delta_{c,c_{gki}}$ and $\delta_{d,d_{gki}}$ (for $c \ne d$) cannot be equal to 1; their possible values are $(1, 0)$, $(0, 1)$ and $(0, 0)$ with respective frequencies n_{gkc}, n_{gkd} and $n_g - n_{gkc} - n_{gkd}$. Thus

$$\begin{aligned} \sum_{i=1}^{n_g}\sum_{c=1}^{C} Z^2_{gk,ic} &= \sum_{c=1}^{C}\left[n_{gkc}\left\{(n_{gkc}-1)\log(n_{gkc}-1) - \frac{1}{n_g}u_{gkc}\right\}^2 + (n_g - n_{gkc})\left\{n_{gkc}\log n_{gkc} - \frac{1}{n_g}u_{gkc}\right\}^2\right] \\ &= \sum_{c=1}^{C}\frac{n_{gkc}(n_g - n_{gkc})}{n_g}u^2_{gkc}, \end{aligned}$$

where $u_{gkc} = n_{gkc}\log n_{gkc} - (n_{gkc}-1)\log(n_{gkc}-1)$, $1 \le g \le G$; $1 \le k \le K$; $1 \le c \le C$. Similarly,

$$\begin{aligned} \sum_{1\le c\ne d\le C} Z_{gk,ic}Z_{gk,id} &= -\sum_{c\ne d}\frac{n_{gkc}n_{gkd}}{n_g}u_{gkc}u_{gkd} \\ &= -\frac{1}{n_g}\left(\sum_{c=1}^{C} n_{gkc}u_{gkc}\right)^2 + \frac{1}{n_g}\sum_{c=1}^{C} n^2_{gkc}u^2_{gkc}. \end{aligned}$$

Therefore we have

$$\widehat{\sigma}^2_{Jgk} = \frac{1}{n_g - 1}\left\{\sum_{c=1}^{C}\frac{n_{gkc}}{n_g}u^2_{gkc} - \left(\sum_{c=1}^{C}\frac{n_{gkc}}{n_g}u_{gkc}\right)^2\right\}$$

$$= \frac{1}{n_g - 1}\left\{\sum_{i=1}^{C}\frac{n_{gkc}}{n_g}[n_{gkc}\log n_{gkc} - (n_{gkc}-1)\log(n_{gkc}-1)]^2 - \left(\sum_{i=1}^{C}\frac{n_{gkc}}{n_g}[n_{gkc}\log n_{gkc} - (n_{gkc}-1)\log(n_{gkc}-1)]\right)^2\right\}, \quad g = 1,\ldots,G;\ k = 1,\ldots,K. \tag{A.2}$$

Moreover, for $k \neq q$, $(\delta_{c,c_{gki}}, \delta_{d,d_{gqi}})$ can be assumed the values (1, 1), (1, 0), (0, 1) and (0, 0) with respective frequencies $n_{gkq,cd}$, $n_{gkc} - n_{gkq,cd}$, $n_{gqd} - n_{gkq,cd}$ and $n_g - n_{gkc} - n_{gqd} + n_{gkq,cd}$. Therefore, for $k \neq q$, the covariance term is

$$\frac{1}{n_g(n_g-1)}\sum_{c=1}^{C}\sum_{d=1}^{C}\left\{n_{gkq,cd} - \frac{n_{gkc}n_{gqd}}{n_g}\right\}u_{gkc}u_{gqd}.$$

Thus, the jackknife variance estimator of $\mathrm{I}_E(\mathbf{\Pi}_g)$ is

$$\widehat{\sigma}^2_{Jg} = \frac{1}{n_g(n_g-1)K^2}\sum_{k=1}^{K}\sum_{q=1}^{K}\sum_{c=1}^{C}\sum_{d=1}^{C}\left\{n_{gkq,cd} - \frac{n_{gkc}n_{gqd}}{n_g}\right\}u_{gkc}u_{gqd}, \quad g = 1,\ldots,G, \tag{A.3}$$

where the u_{gkc} are defined as before.

References

[1] P.K. Sen, M.T. Tsai, J.S. Jou, High-dimension, low-sample size perspectives in constrained statistical inference: the SARSCoV RNA genome in illustration, J. Amer. Statist. Assoc. 102 (2007) 686–694.
[2] M.J. Silvapulle, P.K. Sen, Constrained Statistical Inference. Inequality, Order, and Shape Restrictions, Wiley, New York, 2005.
[3] C.W. Gini, Variabilita e mutabilita, in: Studi Economico-Giuridici della R. Univ. Cagliary, vol. 2, 1912, pp. 3–159.
[4] E.H. Simpson, The measurement of diversity, Nature 163 (1949) 688.
[5] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423. 623–656.
[6] H.P. Pinheiro, F. Seillier-Moiseiwitsch, P.K. Sen, J. Eron, Genomic sequences and quasi-multivariate CATANOVA, in: Handbook of Statistics, vol. 18, 2000, pp. 713–746.
[7] T.Y. Tzeng, W. Byerley, B. Devlin, K. Roeder, L. Wasserman, Outlier detection and false discovery rates for whole-genome DNA matching, J. Amer. Statist. Assoc. 98 (2003) 236–246.
[8] D.J. Schaid, S.K. McDonnell, S.J. Hebbring, J.M. Cunningham, S.N. Thibodeau, Nonparametric tests of association of multiple genes with human disease, Am. J. Hum. Genet. 76 (2005) 789–793.
[9] S. Kullback, R. Leibler, On information and sufficiency, Ann. Math. Statist. 22 (1951) 79–86.
[10] H.P. Pinheiro, A.S. Pinheiro, P.K. Sen, Comparison of genomic sequences using the Hamming distance, J. Statist. Plann. Inference 130 (2005) 325–339.
[11] W. Hoeffding, A class of statistics with asymptotically normal distribution, Ann. Math. Statist. 19 (1948) 293–325.
[12] P.K. Sen, Some invariance principles relating to jackknifing and their role in sequential analysis, Ann. Statist. 5 (1977) 315–329.
[13] S.N. Roy, On a heuristic method of test construction and its use in multivariate analysis, Ann. Math. Statist. 24 (1953) 220–238.
[14] L.H.Y. Chen, Poisson approximation for dependent trials, Ann. Probab. 3 (1975) 534–545.
[15] P.K. Sen, Kendall's tau in high-dimensional genomic parsimony, in: IMS Collection, vol. 3, 2008, pp. 251–266.
[16] R. Arratia, L. Goldstein, L. Gordon, Poisson approximation and the Chen–Stein method: rejoinder, Statist. Sci. 5 (1990) 432–434.
[17] M. Sibuya, Bivariate extreme statistics, I, Ann. Inst. Statist. Math. 11 (1959) 195–210.
[18] S.H. Yeh, H.Y. Wang, C.Y. Tsai, C.L. Kao, J.Y. Yang, H.W. Liu, I.J. Su, S.F. Tsai, D.S. Chen, P.J. Chen, Characterization of severe acute respiratory syndrome coronavirus genomes in Taiwan: molecular epidemiology and genome evolution, Proc. Natl. Acad. Sci. 101 (2004) 2542–2547.
[19] M.T. Tsai, P.K. Sen, Asymptotically optimal tests for parametric functions against ordered functional alternatives, J. Multivariate Anal. 95 (2005) 37–49.