AUSTRIAN JOURNAL OF STATISTICS
Volume 35 (2006), Number 2&3, 197–214

Robust Statistical Inference for High-Dimensional Data Models with Application to Genomics

Pranab Kumar Sen
University of North Carolina, Chapel Hill, U.S.A.

Abstract: In high-dimension (K) low sample size (n) environments, often nonlinear, inequality, order or general shape constraints crop up in complex ways, and as a result, likelihood based optimal statistical inference procedures may not exist, at least, may not be in manageable form. While some of these inference problems can be treated in asymptotic setups, the curse of dimensionality (i.e., $K >> n$ with often n small) calls for a different type of asymptotics (in K) with different perspectives. Roy's union-intersection principle provides some alternative approaches, generally more amenable for $K >> n$ environments. This scenario is appraised with two important statistical problems in genomic studies: a large number of (possibly dependent) genes with heterogeneity amidst a smaller sample create impasses for standard robust inference. These perspectives are examined here in a nonstandard statistical analysis.

Keywords: Dimensional Asymptotics, Genetic Variability, Hamming Distance, Likelihood Principle, Union-Intersection Principle, Subgroup-Decomposability, Microarrays, Nucleotides.

1 Introduction

For large and possibly inequality, order or shape constrained parameter (and/or sample) spaces, even in parametric setups, exact (and optimal) statistical inference based on the classical *likelihood principle* (LP) may be computationally (or even theoretically) too difficult to formulate; often they may not exist in adaptable forms. Without much success with finite-sample optimality, statistical inference in such nonstandard/nonregular environments has mostly taken an asymptotic (n large) approach that goes far beyond the parametrics, albeit optimality properties may not transpire universally. Roy (1953) *union-intersection principle* (UIP) seems to have some relative advantages in this setup (Silvapulle and Sen, 2004, Tsai and Sen, 2005).

For *high-dimension low sample-size* (HDLSS) (i.e., $K >> n$) environments (with often n small), inference perspectives are quite different than conventional models where K is small while $n >> K$. *Robustness* perspectives are even more nonstandard: Departures from model assumptions can take place in more complex and involved ways. The usual concepts of *influence function, breakdown point* and *error-contamination* (all posed in local robustness contexts) need to be addressed in a far more complex setup: sparse activity in high-dimensional data setups may distort local robustness perspectives considerably, and thereby, call for more complex and generally highly nonstandard statistical measures.

The ongoing evolution of information and bio-technology has created an abundance of complex and enormously large dimensional data models in some interdisciplinary research setups. The $K >> n$ with n possibly small scenario dominates in *genomics* (and

bioinformatics at large), a different kind of asymptotics (in K, not n) being the focal point of statistical modelling and analysis in this context. Two important and challenging statistical problems in genomics, namely, (i) high-dimensional *gene expressions* in *microarray* studies, and (ii) purely qualitative *nucleotides* in *DNA / RNA* studies, are appraised here with the help of a *subgroup decomposability* characterization (Sen, 1999; H. P. Pinheiro et al., 2005; A. S. Pinheiro et al., 2005).

The *curse of dimensionality* has led to some rather challenging tasks for valid and effective statistical appraisals in genomic studies. Most of the standard statistical inference tools are of limited utility in such $K >> N$, with n possibly small, setups. The response variables (even when continuous) may be distinctly nonnormal, (usually) count or discrete, and in some cases, are purely qualitative. This feature with the high-dimensionality makes it unreasonable to adopt standard (continuous / discrete or categorical) multivariate models where the number of associated parameters may outnumber the sample size, and thus creating roadblocks in statistical analysis. To illustrate this drawback, we consider in Section 2 a simple multivariate analysis of variance (MANOVA) model and examine the difficulties arising when $K >> n$. An alternative approach based on robust estimation of dispersion and a suitable subgroup decomposability property is explored in this setup. Section 3 deals with (differential) gene expressions in microarray studies where quantitative (usually count), possibly nonnormal variables in a huge number (dimension) vitiate the adaptability of standard MANOVA tools. We consider some nonparametric resolutions based on suitable *stochastic ordering* characterizations. Section 4 relates to some recent findings of A. S. Pinheiro et al. (2005) where the *Hamming distance* in a purely categorical setup is incorporated in the subgroup decomposition. Section 5 deals with the general distributional asymptotics pertaining to the $K >> n$ with possibly n small environment. The concluding section deals with some general remarks.

2 $K >> n$ Impasses in MANOVA

Consider a one-way MANOVA model with $G(\geq 2)$ groups of samples of sizes $n_1, \ldots, n_G$, respectively. The sample observations (K-vectors) in the gth group are denoted by $\mathbf{X}_{g1}$, $\ldots, \mathbf{X}_{gn_g}$, and are assumed to be independent and identically distributed (i.i.d.) with a K-variate distribution F_g; the mean vector and dispersion matrix for this distribution (assumed to be finite and positive definite) are denoted by $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ respectively, for $g = 1, \ldots, G$; all these $n = \sum_{g=1}^{G} n_g$ r.v.'s are assumed to be independent. We want to test for the null hypothesis

$$H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_G \,, \tag{1}$$

against the set of alternatives

$$H_1 : \max\{||\boldsymbol{\mu}_j - \boldsymbol{\mu}_l|| : 1 \leq j < l \leq G\} > 0 \,, \tag{2}$$

with some further regularity assumptions on the $\boldsymbol{\Sigma}_g$. In the conventional case, we assume that the F_g are all multinormal and further the $\boldsymbol{\Sigma}_g$ are all equal. A further condition needed to have manageable distribution theory of the test statistic is

$$n - G > K \quad \text{i.e.} \quad n > G + K \,. \tag{3}$$

The classical MANOVA tests are based on the following two (random) matrices:

$$\mathbf{S}_W = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\mathbf{X}_{ig} - \bar{\mathbf{X}}_g)(\mathbf{X}_{ig} - \bar{\mathbf{X}}_g)' ,$$
$$\mathbf{S}_B = \sum_{g=1}^{G} n_g (\bar{\mathbf{X}}_g - \bar{\mathbf{X}}_.)(\bar{\mathbf{X}}_g - \bar{\mathbf{X}}_.)' , \tag{4}$$

where $\bar{\mathbf{X}}_g$ is the gth group mean vector, $g = 1, \ldots, G$ and $\bar{\mathbf{X}}_.$ is the combined group mean vector. These are known as the 'within group' and 'between group' (centered) sum of product matrices. The Lawley-Hotelling trace statistic (= trace$(\mathbf{S}_B \mathbf{S}_W^{-1})$), Wilks' likelihood ratio criterion (= $|\mathbf{S}_B|/|\mathbf{S}_B + \mathbf{S}_W)$ and the Roy largest root criterion (= largest characteristic root of $\mathbf{S}_B \mathbf{S}_W^{-1}$) are all functions of the characteristic roots of $\mathbf{S}_B \mathbf{S}_W^{-1}$, and are all affine-invariant statistics. For $G = 2$, all the three statistics are equivalent and the test for H_0 based on either of them is best (i.e., uniformly most powerful) invariant. For $G \geq 3$, not only these statistics might differ from each other but also may not possess this best invariance property. Asymptotically, the likelihood ratio test has the best average power property over suitable ellipsoidal contours in the parameter space. Apart from the basic requirement that $n > K + G$, these tests being based on second order sample moments are generally quite *nonrobust* for gross-error contamination, outliers, non(multi-)normality and possible heterogeneity of the $\mathbf{\Sigma}_g$. These nonrobustness aspects are shared by the sample mean vector $\mathbf{X}_g$ but to a relatively lesser extent.

Even under the assumed multi-normality condition, the rank of $\mathbf{S}_W$ is $(n - g) \cap K$ and the rank of $\mathbf{S}_B$ is $(G - 1) \cap K$. So, if $n - G < K$, i.e., in the $K >> n$ environment, both $\mathbf{S}_W$ and $\mathbf{S}_B$ are highly singular matrices, creating impasses for the proper definition of any of these three statistics. Further, because of this degeneracy, there is no simple way of finding out their null distributions in a manageable form (while the nonnull distribution theory is admittedly much more complex). The situation is totally out of hand when n is small but K is very large, as is the case in genomics (we shall see in the next section). We therefore take an alternative approach based on subgroup-decomposability which does not put much emphasis on the high-dimensional homoscedasticity (i.e., homogeneity of the $\mathbf{\Sigma}_g$ when K is greater than n), and amends well in the $K >> n$ environment.

Assume that the distribution F_g (not necessarily multi-normal) admits finite mean vector $\boldsymbol{\mu}_g$ and finite (positive definite) dispersion matrix $\mathbf{\Sigma}_g$, for $g = 1, \ldots, G$. Then based on the (symmetric) kernel (matrix of degree 2):

$$\psi(\mathbf{a}, \mathbf{b}) = \frac{1}{2}(\mathbf{a} - \mathbf{b})(\mathbf{a} - \mathbf{b})' , \quad \mathbf{a}, \mathbf{b} \in \mathbb{R}^K , \tag{5}$$

we consider some estimable parameters (matrices) basically related to the within and between group dispersion matrices. Let

$$\mathbf{\Gamma}_{gg} = E\psi(\mathbf{X}_{g1}, \mathbf{X}_{g2}) = \mathbf{\Sigma}_g , \quad g = 1, \ldots, G ; \tag{6}$$

$$\mathbf{\Gamma}_{gg'} = E\psi(\mathbf{X}_{g1}, \mathbf{X}_{g'1}) = \frac{1}{2}(\mathbf{\Gamma}_{gg} + \mathbf{\Gamma}_{g'g'}) + \frac{1}{2}(\boldsymbol{\mu}_g - \boldsymbol{\mu}_{g'})(\boldsymbol{\mu}_g - \boldsymbol{\mu}_{g'})' , \tag{7}$$

for $g \neq g' = 1, \ldots, G$. Therefore, noting that the second matrix on the right hand side is p.s.d. (of rank 1), we claim that for every pair $(g, g') : 1 \leq g < g' \leq G$,

$$\boldsymbol{\Delta}_{gg'} = 2\boldsymbol{\Gamma}_{gg'} - \boldsymbol{\Gamma}_{gg} - \boldsymbol{\Gamma}_{g'g'} \tag{8}$$

is p.s.d., and for every $\boldsymbol{\lambda} \in \mathbb{R}^K$,

$$\boldsymbol{\lambda}'\boldsymbol{\Delta}_{gg'}\boldsymbol{\lambda} = ||\boldsymbol{\lambda}'(\boldsymbol{\mu}_g - \boldsymbol{\mu}_{g'})||^2 \geq 0\,, \tag{9}$$

where the equality sign holds only when $\boldsymbol{\lambda}'(\boldsymbol{\mu}_g - \boldsymbol{\mu}_g) = 0$. Motivated by this feature, we consider a (Kiefer-)class $\Phi = \{\phi(\cdot)\}$ of nonnegative real valued functions $\phi(\mathbf{A})$ of a $K \times K$ p.s.d. matrix $\mathbf{A}$ satisfying the inequality

$$\phi(\boldsymbol{\Gamma}_{gg'}) \geq \frac{1}{2}\left(\phi(\boldsymbol{\Gamma}_{gg} + \phi(\boldsymbol{\Gamma}_{g'g'})\right)\,, \quad \forall\, \phi \in \Phi\,, \tag{10}$$

where the equality sign holds only when $\boldsymbol{\mu}_g = \boldsymbol{\mu}_{g'}$, $g \neq g' = 1, \ldots, G$.

Note that an optimal, unbiased, nonparametric estimator of $\boldsymbol{\Gamma}_{gg}$ is

$$\mathbf{S}_{gg} = \binom{n_g}{2}^{-1} \sum_{1 \leq i < j \leq n_g} \frac{1}{2}(\mathbf{X}_{gi} - \mathbf{X}_{gj})(\mathbf{X}_{gi} - \mathbf{X}_{gj})'\,, \tag{11}$$

for $g = 1, \ldots, G$. Similarly, we have

$$\mathbf{S}_{gg'} = (n_g n_{g'})^{-1} \sum_{i=1}^{n_g} \sum_{j=1}^{n_{g'}} \frac{1}{2}(\mathbf{X}_{gi} - \mathbf{X}_{g'j})(\mathbf{X}_{gi} - \mathbf{X}_{g'j})' \tag{12}$$

an optimal unbiased nonparametric estimator of $\boldsymbol{\Gamma}_{gg'}$, for $g \neq g' = 1, \ldots, G$. Note that these estimates are (Hoeffding, 1948) (generalized) U-*statistics* and they adapt well in many environments, including $K >> n$. In the present context, G, the number of groups (samples) is assumed to be fixed (and usually small). In this respect, we incorporate a subgroup decomposability characterization (similar to the ANOVA decomposition), considered in Sen (1999) and H. P. Pinheiro et al. (2005), A. S. Pinheiro et al. (2005). In their case, the specific case of $\phi(\mathbf{A})$ used is the trace criterion (i.e., trace$(\mathbf{A}) = \sum_{j=1}^{K} a_{jj}$) which is analogous to the Hamming distance, and we shall discuss that briefly in the next section.

In addition to the statistics $\mathbf{S}_{gg'}$, $1 \leq g \leq g' \leq G$, we consider some pooled sample statistics defined below. We pool the G sample observations and denote the n observation (vectors) as $\mathbf{X}_r^*$, $1 \leq r \leq n$, where the first n_1 observations correspond to the first group, the next n_2 to the second, and so on (the last n_G to the Gth group). Then, based on the same kernel $\psi(\cdot)$, we have the combined sample U-statistic (matrix)

$$\begin{aligned} \mathbf{S}_0 &= \binom{n}{2}^{-1} \sum_{1 \leq r < s \leq n} \psi(\mathbf{X}_r^*, \mathbf{X}_s^*) \\ &= (n-1)^{-1} \sum_{r=1}^{n} (\mathbf{X}_r - \bar{\mathbf{X}}_0)(\mathbf{X}_r - \bar{\mathbf{X}}_0)'\,, \end{aligned} \tag{13}$$

where $\bar{\mathbf{X}}_0$ is the pooled sample mean vector and whenever there is no confusion, we denote the $\mathbf{X}_r^*$ by $\mathbf{X}_r$. Then, by routine steps we obtain that

$$\binom{n}{2}\mathbf{S}_0 = \sum_{g=1}^{G}\binom{n_g}{2}\mathbf{S}_{gg} + \sum_{1\le g\ne g'\le G} n_g n_{g'}\mathbf{S}_{gg'}\,, \tag{14}$$

and using the identity that

$$\binom{n_g}{2}\binom{n}{2}^{-1} = \frac{n_g}{n} - \frac{n_g(n-n_g)}{n(n-1)}\,, \quad \forall\, g = 1,\ldots,G\,,$$

we have

$$\mathbf{S}_0 = \sum_{g=1}^{G}\frac{n_g}{n}\mathbf{S}_{gg} + \sum_{1\le g<g'\le G}\frac{n_g n_{g'}}{n(n-1)}(2\mathbf{S}_{gg'} - \mathbf{S}_{gg} - \mathbf{S}_{g'g'})\,. \tag{15}$$

Note that when $\phi(\mathbf{A}) = \text{trace}(\mathbf{A})$, the last equation leads us to the following

$$\phi(\mathbf{S}_0) = \sum_{g=1}^{G}\frac{n_g}{n}\phi(\mathbf{S}_{gg}) + \sum_{1\le g<g'\le G}\frac{n_g n_{g'}}{n(n-1)}(2\phi(\mathbf{S}_{gg'}) - \phi(\mathbf{S}_{gg}) - \phi(\mathbf{S}_{g'g'}))\,, \tag{16}$$

as $\phi(\cdot)$ is additive. We denote the first and second terms on the right hand side of the last equation by $\hat{\Delta}_W$ and $\hat{\Delta}_B$ respectively; they represent the within and between group components in this subgroup representation. Note that this subgroup decompasibility is somewhat different from the classical MANOVA decomposability, and here we are not imposing the affine invariance which underlies the MANOVA case. The between group component $\hat{\boldsymbol{\Delta}}_B$ is an unbiased estimator of

$$\Delta_B = \sum_{1\le g<g'\le G}\frac{n_g n_{g'}}{n(n-1)}\,\text{trace}(2\boldsymbol{\Gamma}_{gg'} - \boldsymbol{\Gamma}_{gg} - \boldsymbol{\Gamma}_{g'g'})\,, \tag{17}$$

which is a nonnegative entity, and is zero when the nullhypothesis $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_G$ holds, irrespective of the homogeneity of the $\boldsymbol{\Sigma}_g$. Thus, Δ_B is positive when the mean vectors are not all the same, irrespective of possible heterogeneity of the dispersion matrices. This feature makes Δ_B insensitive to possible heteroscedasticity of the G groups. The sample counterpart $\hat{\Delta_B}$ being unbiased, has under $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_G$ zero expectation, and hence takes on both positive and negative values; it may also take on negative as well as positive values when H_0 does not hold, but it will be stochastically more positive. The within group component $\hat{\Delta}_W$ is insensitive to the null hypothesis H_0 being true or not, and thereby serves as a good scaling factor for the between group component. Hence, it seems logical to use the following (analogous to the ANOVA) test statistic

$$\mathcal{L}_1^* = \hat{\Delta}_B/\hat{\Delta}_W\,, \tag{18}$$

for testing H_0 against $H_1 : \boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_G$ not all equal, rejecting the null hypothesis for large positive values of $\mathcal{L}_1^*$. By its formulation, the test is robust to possible heteroscedasticity of the dispersion matrices. However, being based on the sample second order moments,

this test (like the ANOVA tests) is likely to be nonrobust to gross-error contamination or outliers.

It is, therefore, of some interest to consider some alternative test tests based on a similar subgroup decomposability but likely to be more robust in this respect. With that in mind, we consider a (symmetric) kernel (of degree 2)

$$\psi(\mathbf{a},\mathbf{b}) = \sum_{k=1}^{K} |a_k - b_k| \,, \quad \mathbf{a},\mathbf{b} \in \mathbb{R}^K \tag{19}$$

and note that

$$E\psi(\mathbf{X}_{gi},\mathbf{X}_{gj}) = \sum_{k=1}^{K} E|X_{gki} - X_{gkj}| = \sum_{k=1}^{K} \gamma^*_{gg,k} \,, \tag{20}$$

where $\gamma^*_{gg,k}$ is the Gini mean difference for the kth marginal distribution of F_g, the joint distribution of $\mathbf{X}_{g1}$, for $g = 1,\ldots,G$. In the same vein, we have

$$E\psi(\mathbf{X}_{gi},\mathbf{X}_{g'j}) = \sum_{k=1}^{K} E|X_{gki} - X_{g'kj}| = \sum_{k=1}^{K} \gamma^*_{gg',k} \,, \tag{21}$$

for $g \neq g' = 1,\ldots,G$. The additively decomposition for the mean square distance (displayed in (5) - (7)) may not pertain to the case of the $\gamma_{gg',k}$. Nevertheless, under certain mild conditions, an inequality like in (10) holds.

Consider first a shift model (without necessarily assuming that all the underlying (marginal) distributions are symmetric). Thus, we take

$$F_g(\mathbf{x}) = F_0(\mathbf{x} - \boldsymbol{\mu}_g) \,, \quad \mathbf{x} \in \mathbb{R}^K \,,\ g = 1,\ldots,G \,, \tag{22}$$

where F_0 has not necessarily marginal distributions symmetric about 0. In this setup, note that for every g, k, $X_{gki} - X_{gkj}$, $i \neq j$ has a distribution (say F_0^*) symmetric about 0, so that the $\gamma_{gg,k}$ do not depend on the $\boldsymbol{\mu}_g$, and

$$\gamma^*_{gg,k} = \gamma^*_{0,k} \,, \quad \forall\, g = 1,\ldots,G;\ k = 1,\ldots,K \,, \tag{23}$$

where, F_0^* is obtained from F_0 by convolution, and of course, the $\gamma^*_{0,k}$ may vary from one $k = 1,\ldots,K$ to another. Further, for this shift model, for any pair $(g,g') :\ g \neq g' (= 1,\ldots,G)$,

$$(\mathbf{X}_{gi} - \mathbf{X}_{g'j}) - (\boldsymbol{\mu}_g - \boldsymbol{\mu}_{g'}) \tag{24}$$

has a distribution (same as that of $\mathbf{X}_{gi} - \mathbf{X}_{gj}$) which is symmetric about $\mathbf{0}$. Thus, $X_{gki} - X_{g'kj}$ has the median $\mu_{gk} - \mu_{g'k}$ (the same as their mean). Further, by the well known fact that for any distribution, the mean absolute deviation is a minimum about the median (= mean under symmetry), we obtain that for every $k = 1,\ldots,K$, $1 \leq g \neq g' \leq G$,

$$\begin{aligned} \gamma^*_{gg',k} &= E|X_{gki} - X_{g'kj}| \\ &\geq E|X_{gki} - X_{g'kj} - \mu_{gk} + \mu_{g'k}| \\ &= \gamma^*_{gg,k} = \gamma^*_{g'g',k} = \frac{1}{2}(\gamma^*_{gg,k} + \gamma^*_{g'g',k}) \,, \end{aligned} \tag{25}$$

where the equality sign holds only when $\mu_{gk} = \mu_{g'k}$. Thus, in this setup, the subgroup decomposability holds though the strict additively decomposability may not. This shows that under the homogeneity of the $\mathbf{\Sigma}_g$, even without normality, the Gini mean difference based measures satisfy the subgroup decomposability and an ANOVA type test can be incorporated under suitable regularity conditions.

Consider next an intermediate case where the marginal distributions (of the G K-variate distributions $F_1, \ldots, F_G$) are normal but not necessarily homoscedastic. Then, for every $g = 1, \ldots, G$ and $k = 1, \ldots, K$, and for $i \neq j$, the difference $X_{gki} - X_{gkj}$ has normal distribution with 0 mean and variance $2\sigma_{g,kk}$, so that $\gamma^*_{gg,k} = \sqrt{2/\pi}\sqrt{2\sigma_{gg,k}}$. On the same count, for $g \neq g'$, $X_{gki} - X_{g'kj}$ is normal with mean $\mu_{gk} - \mu_{g'k}$ and variance $(\sigma_{gg,k} + \sigma_{g'g',k})$, so that

$$E|X_{gki} - X_{g'kj} - \mu_{gk} + \mu_{g'k}| = \sqrt{\frac{2}{\pi}}\left(\sqrt{\sigma_{gg,k} + \sigma_{g'g',k}}\right) , \tag{26}$$

where by the classical moment inequality,

$$\frac{1}{2}(\sigma_{gg,k} + \sigma_{g'g',k}) \geq (\frac{1}{2}(\sqrt{\sigma_{gg,k}} + \sqrt{\sigma_{g'g',k}}))^2 , \tag{27}$$

where the equality sign holds holds only when $\sigma_{gg,k} = \sigma_{g'g',k}$. The last two expressions lead us to the following.

$$\begin{aligned}
\gamma^*_{gg',k} &= E|X_{gki} - X_{g'kj}| \geq \sqrt{\frac{2}{\pi}}(\sigma_{gg,k} + \sigma_{g'g',k})^{1/2} \\
&\geq \frac{1}{2}(\sqrt{\frac{2}{\pi}}(\sqrt{2\sigma_{gg,k}} + \sqrt{2\sigma_{g'g',k}}) \\
&= \frac{1}{2}(\gamma^*_{gg,k} + \gamma^*_{g'g',k}) ,
\end{aligned} \tag{28}$$

where in the first line, the equality sign holds only when $\mu_{gk} = \mu_{g'k}$ while in the penultimate line, the equality sign holds only when $\gamma^*_{gg,k} = \gamma^*_{g'g',k}$, for $k = 1, \ldots, K$, $g \neq g' = 1, \ldots, G$. Thus, even under H_0, $\gamma^*_{gg',k} \geq (\gamma^*_{gg,k} + \gamma^*_{g'g',k})/2$ with the equality sign holding only when $\gamma^*_{gg,k} = \gamma^*_{g'g',k}$ that is in the marginal homoscedastic case. If we take, as before, the sample counterparts of these $\gamma^*_{gg',k}$ by $m^*_{gg',k}$, that is

$$m^*_{gg,k} = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n_g} |X_{gki} - X_{gk,j}| , \tag{29}$$

$$m_{gg',k} = (n_g n_{g'})^{-1} \sum_{i=1}^{n_g} \sum_{j=1}^{n_{g'}} |X_{gki} - X_{g'kj}| , \tag{30}$$

for $g \neq g' = 1, \ldots, G$; $k = 1, \ldots, K$, then the 'between group' component may be defined as before by

$$\sum_{1 \leq g < g' \leq G} \frac{n_g n_{g'}}{n(n-1)} \sum_{k=1}^{K} \left[2m_{gg',k} - m_{gg,k} - m_{g'g',k}\right] . \tag{31}$$

Note that in the general heteroscedastic case, even under H_0, (31) has a nonnegative expectation that depends on the degree of divergence of the $\gamma^*_{gg',k}$, $g, g' = 1, \ldots, G$, $k = 1, \ldots, K$. A similar feature holds for the between group component under the Huber contamination model as well as general mixture normal distributions. This implies that in such heteroscedastic environments, the use of (31) entails the estimation of the centering constant, requiring in turn the estimation of the variances and covariances by some other methods. Thus, we will confront nonrobustness aspects possibly to a similar extent as in the case of the conventional variances. On the other hand, if we desire to have good robustness properties without the normality assumption but retaining the homoscedasticity assumption, the Gini mean difference based testing procedure can be advocated as a strong contender to the mean square errors based measure.

The main advantage of using the Hamming-distance type measures as adapted for quantitative responses (or trace-criterion based measures) comes from its amenability in the $K >> n$ environment with flexibility to choose a more robust version of the sample second order moments. We shall illustrate this point more with the gene expression model in the next section.

3 Microarray Gene Expression Model

Microarray technology allows studies of typically thousands of genes (K), possibly differentially expressed under diverse biological / experimental setups, simultaneously with only a few (n) arrays. Such experiments are excessively costly, thus preempting the possibility of having a very large number of arrays, and resulting in the $K >> n$ environment. Gene expression differentials under different environment cast light on plausible *gene-environment interaction* (or association), and thus may help *mapping disease genes* whenever the arrays are so designed (e.g., normal vs. HIV positive patients at different stages of AIDS infliction). Nevertheless, there are statistical challenges in microarray data models (Sebastiani et al., 2003).

Typically, in an array there is a large number (K) of genes whose expression levels are measured by their color intensity (or luminosity) as a quantitative variate on the $[0, 1]$ scale or as percentage ranging from 0 to 100. A gene associated (causally or statistically) with a target disease is termed a *disease gene* (DG), and others as nondisease genes (NDG). Typically, only a few are DG while the vast majority NDG. A NDG is expected to have a low gene expression level while DG is likely to have a high expression level. Thus, there may be a natural *stochastic ordering* of the gene expression levels of DG's with varying disease severity while the NDG expression levels are expected to be stochastically unaffected by such disease level differentials. This makes it appealing to incorporate MANOVA models to test for gene-environment interaction, albeit in the $K >> n$ environment with two different classes of genes (DG and NDG).

Suppose that there are Q DG and the remaining $K - Q$ NDG. Typically, $Q << K$ and it is unknown. In that setup, the primary statistical task is to ascertain the Q DG's in a statistical manner; both Q and the tags of the DG's are unknown. Moreover, the expression levels of all the K genes may not be statistically independent. However, for the NDG, the expression levels being stochastically small, such inter-dependence is

expected to be small and hence could be taken as statistically independent. But for the DG, the stochastic interdependence may not be negligible. In the same vein, stochastic dependence between DG and NDG's can be ignored. Therefore, in our modelling in a simple setup, we can assume that the gene expression level of the kth gene, denoted by X_k, $k = 1, \ldots, K$, though not stochastically independent, satisfy the following condition:

$$K^{-1}\text{Var}\left(\sum_{k=1}^{K} X_k\right) < \infty\,. \tag{32}$$

Or in other words, possible DG clustering effect does not alter the convergence order of the variance of the sum. If Q is $O(K^{1/2})$ (or of smaller order), then (32) holds when the NDG's are assumed to be uncorrelated with each other and with the DG's as well. It is also possible to conceive of other form of clustering of the genes for which the sizes of the clusters are not large (while the number of clusters could be large) and the clusters are uncorrelated. Further, if a suitable mixing condition (without possibly the stationarity) can be assumed then under quite general conditions (Yoshihara, 1993) (32) can be validated. We shall make more comments on this assumption in a later section. Of course, when the arrays differ with respect to biological or environmental setups, the variance functions also do so. In this sense, conventional MANOVA models may not be appropriate in microarray studies.

We consider G groups of arrays (relating to $G \geq 2$ possibly different biological / environmental setups), the gth group having $n_g \geq 2$ arrays, $1 \leq g \leq G$; all these $n = \sum_g n_g$ arrays are assumed to be stochastically independent. In each array there are K (a large number of) genes whose expression levels are measured simultaneously by microarray technology. As these levels (denoted by l) are typically between 0 and 1, we use the log-transformation, i.e., $-\log(1-l)$ which will have the range $\mathbb{R}^+ = [0, \infty)$. It is also possible to use some other transformation like the *logit* which corresponds to $\log\{l/(1-l)\}$ or the *normit* which is defined by $\phi^{-1}(l)$, ϕ being the standard normal distribution function. In both the latter cases, the range is transformed to the real line $\mathbb{R}$. For the ith microarray in the gth group, the (K) vector of (possibly transformed) expression levels is denoted by $\mathbf{X}_{gi} = (X_{gi1}, \ldots, X_{giK})'$ where X_{gik} corresponds to the kth gene, for $k = 1, \ldots, K$, $i = 1, \ldots, n_g$, $g = 1, \ldots, G$.

It is usually assumed, albeit often not quite justifiably, that $\mathbf{X}_{gi}$ has a K-variate normal distribution with mean vector $\boldsymbol{\mu}_g$ and dispersion matrix $\boldsymbol{\Sigma}_g$, for $g = 1, \ldots, G$; the $\boldsymbol{\mu}_g$ may then be allowed to be possibly different for different groups but it is traditionally addumed that the $\boldsymbol{\Sigma}_g$ are all the same (but unknown). Thus, the homogeneity of the $\boldsymbol{\Sigma}_g$ and multinormality of the underlying distributions constitute the basic regularity assumptions, thus relating possible group divergence solely in terms of the variability of the $\boldsymbol{\mu}_g$, $1 \leq g \leq G$. Even so, as explained in the previous section, for $K >> n - G$, classical MANOVA tools become unusable. Faced with this discouraging feature of standard multivariate analysis, we consider an alternative approach along the lines of the preceding section. Such a quasi-marginal approach is amenable for the $K >> n$ (with even n small) environment and also for plausible stochastic dependence among the K genes (positions) in an array.

For the kth gene (position), consider the n (random) variables X_{gik}, $i = 1, \ldots, n_g$; $g = 1, \ldots, G$ where k ranges over $1, \ldots, K$. Let T_{nk} be a suitable ANOVA test statistic based on the 'within group' and 'between group' subgroup decomposability (in the sense

of the developments in the preceding section. Both the mean square errors and Gini's mean difference criteria are bonafide members of this class. Note that we are not to assume normality of the distributions, and even sometimes, homogeneity of their scale parameters. In this context, it is also possible to use suitable rank statistics or some robust M-statistics, provided they pertain to subgroup decomposability in a proper sense. We illustrate this with a very simple case of the Wilcoxon-Mann-Whitney (WMW) statistics. For the pair (g, g') of groups, let $W_{gg'k}$ be the WMW statistic, for $1 \leq g < g' \leq G$. Note that, conventionally, we may let $W_{ggk} = 1/2, \forall\ g = 1, \ldots, G$. In the same vein, for the combined group, we have $W_{0k} = 1/2$. Now, if the G groups are stochastically ordered (as is the case when the biological / environmental factors across the groups are in increasing level of dominance), then $W_{gg'k} \geq 1/2, \forall\ 1 \leq g < g' \leq G$, with at least one strict inequality. Thus, if we define

$$\begin{aligned} T_{nk} &= \sum_{1 \leq g < g' \leq G} \frac{n_g n_{g'}}{n(n-1)} \{2W_{gg'k} - W_{ggk} - W_{g'g'k}\} \\ &= \sum_{1 \leq g < g' \leq G} \frac{n_g n_{g'}}{n(n-1)} \{2W_{gg'k} - 1\}\,, \end{aligned} \tag{33}$$

for $k = 1, \ldots, K$, then the subgroup decomposability property holds. (Note that $W_{gg'k} + W_{g'gk} = 1$, for every $g \neq g' = 1, \ldots, G$, so that in the above sum, we may not be able to replace the range $1 \leq g < g' \leq G$ by $1 \leq g \neq g' \leq G$.)

Within this general framework, we let $\mathbf{T}_n = (T_{n1}, \ldots, T_{nK})'$ and consider the null hypothesis H_0 that the G groups are statistically homogeneous, and the alternative hypothesis H_1 relates to possible group divergence with special emphasis on dispersions (for which the subgroup decomposability was formulated in Section 2). Note that for the MW statistics, the dispersion was measured in terms of plausible stochastic ordering of the G groups. We define $\boldsymbol{\tau}_n = (\tau_{n1}, \ldots, \tau_{nK})'$ with the elements

$$\tau_{nk} = E\{T_{nk} | H_0\}\,, \quad k = 1, \ldots, K\,. \tag{34}$$

Under the null hypothesis, we may set without any loss of generality

$$\boldsymbol{\tau}^0_{nk} = E\{\mathbf{T}_n | H_0\} = \mathbf{0}\,. \tag{35}$$

On the other hand, under alternatives of possible group divergence,

$$\boldsymbol{\tau}^*_n = E\{\mathbf{T}_n | H_1\} \geq \mathbf{0}\,, \tag{36}$$

where $||\boldsymbol{\tau}^*_n|| > 0$. Next, we consider a set of nonnegative weights attaching plausible (prior) importance to the K genes. Let $\mathbf{w} = (w_1, \ldots, w_K)'$ satisfying

$$\mathbf{w} \geq \mathbf{0}\,; \quad \mathbf{w}'\mathbf{1} = 1\,. \tag{37}$$

It should be kept in mind that generally there are a proportionately smaller number of genes associated with a specific disease / disorder (or a small group of them), so that if some prior knowledge is acquired on these candidate genes, one can attach greater weight

to them and much less on the rest. Nevertheless, we assume that the Noether condition holds, i.e., as K becomes large,

$$\max\{||\mathbf{w}||^{-1}w_k : 1 \leq k \leq K\} \to 0\,. \tag{38}$$

If no prior information on candidate genes is available, we may take $\mathbf{w} = K^{-1}\mathbf{1}$, i.e., uniform weight to all the K genes. Consider then the convex combination

$$T^*_{wn} = \mathbf{w}'\mathbf{T}_n\,, \tag{39}$$

and note that by construction,

$$E\{T^*_{wn}|H_0\} = 0\,, \quad E\{T^*_{wn}|H_1\} \geq 0\,. \tag{40}$$

This intuitively suggests that a one-sided test (for H_0 vs. H_1) with the rejection of H_0 for large positive values of T^*_{wn} should be in order. In this respect, the crux of the problem is to find suitable critical levels of T^*_{wn} that would correspond to some preassigned level of significance $\alpha(0 < \alpha < 1)$. This is particularly needed in the context where the K genes may not be statistically independent, nor even have marginally homogeneous distributions.

4 Hamming Distance and Qualitative Data Models

Let us consider now the case of qualitative data models, as is typically encountered in DNA nucleotide or RNA protein data models. Here usually there is a large number (K) of genes or positions and in each position the response variables corresponds to one of the C possible outcomes (e.g., $C = 4$ for A, C, G and T for nucleotides, or the codons, some 20 in number, for RNA data). These outcomes are not ordered, even in a weak sense, nor the positions are likely to be stochastically independent. Thus, we encounter a multi-dimensional categorical data model with C^K possible outcomes, implying that the probability law is defined on a (C^K-)-dimensional simplex, while we may have a handful (n) of sequences (or samples) where typically, $K >> n$, so that C^K is even of much higher order compared to n. Since not all the C^K cells are likely to be equally likely, some (if not many) of the cells will have very low probability (or frequency count), creating impasses for standard discrete multivariate analyzes to be properly applicable in this setup. This is essentially the curse of dimensionality problem in qualitative genomics. However, mutations due to possible environmental factors are more likely to be associated with increased genetic variability, causing gene-environment interaction (viz., Coffin, 1986; Hahn et al., 1986). As such, we conceive of a similar G group paradigm corresponding to possibly different biological/environmental setups, and would like to have suitable subgroup-diversity analysis.

As in Section 3, let $\mathbf{X}_{gi} = (X_{gi1}, \ldots, X_{giK})'$ stand for the observation vector for the ith unit in the gth group, for $i = 1, \ldots, n_g(\geq 2)$; $g = 1, \ldots, G$, where each X_{gik} takes on one of the C possible realizations, labelled as $1, \ldots, C$ (without any ordering of the labels). The probability law of $\mathbf{X}_{gi}$ is denoted by $\boldsymbol{\pi}_g = \{\pi_g(\mathbf{c}) : \mathbf{c} \in \mathcal{C}\}$ where $\mathcal{C} = \{(c_1, \ldots, c_K) : c_k = 1, \ldots, C,\ k = 1, \ldots, K\}$, and $g = 1, \ldots, G$. Basically, we are

interested in testing for the null hypothesis (H_0) of homogeneity of the $\boldsymbol{\pi}_g$, $g = 1, \ldots, G$. The class of alternative hypotheses (of possible heterogeneity of these laws) is so big that in the environment $K >> n$, conventional categorical data model tests are of very little power. Therefore, we need to address specific subclasses of alternatives having meaningful genomic as well as statistical interpretations, and for such directed alternatives want to develop more powerful tests.

For each sequence (say, $\mathbf{X}_{gi}$ we may conceive of a transition model where at site k, the response category is c_k and there is a transition from c_k to c_{k+1} from the site k to $k+1$, for $k = 1, \ldots K-1$, and each c_k assuming one of the C possible response labels $1, \ldots, C$. We also consider the marginal (multinomial) law for the very first site by $\boldsymbol{\pi}_{g1}$. It might be tempting to conceive suitable Markov chains to describe this stochastic flow of responses from one site to the next one. Even so, the stationarity of the transition probabilities may not be generally tenable, and as a result, the total number of parameters arising in this modelling would be tremendously large, creating a similar impasse for standard statistical analysis for simple stochastic processes to be genuinely applicable. On top of that the usual $K >> n$ environment can totally vitiate the use of nonstationary Markov chain models in this context.

With established affinity of bio-diversity and genetic variability measures (see for example, Chakraborty and Rao, 1991 for an excellent review), it is natural to incorporate for each coordinate the celebrated Gini-Simpson index (Gini, 1912; Simpson, 1949; H. P. Pinheiro et al., 2005; A. S. Pinheiro et al., 2005), and combine them linearly for a composite measure. This is essentially the use of the Hamming distance in such qualitative group divergence studies. For a pair of observations, say $\mathbf{X}_i$ and $\mathbf{X}_j$, each having K coordinates with each coordinate taking on the labels $1, \ldots, C$, let us define the Hamming distance as

$$d_H(\mathbf{X}_i, \mathbf{X}_j) = K^{-1} \sum_{k=1}^{K} I(X_{ki} \neq X_{kj}) \,, \tag{41}$$

which can only take on the values $0/K, 1/K, \ldots, K/K$ with a probability law that depend on the joint law of $\mathbf{X}_i, \mathbf{X}_j$. It is clear that

$$\delta_H = E\{d_H(\mathbf{X}_i, \mathbf{X}_j)\} = K^{-1} \sum_{k=1}^{K} P\{X_{ki} \neq X_{kj}\} \,. \tag{42}$$

Thus, if we denote the marginal multinomial law for X_{gi} by $\boldsymbol{\pi}_{gk} = (\pi_{gk1}, \ldots, \pi_{gkC})'$ for $k = 1, \ldots, K$; $g = 1, \ldots, G$, then we have

$$\begin{aligned} \mathcal{H}_{gg} &= E\{d(\mathbf{X}_{gi}, \mathbf{X}_{gj}\} = K^{-1} \sum_{k=1}^{K} \{1 - \boldsymbol{\pi}'_{gk} \boldsymbol{\pi}_{gk}\} \\ &= K^{-1} \sum_{k=1}^{K} \sum_{c=1}^{C} \pi_{gkc}(1 - \pi_{gkc}) \end{aligned} \tag{43}$$

which is the arithmetic mean of the K gene-wise Gini-Simpson indexes. In the same vein,

we define

$$\begin{aligned}\mathcal{H}_{gg'} &= E\{d(\mathbf{X}_{gi}, \mathbf{X}_{g'j}) \\ &= K^{-1}\sum_{k=1}^{K}\{1 - \boldsymbol{\pi}'_{gk}\boldsymbol{\pi}_{g'k}\} \\ &= K^{-1}\sum_{k=1}^{K}\sum_{c=1}^{C}\pi_{gkc}(1-\pi_{g'kc})\,, \end{aligned} \tag{44}$$

as the population Hamming distance between the gth and g'th groups, for $g \neq g' = 1, \ldots, G$.

Using the last two equations, we immediately obtain that for every pair (g, g') of groups,

$$2\mathcal{H}_{gg'} \geq \mathcal{H}_{gg} + \mathcal{H}_{g'g'}\,, \quad 1 \leq g \neq g' \leq G\,, \tag{45}$$

where the equality sign holds only when $\boldsymbol{\pi}gk = \boldsymbol{\pi}_{g'k}$, $\forall k = 1, \ldots, K$, i.e., all the G sets of marginal multinomial laws are the same. As such, we can conceive of a generalized Hamming distance measure for the G groups as

$$\sum_{1\leq g\neq g'\leq G}\alpha_{gg'}\{2\mathcal{H}_{gg'} - \mathcal{H}_{gg} - \mathcal{H}_{g'g'}\} = \mathcal{H}_{B,\boldsymbol{\alpha}}\,, \tag{46}$$

where $\boldsymbol{\alpha} = (\alpha_{gg'}\,,\ 1 \leq g \neq g' \leq G)'$ is a nonnegative vector, and the index B stands for the 'between group' variability. This is nonnegative and is 0 only when the K marginal probability laws for each of the G populations are the same.

As we have noted in the preceding section, the K positions or genes may not be equally important for a specific genomic study or probe. Moreover, they may neither be independent nor even marginally identically distributed. Therefore, as in (37)-(38) we consider a convex combination

$$\mathcal{H}_{gg'}(\mathbf{w}) = \sum_{k=1}^{K} w_k P\{X_{gki} \neq X_{g'kj}\}\,, \quad g, g' = 1, \ldots, G\,. \tag{47}$$

As such, we extend (4.6) to a more flexible measure

$$\sum_{1\leq g<g'\leq G}\alpha_{gg'}\{2\mathcal{H}_{gg'}(\mathbf{w}) - \mathcal{H}_{gg}(\mathbf{w}) - \mathcal{H}_{g'g'}(\mathbf{w}) = \mathcal{H}_{B;\mathbf{w},\boldsymbol{\alpha}}\,, \tag{48}$$

which has the same nonnegativity property as in (4.6). For a pair (i, j) of observations $\mathbf{X}_i$ and $\mathbf{X}_j$, we define a symmetric kernel (of degree 2)

$$\phi(\mathbf{X}_i, \mathbf{X}_j) = \sum_{k=1}^{K} w_k I(X_{ki} \neq X_{kj}) \tag{49}$$

and note that $E(\phi(\mathbf{X}_{gi}, \mathbf{X}_{gj}) = \mathcal{H}_{gg}(\mathbf{w})$ and $E(\phi(\mathbf{X}_{gi}, \mathbf{X}_{g'j}) = \mathcal{H}_{gg'}(\mathbf{w})$, for $g \neq g' =$

$1, \ldots, G$. As such, if we obtain the corresponding U-statistics as

$$U_{gg} = \binom{n_g}{2}^{-1} \sum_{1 \le i < j \le n_g} \phi(\mathbf{X}_{gj}, \mathbf{X}_{gj}) = \sum_{k=1}^{K} w_k \{ \sum_{c=1}^{C} \frac{n_{gkc}(n_g - n_{gkc})}{n_g(n_g - 1)} \}, \quad g = 1, \ldots, G; \tag{50}$$

here n_{gkc} stands for the number of sequences in the gth group for which in the kth position the observed response label is c, for $c = 1, \ldots, C$; $k = 1, \ldots, K$; $g = 1, \ldots, K$. Also,

$$U_{gg'} = (n_g n_{g'})^{-1} \sum_{i=1}^{n_g} \sum_{j=1}^{n_{g'}} \phi(\mathbf{X}_{gi}, \mathbf{X}_{g'j}) = \sum_{k=1}^{K} w_k \{ \sum_{c=1}^{C} \frac{n_{gkc}(n_{g'} - n_{g'kc})}{n_g n_{g'}} \}, \quad g \neq g' = 1, \ldots, G. \tag{51}$$

Let $n = n_1 + \ldots + n_G$ and let U_0 be the pooled group U-statistic corresponding to the same kernel. Then, we have after some routine computations,

$$\begin{aligned} U_0 &= \binom{n}{2}^{-1} \{ \sum_{g=1}^{G} \binom{n_g}{2} U_{gg} + \sum_{1 \le g \neq g' \le G} n_g n_{g'} U_{gg'} \} \\ &= \sum_{g=1}^{G} \frac{n_g}{n} U_{gg} + \sum_{1 \le g < g' \le G} \frac{n_g n_{g'}}{n(n-1)} \{ 2U_{gg'} - U_{gg} - U_{g'g'} \} \\ &= W_n + B_n, \quad \text{say}, \end{aligned} \tag{52}$$

where B_n, the last term on the right hand side denotes the 'between group' component and the first term W_n as the 'within group' one (H. P. Pinheiro et al., 2005). Under the null hypothesis, B_n has zero expectation and it is positive under alternatives. Also, W_n has expectation equal to the average of the $\mathcal{H}_{gg}(\mathbf{w})$ which is always a positive quantity. Let then

$$\mathcal{L}_n = B_n / W_n . \tag{53}$$

Based on the above (subgroup) decomposition , we consider $\mathcal{L}_n$ as a test statistic (for testing the null hypothesis of homogeneity of the G groups against possible heterogeneity with special emphasis on their diversity). It is also possible to use B_n as a test statistic, but W_n serves as a scale factor, and hence, we prefer to use it. The critical region is to be set by the right hand tail of B_n.

The above measure, however, is an unweighted one. As in (4.8), we extend this as

$$\begin{aligned} \mathcal{L}_{n,\boldsymbol{\alpha}} &= \sum_{1 \le g < g' \le G} \alpha_{gg'} \{ 2U_{gg'} - U_{gg} - U_{g'g'} \} \\ &= \sum_{k=1}^{K} w_k \sum_{1 \le g < g' \le G} \alpha_{gg'} \{ 2U^{(k)}_{gg'} - U^{(k)}_{gg} - U^{(k)}_{g'g'} \} \\ &= \sum_{k=1}^{K} w_k T_{nk}(\boldsymbol{\alpha}), \quad \text{say}, \end{aligned} \tag{54}$$

where the $U_{gg'}^{(k)}$ are the coordinatewise sample Gini-Simpson indexes, so that $T_{nk}(\boldsymbol{\alpha})$ is a statistic based on the kth position alone, based on across the G groups of observations. Note that $\mathcal{L}_{n,\boldsymbol{\alpha}}$ is an unbiased estimator of $\mathcal{H}_{B;\mathbf{w},\boldsymbol{\alpha}}$. As such, we would use $\mathcal{L}_{n,\boldsymbol{\alpha}}$ as a test statistic for testing the null hypothesis of homogeneity of the G groups against possible heterogeneity with emphasis on their diversities.

5 $K >> n$ Distributional Asymptotics

In the preceding two sections, in a quasi-marginal approach, we have advocated a convex combination of of a set of gene-wise (weighted and centered) group divergence measures as a plausible test statistic. The crux of the problem is to find the distribution theory under $K >> n$ environment, where n could be even small, but K is typically very large. Keeping this scenario in mind we define a general statistic as

$$T_n = \sum_{k=1}^{K} w_k T_{nk} = \mathbf{w}'\mathbf{T}_n\,, \tag{55}$$

where $\mathbf{w}$ is a nonnegative K vector, and the coordinatewise statistics T_{nk} have all null expectation under H_0 and positive under alternatives. However, the T_{nk} may neither be independent nor marginally identically distributed.

If n were large and $K >> n$, one could have exploited the asymptotics for the T_{nk}, and this has been systematically studied by A. S. Pinheiro et al. (2005). But, as has been noted earlier, in the present context it might not be reasonable to assume that n is adequately large to justify this approach. Even so, the asymptotics do work out well when the null hypothesis does not hold. Under the null hypothesis, the set of (generalized) U-statistics appearing in the expression of the T_{nk} end up with a degenerate case where the first-order projections vanish and thereby create roadblocks for the elegant Hoeffding (1948) projection to pave the way for asymptotic normality. For this reason, A. S. Pinheiro et al. (2005) expressed (under H_0) the test statistic T_n in an equivalent (in distribution) form

$$T_n^* = \sum_{1\le r<s\le n} \eta_{nrs}\psi(\mathbf{X}_r^*, \mathbf{X}_s^*)\,, \tag{56}$$

where the $\mathbf{X}_r^*$, $r = 1,\ldots,n$ are i.i.d. random vectors and the (symmetric and square integrable) kernel $\psi(\cdot)$ is first-order stationary in the sense that

$$E\{\psi(\mathbf{X}_r^*, \mathbf{X}_s^*)|\ \mathbf{X}_r^*\} = 0 \quad (a.e.)\,, \tag{57}$$

so that the different terms in (5.2) are uncorrelated. The η_{nrs} are nonstochastic constants for which $\sum_{1\le r<s\le n}\eta_{nrs} = 0$ and without any loss of generality, we may set that $\sum_{1\le r<s\le n}\eta_{nrs}^2 = 1$. Having observed this quasi-U-statistics structure, they exploited a martingale-array characterization that provides an easy way to the asymptotic normality result (under H_0). Further, in the $K >> n$ environment the rate of this convergence is $O(n\sqrt{K})$ instead of the usual $O(\sqrt{nK})$.

Based on this observation, it is intuitive that when n is small but K is very large, a similar asymptotic normality result should hold, albeit the rate of convergence would

probably be $O(\sqrt{K})$. Indeed, if the genes are statistically independent this result would follow directly from the classical central limit theorems for triangular arrays of centered random variables. This result to extended to a class of dependent sequences without imposing weak stationarity conditions. This large K small n scenario is appraised in specific situations in Sen et al. (2005) and A. S. Pinheiro et al. (2005) with specific emphasis on generalized U-statistics (see also Schaid et al., 2005) where n is allowed to be large). Here, we consider a general approach that addresses the asymptotics in a more general setup.

We let $Z_{nK,k} = w_k T_{nk}$, $k = 1, \ldots, K$. Our T_n^* is then expressible as $\sum_{k=1}^K Z_{nK,k}$. Note that the Z_{nk} are gene specific statistics based on the entire set of n observations, so that they are not generally independent. Moreover, their distributions depend on n as well as the K-dimensional joint laws for each of the G subgroup populations. Under the null hypothesis, these distributions are all the same, and hence the dependence is through $n_1, \ldots, n_G$ and the common multidimensional categorical data model probability law that pertain to them. Whenever K is large (but n is fixed), we can regard some weak dependence pattern underlying them, and in particular, we can consider the usual ($\phi, \psi, *-$, or regular) mixing conditions under which central limit theorems apply to an array of such weakly dependent variables. If we assume that the individual $Z_{nK,k}$ have finite second order moments, they satisfy an appropriate mixing condition, and in addition,

$$\sum_{k=1}^K w_k^2 = O(K)\,; \qquad \max\{w_k^2/(\sum_{i=1}^K w_i^2 : 1 \le i \le K\} \to 0\,, \quad \text{as } K \to \infty\,, \tag{58}$$

then we can apply such asymptotic results. We refer to Yoshihara (1993) where general asymptotic normality results for weakly dependent sequences under diverse mixing-conditions have been considered in detail; the only difference being the asymptotics in n are to be replaced by the asymptotics in K. In a somewhat different context (empirical Bayes methodology), the condition (32) is also justified in two recent papers (Qui, Brooks, et al., 2005; Qui, Klebanov, and Yakovlev, 2005), although in the present non-Bayesian setup, stationarity is not imposed in a prior distributional form. Although, there has been some work on high dependence between gene expressions, in the setup of our Section 3, with the classification of genes as DG and NDG, a high correlation may only be expected among the disease genes, so that if the number of DG's is small compared to K, then as has been explained in Section 3, such high correlation does not pose any threat to the assumption (32).

With respect to the MANOVA model in Section 2, we need in this context finiteness of moments of order 4; for the Gini mean difference based statistics, second moment condition would suffice. For the Hamming distance based statistics, we have bounded kernel and hence moments of all finite order exist. In that case, the convergence to normality under appropriate mixing-conditions is even expected to be faster.

6 Concluding Remarks

The main motivation of this study stems from the urge to exploit the subgroup decomposability of multi-group high dimensional low sample size data models and advocate

an approach that allows for the inter-position stochastic dependence in some plausible manner. This approach has some genuine utility in genomics data modelling and analysis. Further, keeping robustness perspectives in mind, flexibility of choice of a suitable statistic that is robust in a well defined manner is retained. A greater complexity arises for qualitative data models in the $K >> n$ setup and in that respect, Hamming-distance type measures have been observed to be very useful.

Such methodological studies, albeit very useful, need to be supported by extensive simulation studies to justify the adequacy of the contemplated asymptotics. In some cases, when n is large, but K is even larger, a conventional (Hoeffding, 1948) decomposition is useful in this respect. However, it should be noted that the residual term in this decomposition (allowing only the first-order projection) is $O_p(n^{-1})$ while the order of convergence to asymptotic normality is typically $O(n\sqrt{K})$. As such, the contribution of the residual term when standardized would be $O(\sqrt{K})$ and would not be negligible. To overcome this difficulty, A. S. Pinheiro et al. (2005) have considered a martingale array characterization (under the null hypothesis) that clearly demonstrates how the asymptotics could be worked out without encountering this impasse. In simulation studies, it is therefore imperative to appraise such order of convergence first; otherwise, the simulated variance factor could be considerably positively biased, resulting in some conservativeness of statistical conclusions to be drawn. Robust simulation methods in genomics are therefore very essential in genomic studies.

Acknowledgements

The author is grateful to the reviewers for helpful comments and for drawing attention to two very recently published articles on mixing conditions. The research was supported by the Boshamer Foundation, University of North Carolina, Chapel Hill.

References

Chakraborty, R., and Rao, C. R. (1991). Measurement of genetic variation in evolutionary studies. In C. R. Rao and R. Chakraborty (Eds.), *Statistical Methods in Biological and Medical Sciences. Handbook of Statistics* (Vol. 8, p. 271-316). Amsterdam: Elsevier.

Coffin, J. M. (1986). Genetic variation in AIDS viruses. *Cell*, *46*(3), 1-4.

Gini, C. W. (1912). Variabilita e mutabilita. *Studi Economico-Giuridici della R. Universita de Cagliary*, *46*(2), 3-159.

Hahn, B. H., Shaw, G. M., Taylor, M. E., Redfield, R. R., and Markham, P. (1986). Genetic variation in HTLV-III / LAV over time in patients with AIDS. *Science*, *232*, 548-1553.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, *19*, 293-325.

Pinheiro, A. S., Sen, P. K., and Pinheiro, H. P. (2005). Decomposability of high-dimensional diversity measures: quasi U-statistics, martingales, and nonstandard asymptotics. *Submitted for publication*.

Pinheiro, H. P., Pinheiro, A. S., and Sen, P. K. (2005). Comparison of genomic sequences using the hamming distance. *Journal of Statistical Planning and Inference*, *130*, 225-239.

Qui, X., Brooks, A., Klebanov, L., and Yakovlev, A. (2005). The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics*, *6*, 120.

Qui, X., Klebanov, L., and Yakovlev, A. (2005). Correlations between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology*, *4*, 34.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics*, *24*, 220-238.

Schaid, D. J., McDonnell, S. K., Hebbring, S. J., Cunningham, J. M., and Thibodeau, S. N. (2005). Nonparametric tests of association of multiple genes with human disease. *American Journal of Human genetics*, *76*, 789-793.

Sebastiani, P., Gussoni, E., Kohane, I. S., and Ramoni, M. F. (2003). Statistical challenges in functional genomics. *Statistical Science*, *18*, 33-70.

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

Sen, P. K., Tsai, M.-T., and Jou, Y.-S. (2005). High-dimension low sample size perspectives in constrained statistical inference: The SARSCoV RNA genome in illustration. *Submitted for publication*.

Silvapulle, M. J., and Sen, P. K. (2004). *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*. New York: Wiley.

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

Tsai, M.-T., and Sen, P. K. (2005). Asymptotically optimal tests for parametric functions against ordered functional alternatives. *Journal of Multivariate Analysis*, *95*, 37-49.

Yoshihara, K. I. (1993). *Weakly Dependent Stochastic Sequences and Their Applications: Order Statistics Based on Weakly Dependent Data* (Vol. III). Tokyo: Sanseido.

Author's address:

Pranab Kumar Sen
Department of Biostatistics
University of North Carolina
School of Public Health, CB #7420
3105E McGavran-Greenberg Hall
Chapel Hill, NC 27599-7420
USA
E-mail: pksen@bios.inc.edu
Fax +919-966-3804