

# Chapter 1

## High-Dimensional Discrete Statistical Models: UIP, MCP and CSI in Perspectives

Pranab Kumar Sen

*Departments of Biostatistics, and Statistics & Operations Research, University of North Carolina at Chapel Hill, NC 27599-7420, USA*

*E-mail: pksen@bios.unc.edu*

The ongoing informatics evolution has posed some challenging statistical problems. Most statistical models arising in applications in bioinformatics, data mining and a variety of other computer-intensive interdisciplinary research fields are complex in their designs, sampling plans, and associated probability laws. The curse of dimensionality is so overwhelming that conventional likelihood ratio based statistical inference may not be much useful. Further, such models are typically constrained by inequality, order, functional, shape or other restraints, often on functional parameters, and as a result, optimal statistical inference procedures are hard to find or may not even exist. Although, the use of restricted-, profile-, pseudo-, partial-, or quasi-likelihood has been advocated in such constrained environments, their optimality properties are generally not known, and at least, are hard to establish. S. N. Roy's ingenious union-intersection principle along with innovations in discrete multivariate analysis provide an alternative avenue, often having some computational advantages, increased scope of adaptability, and flexibility beyond conventional likelihood paradigms. This scenario is appraised here with some illustrative examples of high-dimensional discrete multivariate analysis of covariance models arising in genomic studies in parametrics as well as beyond parametrics setups.

### Contents

1.1 Introduction . . . . .	2
1.2 Preliminary Notion . . . . .	4
1.3 CATANOCOVA . . . . .	7
1.4 UIP and CSI . . . . .	9
1.5 Statistical Reasoning for HDDSM . . . . .	14
1.6 UIP and MCP in HDDSM . . . . .	17
References . . . . .	23

## 1.1. Introduction

During the last twelve years before his premature demise, Professor S. N. Roy made most significant and innovative contributions in three important areas in multivariate statistical analysis, namely, *multiple comparisons procedures* (MCP), *union-intersection principle* (UIP) and discrete multivariate analysis. During 1930 - 1960's, classical multivariate analysis stole the limelight of most innovative and sophisticated statistical analysis, albeit being mostly confined either to samples from finite-dimensional multivariate normal populations or simple and low multidimensional categorical data models. The present study reflects some of Roy's innovative perspectives in finite dimensional data models with an eye on their fruitful incorporation in high dimensional perspectives.

Statistical models usually advocated for complex problems arising in some interdisciplinary research and many real life applications are rarely simple so as to make room for routine incorporation of conventional or standard (i.e., likelihood based) statistical inference tools. The ongoing evolution of genomics (and bioinformatics, in general) has indeed posed some enormously large dimensional statistical models where the sample size may often be relatively much smaller. Such high-dimensional low sample size (HDLSS) models generally involve complex designs, sampling plans, and the underlying stochastics relate to probability laws which, typically, not only involve a multitude of parameters but also with the parameters subjected to various nonlinear restraints. Inequality, order, functional and shape constraints are commonly encountered, in probability as well as sample spaces, where the HDLSS perspectives may complicate the models as well as their statistical resolutions considerably. For beyond parametrics (i.e., nonparametrics and semiparametrics) setups, often, there could be more complex restraints involving functional constraints. Stochastic ordering (dominance) of functional parameters in categorical data models, arising in genomic studies, particularly, in single nucleotide proliferation (SNP) models being a notable example of this kind (Sen et al. 2007). In conventional statistical inference, likelihood, sufficiency and invariance principles play a key role in finite sample methodology, and some of the finite-sample optimality properties usually transpire in large sample cases even without sufficiency, invariance or some other regularity conditions. Nevertheless, even in such asymptotic cases, lacking support of suitable regularity assumptions, particularly in constrained environments, optimal statistical inference may encounter roadblocks of diverse types.

Generally, complex statistical models create impasses for computation of *maximum likelihood estimators* (MLE) and *likelihood ratio tests* (LRT) in closed,

explicit or manageable forms; often, this may become a formidable task. Even so, various algorithms have been developed for such computational convenience, though the finite sample optimality properties of MLE and LRT ranging over the exponential family of densities may not automatically transpire in more complex models where the underlying probability laws are rarely bonafide members of such regular families; even if they are, underlying constraints or implicit structural restraints may take away such optimality properties. (Restricted) RMLE and (restricted) RLRT along with various modifications of the likelihood function have therefore been advocated for such complex models, albeit they may not have any universal optimality property parallel to that in simple models. S. N. Roy's (1953) ingenious UIP, having its genesis in the *likelihood principle* (LP), has emerged as a viable alternative, often having some computational advantages, increased scope of applicability (beyond the likelihood paradigm), greater adaptability to nonstandard situations (beyond the parametrics), and good robustness perspectives.

Roy, Gnanadesikan and Srivastava (1971) contains a thorough treatise of UIP in multivariate models with due emphasis on simultaneous confidence sets and multiple hypothesis testing problems relating to the domain of MCP. Though their treatise is mostly confined to continuous multinormal distributional models, the past three decades have witnessed a steady flow of research on UIP in a variety of beyond parametric models. For a general treatise of *constrained statistical inference* (CSI) we refer to Silvapulle and Sen (2004) which recaptured the prior developments in in the classical monograph of Barlow et al. (1972) and its follow-up by Robertson et al. (1988). The major emphasis in Barlow et al. (1972) has been the finite sample methodology with due consideration of the basic role of the likelihood function in such formulations. More in-depth computational aspects are additionally reported in Robertson et al. (1988). The Silvapulle-Sen (2004) treatise goes beyond that into more general setups with adequate asymptotics to simplify the methodology; in line with the Wald-type tests, in CSI such procedures are elaborated along with the basic role of UIP in some important problems. The present study is devoted to a display of the basic role of UIP in *high-dimensional discrete statistical models* (HDDSM) with due emphasis on CSI as well as MCP; this development being very useful in the evolving field of genomics or bioinformatics. A key factor in this respect is the most innovative treatise of some multidimensional categorical data models in Roy (1957) and its subsequent ramification during the past 50 years. Functional parameters arising in HDDSM, as is commonly perceived in genomics studies (Tsai and Sen 2005, Sen et al. 2007), is a basic perspective deserving thorough appraisals. In Section 2, we outline the preliminary notion. In Section 3 an overview of classical categorical multivariate analysis of covariance (CATMANOCOVA) in low to moderate di-

mensional setups (where the sample size  $n$  is still  $\gg K$ , the dimension) is made. In Section 4, the basic formulation of UIP is considered along with some illustrations in some simple models. Section 5 is devoted to HDDSM with due emphasis on some SNP models. Section 6 deals with fruitful incorporation of the Roy (1953) UIP and MCP in such HDDSM's with due emphasis on some applications in genomic models (Sen et al. 2007).

## 1.2. Preliminary Notion

The scenario of statistical modeling and inference changes drastically from parametric to beyond parametric perspectives, and even in the parametric setup, from the single parameter to multiparameter models. The more complex a model is, the more likely is the feature that optimal statistical inference may not exist, and even it exists, it may be harder to implement. The evolving field of genomics is a pertinent illustration of the enormous difficulties which conventional statistical inference tools are encountering in this high-dimensional low-sample size (HDLSS) setups. For (continuous, discrete, as well as, categorical ) distributions belonging to the so called exponential family, generally optimal statistical inference procedures, based on the sufficiency principle, work out well. However, even for the exponential family, if there are too many parameters or if the parameters are constrained in some way, such optimality properties may not transpire. Often invariance structures (with respect to some group of transformations mapping the sample space onto itself, inducing conjugate transformations on the parameter space) allow us to formulate invariant statistical procedures where within the class of such invariant procedures, an optimal one could be found out in a rational way. Sans such exponential families of densities, usually, finite-sample optimal statistical inference procedure may not exist, although, in the conventional asymptotic setup :  $K \ll n$  (generally,  $K$  being fixed but  $n$  is large), asymptotically optimal statistical inference procedures have been prescribed. Whereas in the univariate case, in many situations, a moderately large sample size provides adequate asymptotic theory based methodology for inference tools, as the dimension becomes large, a reasonably good asymptotic approximation may generally require a much larger sample size; the rate of increase of the required sample size being usually much faster than the dimension increase. This basic limitation stands in the way of valid and efficient statistical inference for HDLSS data models.

As an illustration, consider the classical multivariate analysis of variance

(MANOVA) problem in the setup of (multi-)normally distributed errors. Let

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}, \quad (1.1)$$

where the observable stochastic matrix  $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)'$ , with each  $\mathbf{Y}_i$  being a stochastic  $p$ -vector, is related to a known (nonstochastic) matrix  $\mathbf{X}$  (of order  $n \times m$ ) of regression constants, through an unknown (regression) parametric matrix  $\boldsymbol{\beta}$  (of order  $m \times p$ ), and where  $\mathbf{E}' = (\mathbf{e}_1, \dots, \mathbf{e}_n)'$  with each ( $p$ -vector)  $\mathbf{e}_i$  having a multivariate normal distribution with null mean vector and a positive definite (p.d.) but unknown dispersion matrix  $\Sigma$ . Symbolically, we write

$$\mathbf{E} \sim \text{MN}(\mathbf{0}, \mathbf{I}_n \otimes \Sigma), \quad (1.2)$$

where  $\otimes$  stands for the Kronecker product of the two matrices. In this setup, consider the most simple null hypothesis

$$H_0 : \boldsymbol{\beta} = \mathbf{0}, \quad (1.3)$$

against (global) alternatives that

$$H_1 : \boldsymbol{\beta} \neq \mathbf{0}. \quad (1.4)$$

In this normal theory setup, the MLE of  $\boldsymbol{\beta}$  is given by

$$\hat{\boldsymbol{\beta}}_n = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}). \quad (1.5)$$

The residuals are then defined by

$$\begin{aligned} \hat{\mathbf{Y}}_n &= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n \\ &= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{E}. \end{aligned} \quad (1.6)$$

The residual sum of product matrix (of order  $p \times p$ ) is defined by

$$\mathbf{S}_E = (\hat{\mathbf{Y}}_n)'(\hat{\mathbf{Y}}_n). \quad (1.7)$$

Side by side, the sum of product matrix (of order  $p \times p$ ) due to regression is defined as

$$\mathbf{S}_H = (\hat{\boldsymbol{\beta}}_n)'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}}_n) \quad (1.8)$$

The rank of  $\mathbf{S}_E$  (and  $\mathbf{S}_H$ ) is equal to  $\min(p, n - m)$  (and  $\min(m, p)$ ). Thus, in order that  $\mathbf{S}_E$  is of full rank, it is tacitly assumed that  $n - m \geq p$ , while the other matrix is nonsingular when  $m \geq p$ .

First, even for this simple model (belonging to the exponential family), if  $p$  and  $m$  are greater than one, there may not be a uniformly most powerful (UMP) test for  $H_0$  vs.  $H_1$ . As a matter of fact, since  $\Sigma$  is nuisance, we need to confine ourselves to UMP similar regions. To resolve this problem, one considers the class

of tests which are invariant under nonsingular transformations on the observation vectors:  $\mathbf{Y} \rightarrow \mathbf{Z} = \mathbf{YB}$ , for some nonsingular  $p \times p$  matrix  $\mathbf{B}$ . Within this class of invariant tests, the usual likelihood ratio test is best (i.e., uniformly most powerful invariant (UMPI)) test when the rank of  $\mathbf{S}_H$  is equal to 1; for this particular case, all the three test statistics are equivalent and enjoy the same optimality property. However, even in this case, we need the condition that  $n - m \geq p$ . For the general case where both  $m, p$  are greater than 1, there are several test statistics including the classical likelihood ratio test based on  $|\mathbf{S}_E + \mathbf{S}_H|/|\mathbf{S}_E|$ , Hotelling-Lawley trace criterion based on  $\text{Trace}(\mathbf{S}_H \mathbf{S}_E^{-1})$ , Roy's largest root criterion based on  $ch_{max}(\mathbf{S}_H (\mathbf{S}_E)^{-1})$ , and other ramifications. None of these would be UMPI. Since all these test criteria depend on the maximal invariant, the characteristic roots of  $\mathbf{S}_H (\mathbf{S}_E)^{-1}$ , it is imperative that  $n \geq m + p$ .

Instead of the simple null hypothesis  $H_0$  considered above, one may consider a more general null hypothesis where for suitable prespecified matrices  $\mathbf{A}$  and  $\mathbf{C}$  of order  $r \times m$  and  $p \times q$  respectively with  $r \leq m$  and  $q \leq p$ , we set  $H_0 : \mathbf{A}\beta\mathbf{C} = \mathbf{0}$  and the alternatives relate to nonnull matrices on the right hand side. For simultaneous (or multiple) hypotheses testing, we allow  $\mathbf{A}$  to be arbitrary within a class, say  $\mathcal{A}$ , and also  $\mathbf{C}$  arbitrary within another class, say  $\mathcal{C}$ , and desire to control the type I error over these classes as a whole. Similarly, we may want to set a simultaneous confidence set for all  $\mathbf{A}\beta\mathbf{C}$ , allowing  $\mathbf{A} \in \mathcal{A}$  and  $\mathbf{C} \in \mathcal{C}$ , with the property that the coverage probability is some specified  $1 - \alpha$ , for some  $\alpha \in (0, 1)$ . The genesis of UIP lies in this complex, and this will be elaborated in Section 1.4.

Like the univariate one sided alternatives, we could have in the multivariate case some one sided alternatives, or more generally, alternatives specified by some inequality, order or other constraints. CSI typically pertains to such complex statistical environments. The affine invariance structure mentioned earlier may not pertain to such restricted alternatives, and hence, the classical (unrestricted) likelihood based tests sketched above may not be ideal in such CSI problems. Much of the development in CSI rests on RMLE and RLRT, which takes into account the underlying restraint(s), although not much optimality property may be retained in this setup. There could be computational complexities too. Therefore, it may be natural to appraise the interactive role of UIP and CSI. This will be considered in Section 1.4.

The above discussion pertains specifically to multinormal distributions, In many fields of application, such an assumption may be very untenable. In some cases, the random vectors are continuous, and hence, nonparametric models are more reasonable to adopt. In some other cases, we may be confronted with count variables while in many other cases we have categorical (and possibly qualitative) data models. The likelihood principal (LP) may generally encounter roadblocks

in such general setups. In the next section, we consider simple multidimensional categorical data models, albeit in the conventional asymptotic setups, to illustrate some of these basic difficulties with the LP. In Section 1.5, we shall introduce the HDLSS discrete multivariate setups, and appraise these models in the light of MCP and CSI perspectives.

### 1.3. CATANOCOVA

Categorical ANOVA models typically involve product-multinomial distributions which can be presented in their simplest form as follows. Consider  $G$  independent populations, each involving a set of  $C$  categorical responses not necessarily quantitative or even ordered in some way, and let  $\pi_g(c)$ ,  $c = 1, \dots, C$  be the cell probabilities for the  $g$ th population, for  $g = 1, \dots, G$ . Let there be  $n_g$  observations from the  $g$ th population and let  $n_{gc}$  be the cell frequency of the  $c$ th cell, for  $c = 1, \dots, C$ , and  $g = 1, \dots, G$ , all these samples being drawn independently. The joint probability function of these  $n_{gc}$ , known as the product multinomial law, is given by

$$\prod_{g=1}^G \frac{n_g!}{n_{g1}! \cdots n_{gC}!} \prod_{c=1}^C \{\pi_g(c)\}^{n_{gc}}, \quad (1.9)$$

where  $n_{gc}$  are nonnegative integers such that  $\sum_{c=1}^C n_{gc} = n_g$ , and the  $\pi_g = (\pi_g(1), \dots, \pi_g(C))'$ ,  $g = 1, \dots, G$  all belong to the simplex  $\mathcal{S}_{C-1} = \{\mathbf{x} \in [0, 1]^C : \mathbf{x}'\mathbf{1} = 1\}$ . In this nonparametric formulation, an ANOVA model relates to the homogeneity of the  $\pi_g$ . In many cases, we may be interested in various MHT testing problems, possibly in CSI setups, for this CATANOVA model. As an illustration, first, consider the (opinion) response model for

#### *Reduction of US Military Involvement in Iraq*

where we have the following (ordered) response categories: HS = highly supportive, S = generally supportive, N = No opinion, O = opposed and TO = totally opposed. Note that despite an inherent ordering, there is no linear or precise mathematical scale for the ordering. The probability space is the simplex  $S_4$ , generated by the vector of the 5 cell probabilities. Let us now consider the following classification of respondents: Democratic and Republican. The probability vector for the two groups are denoted by  $\pi_D$  and  $\pi_R$  respectively, each belonging to the common simplex  $S_4$ . We frame the null hypothesis of homogeneity as

$$H_0 : \pi_D = \pi_R = \pi \text{ (unknown).}$$

We denote the corresponding cumulative probability vectors by  $\Pi_D$  and  $\Pi_R$  respectively (where the last element in each vector is equal to 1). To reflect a one-sided preference of the Democrats to the Republicans, we consider the following restricted alternative

$$H_1^< : \Pi_D \geq \Pi_R$$

with at least one (of the four coordinates) bearing a strict inequality sign. The conventional  $2 \times 5$  contingency table based  $\chi^2$ -test or even the Fisher exact randomisation test may not be ideal, as they address global alternatives of lack of homogeneity and thereby are more likely to be less powerful for such restricted alternatives. This one-sided hypothesis testing in a multiparameter setup could be even more complicated if there are some covariates or explanatory variables (such as male / female and young, middle-age and senior people, type of employment, educational and racial diversity etc). Some of these problems are discussed in Silvapulle and Sen (2004, Ch.6) in detail.

Suppose that instead of one such query, we have a questionnaire involving  $K$  basic questions, for each of which, we have a set of  $C$  ordered categories of responses. Thus there will be a totality of  $C^K$  possible response category-combinations. The  $K$  questions could be related (as in the classical *item analysis* schemes) when typically they refer to different aspects of a composite health or psychiatric problem. As such, the  $K$  response vectors are expected to be stochastically dependent. Although some restricted alternative hypothesis testing problems may relate to the  $K$  marginal probability laws, without taking into account their interdependence, such hypothesis testing problems can not be treated efficiently. In the context we are more interested the categories may not have any partial ordering and hence, hypotheses are to be formulated in a somewhat different way.

In a parametric formulation, the  $\pi_g(c)$  are expressed in terms of some unknown parameters (vectors)  $\theta_g$  of dimension less than  $C$  (and possibly involving some explanatory or concomitant variates), and as in the classical ANOVA problem, we may like to test for the homogeneity of these parametric vectors, as well as, associated MHT along with CSI formulations. In passing, we may remark that if the response is quantal (i.e., all or nothing) where  $C = 2$ , regression on the explanatory variables may not be in conventional linear models, and a logit or probit transformation is advocated for using generalized linear models for drawing statistical conclusions. On the other hand, if  $C \geq 3$  and the response categories are qualitative, such transformations are usable, and more general regression models are to be sought. Due to such model complexities, the finite sample treatment

of the nonparametric CATANOVA model may not generally be tenable in such parametric formulations. Usually, BAN (best asymptotically normal) estimators are incorporated in an asymptotic setup wherein the  $n_g$  are all taken to be large. Wald-type tests are more commonly used in this large sample size (relative to the dimension) context, including CSI setups; we again refer to Silvapulle and Sen (2004, Ch. 6). The UIP based approach will be outlined in later sections. Our main contention is to appraise HDDSM problems arising in this context, and this will be done in a later section.

#### 1.4. UIP and CSI

S.N. Roy (1953) motivated the UIP through multivariate models with due emphasis on multiple comparisons and simultaneous statistical inference. We illustrate UIP with a general composite hypothesis testing problem that lends itself naturally to CSI as well as HDDSM, typically involving multiparameter models. Consider a general hypothesis testing problem, not necessarily the multinormal or multinomial models treated in earlier sections, or even a parametric model. Let  $H_0$  be the null hypothesis of interest and let  $H_1$  be the alternative one; both of them are composite so that the likelihood function is not completely specified under either of them. As it is the case with composite hypotheses testing problems, there may not be in general an optimal test for testing  $H_0$  vs.  $H_1$ , and in many case, even finding out a similar region may restrict attention to a subclass of tests like invariant tests, conditional tests, etc.. This situation is likely to be worse in CSI where the conceived restraints may preempt the relevance of invariant or conditional tests. However, for a general class of testing problems, including in CSI, it might be possible to express

$$H_0 = \cap_{j \in \mathcal{J}} H_{0j}, H_1 = \cup_{j \in \mathcal{J}} H_{1j} \quad (1.10)$$

where  $\mathcal{J}$  is a suitable index set, and for each  $j \in \mathcal{J}$ , there exists a suitable (and often optimal in a certain sense) test for testing  $H_{0j}$  vs  $H_{1j}$ . In a parametric framework, such a test could be the UMP test whenever the latter exists, could be LMP (locally most-powerful) test in some other case, and in beyond parametrics setups, such a test can be decided on the basis of robustness, validity and efficiency considerations. Further, the index set  $\mathcal{J}$  can be a finite (discrete) set, or it may even be a set in continuum. In this way, there is flexibility in the decomposition of the hypotheses and choice of appropriate test statistics. Bearing in mind the genesis of UIP in LP (Roy, 1953), we consider first the following illustrative example pertaining to multinormal populations where the connection of UIP and LP can be

identified easily.

Let  $\mathbf{X}_1, \dots, \mathbf{X}_n$  be  $n$  independent and identically distributed random vectors having a  $p$ -variate (multi-)normal distribution with unknown mean vector  $\mu$  and dispersion matrix  $\Sigma$ , unknown but positive definite (p.d.). Consider first the null hypothesis  $H_0 : \mu = \mathbf{0}$  versus  $H_1 : \mu \neq \mathbf{0}$ , treating  $\Sigma$  as a nuisance parameter (matrix). There is no UMP test for this hypotheses testing problem. The likelihood ratio test statistic for this problem is a monotone function of the Hotelling  $T^2$ -statistic

$$T^2 = n(\bar{\mathbf{X}}_n)' \mathbf{S}_n^{-1} (\bar{\mathbf{X}}_n), \tag{1.11}$$

where

$$\bar{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i, \mathbf{S}_n = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}}_n)(\mathbf{X}_i - \bar{\mathbf{X}}_n)'. \tag{1.12}$$

The test is affine-invariant and within the class of affine-invariant tests it is UMP. The affine invariance is defined in terms of the invariance of the test under the class of transformations  $\mathbf{X} \rightarrow \mathbf{Y} = \mathbf{B}\mathbf{X}$ ,  $\mathbf{B}$  being nonsingular. Let us look into this picture from a different angle.

Let  $\mathbf{a} \in \mathbf{R}^p$  be an arbitrary  $p$ -vector, and let  $H_{0\mathbf{a}} : \mathbf{a}'\mu = 0$  and  $H_{1\mathbf{a}} : \mathbf{a}'\mu > 0$ . Then note that

$$H_0 = \cap_{\mathbf{a} \in \mathbf{R}^p} H_{0\mathbf{a}}, H_1 = \cup_{\mathbf{a} \in \mathbf{R}^p} H_{1\mathbf{a}}. \tag{1.13}$$

Further note that for testing the null hypothesis  $H_{0\mathbf{a}}$  against  $H_{1\mathbf{a}}$ , a UMP test is based on the Student  $t$ -statistic

$$t(\mathbf{a}) = \sqrt{n}(\mathbf{a}'\bar{\mathbf{X}}_n)/(\mathbf{a}'\mathbf{S}_n\mathbf{a})^{1/2} \tag{1.14}$$

using the right hand side critical region marked by the Student  $t$ -distribution with  $n - 1$  degrees of freedom (DF). The overall hypothesis  $H_0$  is only accepted when all the component hypotheses are accepted and  $H_1$  is accepted when at least one component alternative hypothesis is deemed acceptable, thus implementing the UIP. Therefore, the UIT is based on the test statistic

$$t^* = \sup\{t(\mathbf{a}) : \mathbf{a} \in \mathbf{R}^p\} \tag{1.15}$$

and some routine computations yield that

$$(t^*)^2 = T_n^2. \tag{1.16}$$

Thus, the UIT and LRT are isomorphic for this hypotheses testing problem. If, however, we go to the  $K$ -sample problem of testing the homogeneity of the mean vectors for  $K \geq 3$  or the general MANOVA problem, the UIT and LRT statistics may not be generally isomorphic, albeit both are affine-invariant (and none being UMP invariant). The UIT relates to Roy's largest root criterion.

Let us consider the same one-sample model under a CSI setup. Namely, we consider the same null hypothesis that  $\mu = \mathbf{0}$  against one-sided alternative  $H_1^> : \mu \geq \mathbf{0}$ , i.e., the mean vector lies in the positive orthant  $\mathbf{R}^{p+} = \{\mathbf{x} \in \mathbf{R}^p : \mathbf{x} \geq \mathbf{0}\}$ . The RLRT for this problem, treated by Perlman (1969) and others, rests on the test statistic

$$\lambda_n = \log L_n(\hat{\theta}_n^{(1)}) - \log L_n(\hat{\theta}_n^{(0)}), \quad (1.17)$$

where  $\hat{\theta}_n^{(0)}$  is the RMLE of  $\theta = (\mu, \Sigma)$  under the null hypothesis (which is simple), and  $\hat{\theta}_n^{(1)}$  under the alternative  $H_1^>$ , and this is complex. By the use of the KTL-point formula, one can obtain an expression of the estimates, though not very simple. Further, the estimator of  $\Sigma$  (under the alternative) is not orthogonal to the RMLE of  $\mu$  and it does not have the anticipated Wishart distribution. To eliminate this problem, Perlman (1969) suggested a conservative test by allowing a further maximisation over all  $\Sigma$  in the class of p.d. matrices.

The UIT can be constructed as follows: The alternative hypothesis can be equivalently expressed as  $H_1 = \cup\{\mathbf{a}'\mu > 0, \forall \mathbf{a} \geq \mathbf{0}\} = \cup_{\mathbf{a} \in \mathbf{R}^{p+}} H_{1(\mathbf{a})}$ . Then, for testing  $H_{0\mathbf{a}}$  versus  $H_{1\mathbf{a}}$ , we have an optimal one-sided test based on  $t_n(\mathbf{a})$ , as defined above. Therefore, the UIT test statistic is given by

$$t_n^+ = \sup\{t_n(\mathbf{a}) : \mathbf{a} \in \mathbf{R}^{p+}\}, \quad (1.18)$$

an explicit expression for this UIT test statistic, also based on the KTL-point formula theorem, comparable to the RLRT test statistic is available in the literature (Silvapulle and Sen 2004, Ch. 5). We also refer to Sen and Tsai (1999) for a detailed treatise of the LRT and UIT for one-sided alternatives for multivariate normal mean testing problem when the dispersion matrix is nuisance (i.e., positive definite but arbitrary). It is shown that the UIT statistic can not be smaller than the RLRT statistic, and both of them have a certain amount of conservativeness due to the nuisance dispersion matrix. They suggested a two-stage LRT and UIT to overcome this problem. If we look into the corresponding simultaneous confidence sets, the procedures based on the LP and UIP have possibly different forms, and it is known (Wijsmann, 1979) that the UIP has certain advantages over the LP in this setup.

The UIP has also received due attention in other parametric CSI problems along with new interpretations and motivations. While a comprehensive account of some of these developments is available in the literature (viz., Silvapulle and Sen, 2004), it should be noted that in specific CSI problems, often these UIP motivated procedures have much simplicity and rationality to offer; we refer to Mudholkar and McDermott (1989), McDermott and Mudholkar (1993), Mudholkar et al. (1993,1995), and Srivastava and Mudholkar (2001) and Mudholkar

et al. (2001) for some interesting developments including some robust tests for the multivariate orthant restricted alternative testing problem. Tsai (1995) considered the estimation of covariance matrices under Löwner order restriction, and Tsai (2004) considered the covariance matrices estimation problem under simple tree ordering. Das and Sen (1994) considered the restricted canonical correlation inference problem. In the context of genomics, Sen et al. (2007) have developed restricted alternative tests based on Hamming distance, and have discussed their plausibility for the high-dimension low sample size contexts. Also, nonparametric tests for ordered diversity in a genomic sequence have been considered by Sen (2005). Although, these developments relate to some specific distributional setups or pseudo-likelihood problems, in principle, that goes over to the more general case of densities not belonging to the so called exponential family, as well as, to beyond parametrics. However, computational and distributional complexities may mar the simple appeal of the UIP to a certain extent.

In a parametric setup with quantitative multifactor multiresponse experiments, a detailed account of UIP is available in Roy et al. (1971). This interesting monograph, a culmination of the fundamental ideas of Roy (1953, 1957), has clearly illustrated the basic appeal of UIP in various statistical inference problems, albeit in traditional setups without so much emphasis on CSI. The evolution of CSI during the past four decades has added more impetus to look into UIP in CSI for parametric as well as beyond parametric setups. Towards this end, we consider first the same one-sample multivariate problem as in above but without assuming that the underlying distribution ( $F$ ) is multinormal or of some specified form. Simply assume that  $F$  is diagonally symmetric about its location  $\mu$  (in the sense that both  $\mathbf{X} - \mu$  and  $\mu - \mathbf{X}$  have the same distribution). Consider the hypotheses testing problem:  $H_0 : \mu = \mathbf{0}$  vs.  $H_1^+ : \mu \geq \mathbf{0}$ . If we write  $H_{0j} : \mu_j = 0, j = 1, \dots, p$  and  $H_{1j} : \mu_j \geq 0, j = 1, \dots, p$ , then  $H_0$  and  $H_1^+$  can be expressed as the (finite) intersection and union of the  $H_{0j}$  and  $H_{1j}$  respectively. This is therefore a finite union-intersection formulation ( see J. Roy 1958) for a step-down procedure) that makes sense since we are not seeking for affine or similar invariance in our resolution. For the  $j$ th marginal, we have a univariate testing problem for which (locally or globally) optimal or at least desirable signed rank tests are known to exist. The crux of the problem is however to find the distribution theory for the maximum of these  $p$  possibly correlated statistics. Unfortunately, this distribution depends on the unknown  $F$ , even under the null hypothesis. An easy way to eliminate this impasse is to take recourse to the permutation distribution theory generated by the  $n!2^n$  conditionally equally likely sign-inversions and column permutations, a detailed treatise of this being available in Sen and Puri (1967). Silvapulle and Sen (2004, Sec.5.5) have considered some parallel UIT statistics based on derived

$R$ -estimators of location, and have shown that in an asymptotic setup the multi-normal setup is retained in this formulation as well.

In the context of multiparameter/multivariate hypothesis testing problems, restricted alternatives, often, there is no UMP test, even along specific directions. This feature complicates the construction of usual LRT and RLRT. In a majority of cases, a locally most powerful (LMP) test can be constructed for specific directions, and thus, a UIT based on such LMP statistics can be constructed. The concept extends easily for nonparametric tests where LMPR tests are advocated. Some of these procedures have been discussed in detail in Silvapulle and Sen (2004, Ch. 5). As such, we omit most of these discussions here.

The extension of statistical reasoning from simple parametrics to more complex beyond parametrics setups has been fortified with less emphasis on the likelihood and more on asymptotics to accommodate workable resolutions. This is no exception in CSI too, and the where UIP plays a special role. In most of the complex statistical inference problems, the usual likelihood formulation stumbles into methodological as well as computational difficulties, even in asymptotic setups. For example, in semiparametric inference, the celebrated Cox (1972) proportional hazards model, a partial likelihood approach was innovated to accommodate censoring in a meaningful way, and suitable counting processes along with martingale methodology provided the needed methodological support. Yet, the very basic assumption of proportional hazards may often appear to be rather untenable. More general semiparametrics may require further modification of the LP, and along this line, profile-, partial-, penalized-, pseudo-, and quasi-likelihoods have been advocated in the literature. The usual estimating equations appearing in likelihood formulations have been extended to "generalized estimating equation" (GEE), hence linking the generalised linear models in this broader setup. Empirical likelihoods have also been advocated, along the lines of conventional resampling methods. In all these developments, the validity of exact statistical inference is questioned, loss of information is assessed, and robustness aspects have been focused.

As we look into CSI in such a complex setup, we observe that statistical inference perspectives become even more unclear. For example, in many such problems, suitable scores statistics based on appropriate modification of the likelihood formulation are used in the formulation of test statistics or estimating equations. In constrained environments, these score statistics often require considerable modifications to satisfy the set constraints, and thus resulting in a different distributional problem. Silvapulle (1995) and Silvapulle and Silvapulle (1995) formulated a somewhat different approach, termed the Wald-type tests. Recall that in a conventional regular model, Wald (1943) considered a modification of the likelihood ratio test by considering the unrestricted MLE of the associated parameters and

exploiting their asymptotic normality in a formulation of a quadratic form which is asymptotically equivalent to the LRT for the same hypothesis testing problem. This approach has been systematically explored in Silvapulle and Sen (2004) covering some parametrics as well as beyond parametrics CSI problems. This approach has also been extended to beyond parametrics situations. The only point in this approach is the need for the computation of the restricted estimators as well as the unrestricted ones, even in an asymptotic setup and that in general requires extensive computational tools. It has been observed (Silvapulle and Sen 2004) that in some simple CSI problems it might be more convenient to incorporate the UIP to derive parallel inference tools which may be computationally less cumbersome and yet asymptotically equivalent. In the rest of this study, we shall elaborate this feature of UIP with some specific CSI problems.

### 1.5. Statistical Reasoning for HDDSM

Low dimensional discrete statistical models in general CSI setups have been treated in Silvapulle and Sen (2004), Section 6.5 (pp.306-313) in a general case of  $r \times c$  contingency tables, for  $r, c \geq 2$ . In the LP based formulation (viz., Dardanoni and Forcina, 1998), it is necessary to find the MLE of  $\pi_D$  and  $\pi_R$  under the null as well as alternative hypotheses. The computation of the MLE under  $H_0$  is simple, namely, the pooled group marginal proportions in the  $r$  or  $c$  categories. However, analytical computation of the MLE under general restricted alternatives may be usually quite cumbersome, requiring extensive computational algorithms. Further, once these are done, one has to appeal to the large sample distribution theory of RLRT as customarily given by the conventional chi-square bar distribution. There is a further complication due to the nature of the dispersion matrix (unknown and not of full rank), and hence, as in Perlman's (1969) multinormal mean testing problem against positive orthant alternatives (with unknown arbitrary p.d. dispersion matrix), one has to deal with the least favorable configuration to obtain a conservative  $p$ -value. Side by side, let us consider the UIP approach. We need to find out the (unrestricted) UMLE of the two probability vectors, and the task is comparatively simpler, specially in the same asymptotic setup. We may motivate the approach through the asymptotic joint normality of these UMLE, again a well known result. Once this is done, we have a set of inequality constraints for which the classical Kühn- Tücker- Lagrange (KTL) point formula (viz., Silvapulle and Sen, 2004, pp. 166 - 168), under the Shapiro (2000) regularity conditions, could be used to formulate the appropriate test statistic. This may generally require less intensive computational schemes. Its asymptotic null hypothesis distribution is given by a chi square bar distribution, a convex mixture of chi-square distribu-

tions with degrees of freedom ranging from 0 to  $p$  (here  $p = 4$ ), similar to the case treated in Silvapulle and Sen (2004, p.157). Finally, we note that under the null hypothesis, we have the homogeneity of the two probability vectors, and hence, the exact conditional probability law (given the marginal totals) can be effectively used to find a conditional test. For sample sizes not too large, this provides a better control of the type I error than a conservative testing procedure based on the asymptotics solely. For some details, we may refer to Tsai and Sen (2005) where a more general case of restricted alternative hypothesis testing problem has been treated under the Shapiro (2000) regularity conditions, exploiting the UIP there to a greater extent. Whereas the LP based approach exploits the Wald formulation, the UIP based approach does so through the Rao score statistics, under constrained environments.

The second motivating example relates to a statistical comparison of 4 epidemiologic groups with respect to their SARSCoV genomes, treated in Sen et al. (2007). Following the origin of SARS (severe acute respiratory syndrome) in Southern China, the global epidemic resulted in 8,422 infected people with 916 deaths. The SARS causative agent was identified as a novel coronavirus (SARSCoV), as a single-stranded and positive sense RNA virus with large genome size (around 30kb). Compared to other RNA viruses, the mutation rate in the SARSCoV is moderate but still several orders of magnitude higher than in DNA virus. An appraisal of variations on the viral RNA was therefore sought for molecular, clinical and therapeutic studies. After preliminary data handling, 25 SARS genome with single nucleotide variation were identified including 6 from Beijing, 3 from Hong Kong, 4 from Singapore and 12 from Taiwan. There were 192 screened genes, so that we have 4 groups of 6, 3, 4 and 12 sequences, each with 192 positions and at each position 4 possible response: Nucleotides A, C, G and T. This perfectly fits with our contemplated HDDSM.

Motivated by the above, we consider  $G$  groups of sequences, where in the  $g$ th group, there are  $n_g$  sequences, for  $g = 1, \dots, G$ . In the  $k$ th position, let  $n_{gkc}$  denote the number of sequences in the  $g$ th group with the response category  $c$ , where  $c$  ranges over  $1, \dots, C$  and  $k$  over  $1, \dots, K$ . In the above example,  $K = 192$  and  $C = 4$ . Thus, we have  $G$  stochastic matrices  $((n_{gkc}))_{K \times C}$ , for  $g = 1, \dots, G$ . Although, it could be assumed that the sequences are independent, it is unreasonable to assume that the responses at the  $K$  positions are stochastically independent. To capture the HDDSM, we let  $\mathbf{X}_{gi} = (X_{gi,1}, \dots, X_{gi,K})'$ ,  $i = 1, \dots, n_g$ , where  $X_{gi,k}$  can take on levels  $1, \dots, C$  for each  $k = 1, \dots, K$ . Therefore, there is a set  $\mathcal{C}$  of  $C^K$  possible joint labels  $\mathbf{c} = (c_1, \dots, c_K)$ , where each  $c_k$  can take on the labels  $1, \dots, C$ , so that the cardinality of  $\mathcal{C}$  is  $C^K$ . The corresponding cell probability is denoted by  $\pi_g(\mathbf{c})$ , for  $\mathbf{c} \in \mathcal{C}$ . Note that  $\sum_{\mathbf{c} \in \mathcal{C}} n_{gkc} = n_g$  and  $\sum_{\mathbf{c} \in \mathcal{C}} \pi_g(\mathbf{c}) = 1$ , for

every  $g(= 1, \dots, G)$ . The full multisample, multidimensional, multinomial law is given by

$$\prod_{g=1}^G \frac{n_g!}{\prod_{\mathbf{c} \in C} n_g(\mathbf{c})!} \prod_{\mathbf{c} \in C} [\pi_g(\mathbf{c})]^{n_g(\mathbf{c})}, \tag{1.19}$$

defined over the product simplex  $S_{G \times (C^{K-1})}$ .

Since here the categories  $1, \dots, C$  relate to purely qualitative characteristics (without any implicit ordering), conventional measures of variability are not usable. Rather variation is viewed in the light of mutation rates or other diversity measures, which are functions of the cell probabilities. If we consider the full multinomial model, formulated above, when  $K$  is large, even if  $C$  may not be, for each group there being  $C^K - 1$  cell probabilities, we need the individual  $n_g(\mathbf{c})$  to be at least moderately large, so that  $n_g$  should be  $\gg C^K$ , a condition rarely tenable in HDDSM, specially in genomics context where experiments are excessively costly. For this reason, a full likelihood based statistical inference procedure is impractical in use in HDDSM, and alternative approaches are to be advocated.

We consider here a pseudo-marginal approach wherein for each of the  $K$  marginal probability laws, a convenient measure of diversity is used in a composite way to formulate an overall measure of diversity. Among plausible measures of diversity, we may consider two important ones, namely, the Gini-Simpson index and the entropy measure. For a simple multinomial law with cell probabilities  $\pi_1, \dots, \pi_C$ , the Gini-Simpson Index (GSI) (Gini 1912, Simpson 1949) is defined as

$$I_{GS}(\pi) = 1 - \pi' \pi, \tag{1.20}$$

which attains a maximum value  $(C - 1)/C$  when all the  $\pi_j$  are equal (to  $C^{-1}$ ), and a minimum value 0 when only one of the  $\pi_j$  is equal to 1 and the rest 0. Thus, if we consider a simplex  $S_{C-1}$  then  $\pi$  being defined on this simplex, minimum diversity occurs at the  $C$  vertexes of the simplex while a maximum occurs at the centroid of this simplex. Sen et al. (2007) have exhibited diversity contours based on the GSI. The basic idea is the following: If some gene (position) is not associated with a specific disease/disorder then its variation, as measured by some diversity index, will be stable. On the other hand, for disease-genes, the variation would stochastically differ with more concentration in some specific transitions. Therefore, the average of the marginal diversity measures has an interpretable role to play in this study. Actually, CSI comes in good handy form in such a marginal approach. Let  $\pi_{gk}$  be the vector of cell probabilities of the  $g$ th group at the  $k$ th position, for  $k = 1, \dots, K; g = 1, \dots, G$ . Let us then define

$$\theta_g = K^{-1} \sum_{k=1}^K I_{GS}(\pi_{gk}), g = 1, \dots, G. \tag{1.21}$$

It is also possible to express  $\theta_g$  as  $K^{-1} \sum_{k=1}^K P\{X_{gi,k} \neq X_{gj,k}\}$ , and this is known as the Hamming distance for the probability law  $\pi_g$ . The sample counterparts are easily shown to be (suitable  $U$ -statistics)

$$U_{n_g} = \binom{n_g}{2}^{-1} \sum_{1 \leq i < j \leq n_g} \phi(\mathbf{X}_{gi}, \mathbf{X}_{gj}), \quad g = 1, \dots, G, \tag{1.22}$$

where  $\phi(\mathbf{a}, \mathbf{b}) = K^{-1} \sum_{k=1}^K I(a_k \neq b_k)$  is the Hamming distance between  $\mathbf{a}$  and  $\mathbf{b}$  both being  $K$  vectors. Pinheiro et al. (2005) have incorporated these Hamming distances for the  $G$  groups to test for the homogeneity of the  $\theta_g$ . A subgroup decomposability property (Sen 1999) underlies their formulation.

We combine the  $G$  groups into a single one with  $n$  sequences. Define

$$U_n = \binom{n}{2}^{-1} \sum^* \phi(\mathbf{X}_{gi}, \mathbf{X}_{g'i'}), \tag{1.23}$$

where the summation  $\sum^*$  extends over all possible pairs of vectors from the pooled sample. Further, let

$$U_{n_g, n_{g'}} = (n_g n_{g'})^{-1} \sum_{i=1}^{n_g} \sum_{j=1}^{n_{g'}} \phi(\mathbf{X}_{gi}, \mathbf{X}_{g'j}), \quad g \neq g' = 1, \dots, G; \tag{1.24}$$

these are the generalised  $U$ -statistics for pairs of samples. Then the subgroup decomposability relates to the following:

$$U_n = \sum_{g=1}^G (n_g/n) U_{n_g} + \sum_{1 \leq g < g' \leq G} \frac{n_g n_{g'}}{n(n-1)} \{2U_{n_g, n_{g'}} - U_{n_g} - U_{n_{g'}}\}, \tag{1.25}$$

where the second term, termed the adjusted between group Hamming distance, has nonnegative expectation when the null hypothesis is not true and 0 expectation under the null hypothesis. Thus, we may consider a regular ANOVA-type test based on this decomposition, rejecting the null hypothesis for large positive values. The crux of the problem is to determine the distribution theory of this test statistic under the null hypothesis. We refer to Pinheiro et al. (2005) and Sen et al. (2007) for some details.

### 1.6. UIP and MCP in HDDSM

Sen et al. (2007) have considered CSI problems relating to the Hamming distances for the  $G$  groups presented in the preceding section. Basically, they considered suitable ordered alternatives against the null hypothesis of homogeneity. Thus, the null hypothesis  $H_0$  relates to the homogeneity of the  $\pi_g$ ,  $g = 1, \dots, G$

against alternatives relating to suitable ordering of the Hamming distances  $\theta_g$  having interpretable biological implications. For example, we might have  $H_1^> : \theta_1 \leq \theta_2 \leq \dots \leq \theta_G$  with at least one strict inequality sign being true. Note that there is no linear ordering involved, and moreover, the  $\theta_g$  are all defined on the interval  $[0, (C-1)/C]$  so that translation or scale equivariance property may not hold. For the normal theory models, for such ordered alternative hypothesis testing problems, as treated in detail in Silvapulle and Sen (2004), even if translation or scale equivariance may hold, we may not have an optimal test, specially for partial ordering when there are nuisance scale parameters or dispersion matrices. Roy's UIP has been incorporated to formulate suitable union-intersection test statistics. Further, for small sample sizes and relatively large  $K$ , the exact distribution theory of such test statistics may be difficult to obtain, and the problem becomes even more unmanageable for our contemplated HDDSM. In order to bypass some of these technical difficulties, the UIP is incorporated to formulate suitable test statistics (without necessarily claiming that these are optimal), and a permutation approach is advocated for good approximation for critical levels. If the  $n_g$  were large, Tsai and Sen (2005) prescription for asymptotic CIS would apply here well. However, if the  $n_g$  are relatively small, compared to  $K$ , such approximations are not adequate, and the proposed permutation-sampling provides a better resolution. Recall that under  $H_0$ , all the  $n$  sequences conform to a common multidimensional multinomial probability law, so that all possible partitioning into  $G$  subsets, namely,

$$M_n = \frac{n!}{\prod_{g=1}^G n_g!} \quad (1.26)$$

are equally likely, each having the common probability  $M_n^{-1}$ . This provides the basis for the permutation distribution.

Even if we do so,  $M_n$  could be prohibitively large. For example, in the SARSCoV dataset, we have  $n = 25$ ,  $n_1 = 6, n_2 = 3, n_3 = 4, n_5 = 12$  so that  $M_n$  is unmanageably large. As such, what was done to draw a random sample of 5,000 partitionings from this set, and for each drawn permutation, the UIT statistic was computed. Thus, we arrive at a set of 5,000 values of the test statistic against which the actual sample realisation was compared. This procedure gives a fairly good approximation of the critical level based on the permutation distribution. Of course, this would not give us the exact critical level, but the procedure remains valid for the HDDSM whereas conventional chi-squared bar distributional approximations are much less reliable.

As has been noted earlier, one of the basic problems in high-dimensional data models is the abundance of hypotheses or comparisons, often outnumbering the

sample size. We therefore consider some pertinent remarks on the role of UIP in MCP studies role of UPI in meta analysis, or pooling of information from composite sites, as is currently adapted in in various fields of application, specially in genomic studies. In passing, we should note, however, in such high-dimensional perspectives, the test statistics for different subhypotheses may not be stochastically independent. For multi-center clinical trials, generally conducted under not so homogeneous environment (e.g., different geographical or demographic strata, age / cultural differences), inter-center heterogeneity may account for some extra variation, although all the centers may have a common objective of drawing statistical conclusions that pertain to a broader population. The picture is similar in genomics studies where inter-species heterogeneity may mar the simplicity of usual UIP or MCP approaches. Typically with a huge number of genes and with relatively smaller number of replications, there is a high level of degeneracy of statistical models so that conventional MCP formulations may generally encounter serious roadblocks. At the present there is considerable emphasis on the use of individual gene based statistical analysis and then combining these marginal statistics into some rational statistical inference scheme. Although this is typically in line with conventional meta analysis, possible stochastic dependence among the genes may vitiate standard meta analysis tools. We discuss these problems briefly here.

For motivation, we briefly take a detour to multi-center clinical trials where the clinics can be taken as independent. Consider in this vein,  $C(\geq 2)$  centers, each one conducting a clinical trial with the common goal of comparing a new treatment with an existing one or a control or placebo. Since such centers pertain to patients with possibly different cltural, racial, demographic profiles, diet and physical exercise habits etc. and they may have somewhat different clinical norms too, the intra-center test statistics  $\mathcal{L}_c$ ,  $c = 1, \dots, C$ , used for CSI/RST, though could be statistically independent, might not be homogeneous enough to pull directly. This feature may thus create some impasses in combining these statistics values directly into a pooled one to enhance the statistical information. Meta analysis, in the context of MCP, based on *observed significance levels* (OSL) or  $p$ -values, is commonly advocated in this context. Recall that under the null hypothesis (which again can be interpreted as the intersection of all the center null hypotheses), the  $p$ -values have the common uniform  $(0, 1)$  distribution, providing more flexibility to adopt UIP in meta analysis. Under restricted alternatives, these OSL values are left-tilted (when appropriate UIT are used) in the sense that the probability density is postively skewed over  $(0, 1)$  with high density at the lower tail and low at the

upper. Let us denote the  $p$ -values by

$$P_c = P\{\mathcal{L}_c \geq \text{the observed value} | H_0\}, c = 1, \dots, C. \tag{1.27}$$

The well-known Fisher’s test is based on the statistic

$$F_n = \sum_{c=1}^C \{-2 \log P_c\}, \tag{1.28}$$

which, under the null hypothesis, has the central chi-square distribution with  $2C$  degrees of freedom. This test has some desirable asymptotic (in  $n$ ) properties, albeit in HDLSSM such properties may not be tenable. There are many other tests based on the OSL values. The well known step-down procedure (Roy 1958) has also been adapted in this vein (cf. Mudholkar and Subbaiah 1980, Sen 1983), and they have been amended for CSI and RST as well (cf. Sen 1988). One technical drawback observed in this context is the insensitivity (to small to moderate departures from the null hypothesis) of such tests (including the Fisher’s ) when  $C$  is large, resulting in nonrobust and, to a certain extent, inefficient procedure. In multi-center clinical trials, typically,  $C$  may not be too large, and hence, the extent of nonrobustness and inefficacy of the Fisher method as well as other conventional ones might not be that significant. However, as  $C$  becomes large, these deficiencies can be more apparent. Thus, alternative approaches based on the OSL values have been explored more recently in the literature.

In the evolving field of bioinformatics and genomics, generally, we encounter an excessively high dimensional data set with inadequately small sample size creating impasses for the applicability of standard CSI or even conventional statistical inference tools. On top of that, the OSL values to be combined (corresponding to different genes) may not be independent, and in many cases, due to actual distributional assumptions (e.g., nonparametric ones) may not have strictly uniform distribution under the null hypothesis, creating another layer of difficulty with conventional meta analysis. This led to the development of multiple hypotheses testing in large dependent data models based on OSL values. This field is going through an evolution, and much remains to be accomplished. In this spectrum, the Simes (1986) theorem occupies a focal point. Let there be  $K$  null hypotheses (not necessarily independent)  $H_{0k}$ ,  $k = 1, \dots, K$  with respective alternatives (which possibly could be restricted or constrained as in clinical trials or microarray studies)  $H_{1k}$ ,  $k = 1, \dots, K$ . We thus come across the same UIP scheme by letting  $H_0$  as the intersection of all the component null hypotheses, and  $H_1$  as the union of the component alternatives. Let  $P_k, k = 1, \dots, K$  be the OSL values associated with the hypotheses testing  $H_{0k}$  vs.  $H_{1k}$ , for  $k = 1, \dots, K$ . We denote the

ordered values of these OSL values by

$$P_{K:1}, \dots, P_{K:K}. \quad (1.29)$$

The basic idea is to exploit the information contained in these ordered p-values in a more creative way. If the individual tests have continuous null distributions then the ties among the  $P_k$  (and hence, among their ordered values) can be neglected, in probability. Assuming independence of the  $P_k$ , and uniform distribution for the unordered ones, Simes theorem states that

$$P\{P_{K:k} > k\alpha/K, \forall k = 1, \dots, K | H_0\} = 1 - \alpha. \quad (1.30)$$

Interestingly enough, the Simes theorem is a restatement of the classical Ballot theorem, developed some twenty years earlier (cf. Karlin, 1969): Let  $U_1, \dots, U_K$  be i.i.d. r.v.'s having the Unif.(0, 1) distribution and let  $G_K(u) = K^{-1} \sum_{k=1}^K I(U_k \leq u)$ ,  $u \in (0, 1)$  be the associated empirical d.f. Then, for every  $\gamma \geq 1$ , and every  $K \geq 1$ ,

$$P\{G_K(u) \leq \gamma u, \forall u \in (0, 1)\} = 1 - \gamma^{-1}. \quad (1.31)$$

In any case, granted the unawareness, it is a nice illustration how the UIP is linked to the extraction of extra statistical information through ordered OSL values.

It did not take long time for applied mathematical statisticians to make good uses of the Simes-Ballot theorem in CSI and multiple hypothesis testing problems. The above results pertains to tests for an overall null hypothesis in the UIP setup. Among others, Hochberg (1988) incorporated a variant of the above result:

$$P\{P_{K:j} \geq \alpha/(K - j + 1), \forall j = 1, \dots, K | H_0\} = 1 - \alpha, \quad (1.32)$$

in a multiple testing framework. Benjamini and Hochberg (1995) introduced the concept of false discovery rate (FDR) in the context of multiple hypothesis testing, and illustrated the role of the Simes-Ballot theorem in that context. Just to point out how difficult may be such procedures, let us consider the following illustrative example (Sen et al. 2007). Suppose that  $K = 192$  and there are 4 groups with sample sizes 4,6,3 and 12 respectively. Basically then one has 192 tests in a multiple hypotheses testing setup. Even if we assume multinormality, the sample sizes are so small that the recording of the actual p-values from appropriate tables could be very sensitive for values close to zero (or 1); the tabular values could be quite different from the actual ones. A permutation approach, as has been prescribed in Sen et al. (2007), may be quite conservative and thereby subject to the same limitation. On top of that for any chosen value of  $\alpha$ , the numbers  $\alpha/K$  or  $\alpha/(K - j + 1)$ , for small values of  $j(\geq 1)$  will be so small that these MTP will have too little power. For example, for  $\alpha = 0.05$  and  $K = 192$ , we have  $\alpha/K = 0.00025$

so that we need to have a fairly accurate recording of the actual  $p$ -values, especially near the lower end-point 0. From robustness point of view, this is often a challenging task.

The past ten years have witnessed a phenomenal growth of research literature in this subfield with applications to genomics and bioinformatics. The basic restraint in this respect is the assumption of independence of the  $P_j, j = 1, \dots, K$ , and in bioinformatics, this is hardly the case. Sarkar (1998) and Sarkar and Chang (1997) incorporated the  $MTP_2$  (multivariate total positivity of order 2) property to relax the assumption of independence to a certain extent. Sarkar (2000, 2002, 2004) has added much more to this development with special emphasis on controlling of FDR in some dependent cases. The literature is too large to cite adequately, but our primary emphasis here is to stress how UIP underlies some of these developments and to focus on further potential work.

Combining OSL values, in whatsoever manner, may generally involve some loss of information when the individual tests are sufficiently structured to have coherence that should be preserved in the meta analysis. We have seen earlier how guided by the UIP, progressive censoring in clinical trials provided more efficient and interpretable testing procedures. The classical Cochran-Mantel-Haenszel (CMH) procedure is a very notable example of this line of attack. In a comparatively more general multiparameter CSI setting, Sen (1999b) has emphasized the use of the CMH procedure in conjunction with the OSL values to induce greater flexibility. The field is far from being saturated with applicable research methodology. The basic assumption of independence or specific type of dependence is just a part of the limitations. A more burning question is the curse of dimensionality in CSI problems. Typically, there  $K$  is large and the sample size  $n$  is small, i.e.,  $K \gg n$ . In the context of clinical trials in genomics setups, Sen (2006) has appraised this problem with due emphasis on the UIP. Conventional test statistics (such as the classical LRT) have awkward distributional problems so that usual OSL values are hard to compute and implement in the contemplated CSI problems. Based on the Roy (1953) UIP but on some nonconventional statistics, it is shown that albeit there is some loss of statistical information due to the curse of dimensionality, there are suitable tests which can be implemented relatively easily in high-dimension low sample size environments. In CSI for clinical trials in the presence of genomics undercurrents, there is a tremendous scope for further developments along this line.

## References

1. Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. (1972). *Statistical Inference under Order Restrictions*, John Wiley, New York.
2. Benjamini, Y. and Hochberg, Y. (1995). *Controlling the false discovery rate: A practical and powerful approach to multiple testing*. Jour. Roy. Statist. Soc. B57., 289-300.
3. Cox, D. R. (1972). *Regression models and life tables (with discussion)*. Jour. Roy. Statist. Soc.B 34, 187-220.
4. DeMets, D.L. and Lan, K.K.G. (1983). *Discrete sequential boundaries for clinical trials*. Biometrika 70, 659-663.
5. Dardanoni, V. and Forcena, A. (1998). *A unified approach to likelihood inference on stochastic ordering in a nonparametric context*. Jour. Amer. Statist. Assoc.93, 1112 - 1123.
6. Hochberg, Y. (1988). *A sharper Bonferroni procedure for multiple tests of significance*. Biometrika 75, 800 - 802.
7. Karlin, S. (1969). *A first Course in Stochastic Processes*, Academic Press, New York.
8. McDermott, M. P. and Mudholkar, G. S. (1993). *A simple approach to testing homogeneity of order-constrained means*. Jour. Amer. Statist. Assoc. 88, 1371 - 1379.
9. Mudholkar, G. S., Kost, J. and Subbaiah, P. (2001). *Robust tests for orthant - restricted mean vector*. Commun. Statist. Theor. Meth. 30, 1789 - 1810.
10. Mudholkar, G. S. and McDermott, M. P. (1989). *A class of tests for equality of ordered means*. Biometrika 76, 161 - 168.
11. Mudholkar, G. S. and Subbaiah, P. (1980). *Testing significance of a mean vector - a possible alternative to Hotelling  $T^2$* . Ann. Institut. Statist. Math. 32, 43 - 52.
12. Perlman, M. D. (1969). *One-sided problems in multivariate analysis*. Ann. Math. Statist. 40, 549-567.
13. Pinheiro A.S., Pinheiro, H.P. and Sen, P.K. (2005) *Comparison of genomic sequences by Hamming distance*. Jour. Statist. Plan. Infer. 130, 325-339.
14. Robertson, T., Wright, F.T. and Dykstra, R. (1988). *Order Restricted Statistical Inference*, John Wiley, New York.
15. Roy, J. (1958). *Step-down procedures in multivariate analysis*. Ann. Math. Statist. 29, 1177 - 1188.
16. Roy, S. N. (1953). *On a heuristic method of test construction and its use in multivariate analysis*. Ann. Math. Statist.24, 220-238.
17. Roy, S.N. (1957). *Some Aspects of Multivariate Analysis*, John Wiley, New York, and Asia Publ. House, Bombay.
18. Roy, S.N., Gnanadesikan, R. and Srivastava, J.N. (1971). *Analysis and Design of Certain Quantitative Multiresponse Experiments*, Pergamon Press, New York.
19. Sarkar, S.K. (1998). *Some probability inequalities for ordered  $MTP_2$  random variables: a proof of the Simes conjecture*. Ann. Statist.26, 494-504.
20. Sarkar, S.K. (2000). *A note on the monotonicity of the critical values of a step-up test*. Jour. Statist. Plann. Infer. 87, 241-249.
21. Sarkar, S.K. (2002). *Some results on false discovery rate in multiple testing procedures*. Ann. Statist. 30, 239-257.
22. Sarkar, S.K. (2004). *FDR-controlling stepwise procedures and their false negatives rates*. Jour. Statist. Plann. Infer. 125, 119-137.

23. Sarkar, S.K. and Chang, C.-K. (1997). *The Simes method for multiple hypothesis testing with positively dependent test statistics*. Jour. Amer. Statist. Assoc. 92, 1601-1608.
24. Sen, P.K. (1981). *Sequential Nonparametrics: Invariance Principles and Statistical Inference*, John Wiley, New York.
25. Sen, P.K. (1983). *A Fisherian detour of the step-down procedure*. In *Contributions to Statistics: Essays in honour of Norman L. Johnson*, North Holland, Amsterdam, pp. 367-377.
26. Sen, P.K. (1988). *Combination of statistical tests for multivariate hypotheses against restricted alternatives*. In *Advances in Multivariate Statistical Analysis* (eds. S. Dasgupta and J.K. Ghosh), Ind. Statist. Inst. pp. 377-402.
27. Sen, P. K. (1999 a). *Multiple comparisons in interim analysis*. Jour. Statist. Plann. Infer. 82, 5-23.
28. Sen, P.K. (1999 b). *Some remarks on the Stein-type multiple tests of significance*. Jour. Statist. Plann. Infer. 82, 139-145.
29. Sen, P. K. (2001). *Survival analysis: Parametrics to semiparametrics to pharmacogenomics*. Brazilian Jour. Probab. Statist. 15, 201 - 220.
30. Sen, P. K. (2005). *Nonparametric tests for ordered diversity in a genomic sequence*. Jour. Statist. Res. 39, No. 2, 7 - 21.
31. Sen, P. K. (2006). *Robust Statistical inference for high-dimension low sample size problems with applications to genomics*. Austrian Jour. Statist. 35 197-214.
32. Sen, P. K. (2007). *Union-intersection principle and constrained statistical inference*. Jour.Stat. Plan. Infer. 1bc, in press.
33. Sen, P.K. and Puri, M. L. (1967). *On the theory of rank order tests in the multivariate one-sample problem*. Ann. Math. Statist. 136, 3741-3752.
34. Sen, P. K. and Tsai, M.-T. (1999). *Two-stage likelihood ratio and union-intersection tests for one-sided alternatives multivariate mean with nuisance dispersion matrix*. Jour. Multivar. Anal. 68, 264-282.
35. Sen, P. K., Tsai, M.-T. and Jou, Y.-S. (2007). *High dimension low sample size perspectives in constrained statistical inference: The SARSCoV genome in illustration*. Jour. American Statist. Assoc., 102, in press.
36. Shapiro. A. (2000). *On the asymptotics of constrained local M-estimators*. Ann. Statist. 28, 685-694.
37. Silvapulle, M.J. (1995). *A Hotelling  $T^2$ -type statistic for testing against one-sided hypotheses*. Jour. Multivar. Anal. 55, 312-319.
38. Silvapulle, M.J. and Sen, P. K. (2004). *Constrained Statistical Inference: Inequality, Order and Shape Restrictions*, John Wiley, New York.
39. Silvapulle, M. J. and Silvapulle, P. (1995). *A score test against one-sided alternatives*. Jour. Amer. Statist. Assoc. 90, 342-349.
40. Simes, R.J. (1986). *An improved Bonferroni procedure for multiple tests of significance*. Biometrika 73, 751-754.
41. Tsai, M.-T. and Sen, P.K. (2005). *Asymptotically optimal tests for parametric functions against ordered functional alternatives*. Jour. Multivar. Anal. 95, 37-49.
42. Wald, A. (1943). *Tests of statistical hypotheses concerning several parameters when the number of observations is large*. Trans. Amer. Math. Soc. 54, 426-482.
43. Wijsmann, R.A. (1979). *Constructing all smallest simultaneous confidence sets in a general class with applications to MANOVA*. Ann. Statist. 7, 1003-1018.