Moonsu Kang; Pranab Kumar Sen

Kendall's tau-type rank statistics in genome data

Applications of Mathematics, Vol. 53 (2008), No. 3, 207–221

Persistent URL: `http://dml.cz/dmlcz/140316`

KENDALL'S TAU-TYPE RANK STATISTICS IN GENOME DATA

Moonsu Kang, Pranab K. Sen, Chapel Hill

(Invited)

Abstract. High-dimensional data models abound in genomics studies, where often inadequately small sample sizes create impasses for incorporation of standard statistical tools. Conventional assumptions of linearity of regression, homoscedasticity and (multi-) normality of errors may not be tenable in many such interdisciplinary setups. In this study, Kendall's tau-type rank statistics are employed for statistical inference, avoiding most of parametric assumptions to a greater extent. The proposed procedures are compared with Kendall's tau statistic based ones. Applications in microarray data models are stressed.

Keywords: dimensional asymptotics, genomics, multiple hypotheses testing, microarray data model, nonparametrics, U-statistics

MSC 2010: 62G10, 62G99, 62P99

1. Introduction

In genomic studies as well as in many other interdisciplinary research, high-dimensional low sample size (HDLSS) data models arise with a variety of complexities due to many extraneous factors. For example, in genomics, there may be a huge number of genes not necessarily statistically (as well as biologically) independent, and there may be only a few arrays constituting such HDLSS models. In view of possible gene-environment interaction and the importance of disease gene mapping, there is a genuine need to develop suitable statistical inference procedures based on biologically consistent, plausible assumptions. Gene expression levels across various experiments (or treatments) setups have complex structures, often, marred by inequality, order or other constraints, for which standard procedures, such as mulitivariate analysis of variance (MANOVA) may not work out well. The gene expression levels for different genes, in addition to being possibly dependent, exhibit considerable heterogeneity, thus invalidating the classical MANOVA tools in such HDLSS setups. Therefore, some robust, nonstandard procedures need to be incorporated in such studies.

Recently, Sen [17] has formulated some procedures based on the Kendall [11] tau statistic, albeit in a HDLSS setup, bypassing the heterogeneity and linearity of the regression assumption to a greater extent. This procedure is insensitive to outliers, although it incorporates inter-gene dependence by a tactful use of the Chen-Stein [5], [6] theorem, extended to a discrete time parameter Poisson process approximation. Rank statistics incorporate the relative order of the observations better than the Kendall's tau statistic. Therefore, it is more appealing to use such linear rank statistics in HDLSS models. However, there is some underlying emphasis on linear regression models which are not likely to be tenable in genomic studies. Therefore, we have considered a hybrid of linear rank statistics and Kendall's tau, which have the advantage of providing more appropriate statistical tools in genomic studies.

Section 2 deals with a typical microarray data setup where the proposed statistics are appropriate. Section 3 is devoted to the formulation of these statistics and the study of their properties too. Section 4 deals with multiple hypotheses testing (MHT) problem arising in this context. In Section 5, an extension of the Chen-Stein theorem, as formulated in Sen [17], is incorporated in the formulation of statistical inference tools. The last section is devoted to general discussion along with an illustration of the Lobenhofer et al. [12] study.

2. Preliminaries

Consider a DNA microarray data model with a large number (K) of genes, each having (gene) expression levels on n (small) arrays. Thus the dataset can be represented as an $n \times K$ matrix $\mathbf{X} = ((X_{ik}))$, where X_{ik} stands for the gene expression level of the kth gene in the ith array, $i = 1, \ldots, n$; $k = 1, \ldots, K$. Typically, $K \gg n$, and often, n is small. We write $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$; $\mathbf{X}_i = (X_{i1}, \ldots, X_{iK})'$, $1 \leqslant i \leqslant n$. In some cases, the K-vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ can be taken to be stochastically independent. However, the K coordinate variables in each $\mathbf{X}_i$ may be neither independent nor (marginally) identically distributed. Guided by this feature, we denote the K-variate distribution function of $\mathbf{X}_i$ by $F_i(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^K$, for $i = 1, \ldots, n$. Further, the marginal distribution of X_{ik} is denoted by $F_{ik}(x)$, $x \in \mathbb{R}$, $k = 1, \ldots, K$. Often, the arrays relate to possibly different biological or experimental (environmental) setups. Hence, we can conceive of some design variable t_i, $i = 1, \ldots, n$, which, for example, may relate to disease severity. Without loss of generality, we assume that

$$t_1 \leqslant t_2 \leqslant \ldots \leqslant t_n \tag{2.1}$$

with at least one strict inequality. However, apart from this weak ordering, we do not impose any linear or nonlinear ordering on t_i.

Let us look into the n arrays for the kth gene, thus comprising the set $(X_{1k}, \ldots, X_{nK})$ of responses for $k = 1, \ldots, K$. If the gene k is associated with an experimental variation (such as severity of disease) then the expression level X_{ik} should have a stochastic ordering across the levels $t_1, t_2, \ldots, t_n$, i.e.,

$$F_{1k}(x) \geqslant F_{2k}(x) \geqslant \ldots \geqslant F_{nk}(x), \quad \forall\, x \in \mathbb{R}. \tag{2.2}$$

This stochastic ordering is weaker than the ordered mean or isotonic regression setups. Since the domain of x is typically the interval $[0, 1]$ (or $[0, 100]$) in terms of luminosity of the gene expression levels, shift in location or scale parameters may not be ideal in this setup.

For the same reason, a regression model like $F_{ik}(x) = F_{0k}(x - \beta t_i)$, $1 \leqslant i \leqslant n$; $k = 1, \ldots, K$, may not be appropriate. Thus, in our setup, we would like to develop robust statistics taking into account a plausible stochastic ordering as an alternative to the null distribution of homogeneity of F_{ik} $(1 \leqslant i \leqslant n)$.

It is also very complex to frame such a stochastic ordering on the K variate vector $\mathbf{X}_i$, $1 \leqslant i \leqslant n$. Hence, as in Sen et al. [18] and Sen [17], we will adopt a pseudo-marginal approach as follows. Let $H_{0k}\colon\ F_{1k} \equiv \ldots \equiv F_{nk}$, $1 \leqslant k \leqslant K$ and let $H_{1k}\colon\ F_{1k} \geqslant \ldots \geqslant F_{nk}$, $1 \leqslant k \leqslant K$, for $k = 1, \ldots, K$. Let then

$$H_0 = \bigcap_{k=1}^{K} H_{0k} \quad \text{vs} \quad H_1 = \bigcup_{k=1}^{K} H_{1k}. \tag{2.3}$$

In this setup, we consider a suitable test statistic T_{nk} for testing H_{0k} vs H_{1k}, $1 \leqslant k \leqslant K$, and then incorporate the Roy [13] union-intersection principle (UIP) to formulate overall test procedures. Further, in testing H_{0k} vs H_{1k}, we would like to use a nonparametric statistics having some nice features:

(i) Its distribution under H_{0k} does not depend on the common F_{0k},
(ii) it is insensitive to heterogeneity of the F_{0k} for different k, and
(iii) it is robust and efficient.

Sen [17] used Kendall's [11] tau-statistics for testing H_{0k} vs H_{1k}, $k = 1, \ldots, K$. This can be written as

$$T_{nk} = \sum_{1 \leqslant i < i' \leqslant n} \operatorname{sign}(t_{i'} - t_i) \operatorname{sign}(X_{i'k} - X_{ik}) \tag{2.4}$$

for $k = 1, \ldots, K$. Note that $\operatorname{sign}(t_{i'} - t_i)$ is invariant under any strictly monotone transformation on t_i. Similarly, $\operatorname{sign}(X_{i'k} - X_{ik})$ is invariant under any strictly monotone transformation on X_{ik}. Thus, T_{nk} is doubly-invariant with respect to t_i

as well as X_{ik}, thus providing a robust procedure. On the other hand, a linear rank statistic for testing H_{0k} vs H_{1k} may be defined as

$$L_{nk} = \sum_{i=1}^{n}(t_i - \bar{t}_n)a(R_{ik}), \quad 1 \leqslant k \leqslant K, \tag{2.5}$$

where R_{ik} = rank of X_{ik} among $X_{1k}, \ldots, X_{nk}$ for $i = 1, \ldots, n$; $\bar{t}_n = n^{-1}\sum_{i=1}^{n} t_i$, $a_n(1) \leqslant \ldots \leqslant a_n(n)$ are suitable scores and $\bar{a}_n = n^{-1}\sum_{i=1}^{n} a_n(i)$. Whereas the ranks R_{ik} are invariant under any strictly monotone transformation on X_{ik} in L_{nk}, the assigned values of t_i have a non-invariant property. In this sense, L_{nk} is perceived as somewhat less robust than the Kendall's tau statistics. However, through appropriate choice of $a_n(i)$, a gain in efficiency is possible, if the assumed t_i values are correct (up to a location/scale perturbation location). For some details, we refer to Kang [10].

We intend to propose some linear rank statistics which are of the Kendall's tau-type, i.e., invariant to any strictly monotone transformation on the t_i, while keeping the scores $a_n(i)$ and ranks R_{ik} unchanged. This way, it would compromise robustness and efficiency in a more conceivable manner, in such a nonstandard setup.

3. Proposed statistics

We may rewrite L_{nk} as equivalent to

$$\sum_{1 \leqslant i < i' \leqslant n} (t_{i'} - t_i)(a(R_{i'k}) - a(R_{ik})), \tag{3.1}$$

so that the dependence on the assigned values of t_i becomes clear. In conformity with the Kendall's tau, we consider the statistics

$$T_{nk} = \sum_{1 \leqslant i < i' \leqslant n} \mathrm{sign}(t_{i'} - t_i)(a(R_{i'k}) - a(R_{ik})) \tag{3.2}$$

for $k = 1, \ldots, K$. These are all rank statistics but not linear ones in the conventional sense. Given (2.1), we set $S = \{(i, i')\colon\ t_i < t_{i'};\ 1 \leqslant i < i' \leqslant n\}$ and let N = cardinality of S (so that $(n-1) \leqslant N \leqslant \binom{n}{2}$). Then we have

$$T_{nk} = \sum_{S}(a(R_{i'k}) - a(R_{ik})), \quad k = 1, \ldots, K. \tag{3.3}$$

If instead of (3.2) we would have taken

$$\sum_{1\leqslant i<i'\leqslant n} \operatorname{sign}(t_{i'}-t_i)\operatorname{sign}(a(R_{i'k})-a(R_{ik})) \tag{3.4}$$

then for strictly monotone scores $a_n(\cdot)$, such as the Wilcoxon score ($a_n(i) = i/(n+1)$), we would have

$$\operatorname{sign}[a(R_{i'k})-a(R_{ik})] = \operatorname{sign}[R_{i'k}-R_{ik}] = \operatorname{sign}(X_{i'k}-X_{ik}) \tag{3.5}$$

so that by (3.5), (3.4) would reduce to the Kendall's tau statistic in (2.4). This explains motivation for the T_{nk} in (3.2), combining more information on the scores (through ranks) and invariance on the t_i, $1 \leqslant i \leqslant n$.

From (3.3) we conclude that under H_{0k}, $(R_{1k},\ldots,R_{nk})$ taking all possible ($n!$) permutations of $(1,\ldots,n)$ with equal probabilty $1/n!$ leads to an exact distribution-free statistic, i.e, under H_{0k}, the distribution of T_{nk} does not depend on the underlying (common) F_{0k}. Further, this marginal (null) distribution remains the same for all k ($=1,\ldots,K$), as the set S remains the same for all k. As such,

$$E_0(T_{nk}) = \sum_S E_0(a_n(R_{i'k})-a_n(R_{ik})) = \sum_S \frac{1}{n(n-1)} \sum_{j\neq l=1}^{n} [a_n(j)-a_n(l)] = 0. \tag{3.6}$$

Further, if we let $S_1 = \{((i,j),\,(i',j'))\colon\ t_i < t_j$ and $t_{i'} < t_{j'};\ 1 \leqslant i < j \leqslant n,\ 1 \leqslant i' < j' \leqslant n,\ ((i=i'$ and $j \neq j')$ or $(j=j'$ and $i \neq i'))\}$, $S_2 = \{((i,j),(i',j'))\colon\ t_i < t_j$ and $t_{i'} < t_{j'};\ 1 \leqslant i < j \leqslant n,\ 1 \leqslant i' < j' \leqslant n,\ ((j=i'$ and $i \neq j')$ or $(i = j'$ and $j \neq i'))\}$, $S_3 = \{((i,j),(i',j'))\colon\ t_i < t_j$ and $t_{i'} < t_{j'};\ 1 \leqslant i < j \leqslant n$ and $1 \leqslant i' < j' \leqslant n,\ (i \neq i'$ and $j \neq j')\}$ with cardinalities N_1, N_2 and N_3, respectively, then

$$\begin{aligned}
V_0(T_{nk}) &= E_0\Big[\sum_S (a(R_{i'k})-a(R_{ik}))\Big]^2 \\
&= E_0\Big[\sum_S (a(R_{i'k})-a(R_{ik}))^2\Big] \\
&\quad + E_0\Big[\sum_{S_1} (a(R_{jk})-a(R_{ik}))(a(R_{j'k})-a(R_{i'k}))\Big] \\
&\quad + E_0\Big[\sum_{S_2} (a(R_{jk})-a(R_{ik}))(a(R_{j'k})-a(R_{i'k}))\Big] \\
&\quad + E_0\Big[\sum_{S_3} (a(R_{jk})-a(R_{ik}))(a(R_{j'k})-a(R_{i'k}))\Big] \\
&= A+B+C+D = (2N+N_1-N_2)\times A_n^2 = \omega_n^2
\end{aligned} \tag{3.7}$$

where

$$A_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}[a_n(i) - \bar{a}_n]^2. \tag{3.8}$$

The proof is relegated to the Appendix.

4. Multiple hypotheses testing and UIP

We intend to test for H_0 vs H_1 in (2.2) along with the provision of multiple hypotheses testing. In line with the UIP, we have T_{nk}, $1 \leqslant k \leqslant K$, for the component hypotheses testing problem which we want to incorporate in an overall testing scheme, allowing multiple hypotheses testing. Though, under H_0, the T_{nk} have all a common distribution, symmetric about 0 and independent of the underlying F_{0k}, the behaviour of T_{nk} under alternatives would depend on the stochastic ordering as well as the possible heterogeneity of the F_{ik} for different k. Under the stochastic ordering in (2.2), it is easy to show that

$$\xi_{nk} = E[T_{nk}|H_{1k}] \geqslant 0, \quad \forall\, k = 1, \dots, K, \tag{4.1}$$

although the ξ_{nk} may differ from one k to another. This motivates us to use tests based on the right-hand side p-values for the individual T_{nk}. As in the case of Kendall's tau statistics, for small n, T_{nk} has a discrete distribution, and hence, for the p-values, the distribution (though known) will not be uniform on (0,1) but a discrete one on [0,1]. For large n, the standardized form of T_{nk} (i.e, T_{nk}/ω_n) under H_{0k} has closely standard normal distribution, and hence, the p-values have closely the uniform [0,1] distribution.

The UIP leads us to consider the UI-test statistic

$$T_n^* = \max\{T_{nk}\colon\ 1 \leqslant k \leqslant K\}. \tag{4.2}$$

If T_{nk} were independent, then we would have

$$P_0\{T_n^* \leqslant x\} = [P_0\{T_{n1} \leqslant x\}]^K, \tag{4.3}$$

so that the marginal distribution of T_{n1} could be used to compute the significance level for T_n^*. If T_{nk} are nonnegatively associated then in (4.3), the "=" sign can be replaced by" $\geqslant$" sign, so that the same critical level can be used, albeit giving a somewhat conservative test. For large K, if $n = O(\log K)$ (or larger), then the Bonferroni bound (Sen [17]) provides good approximation.

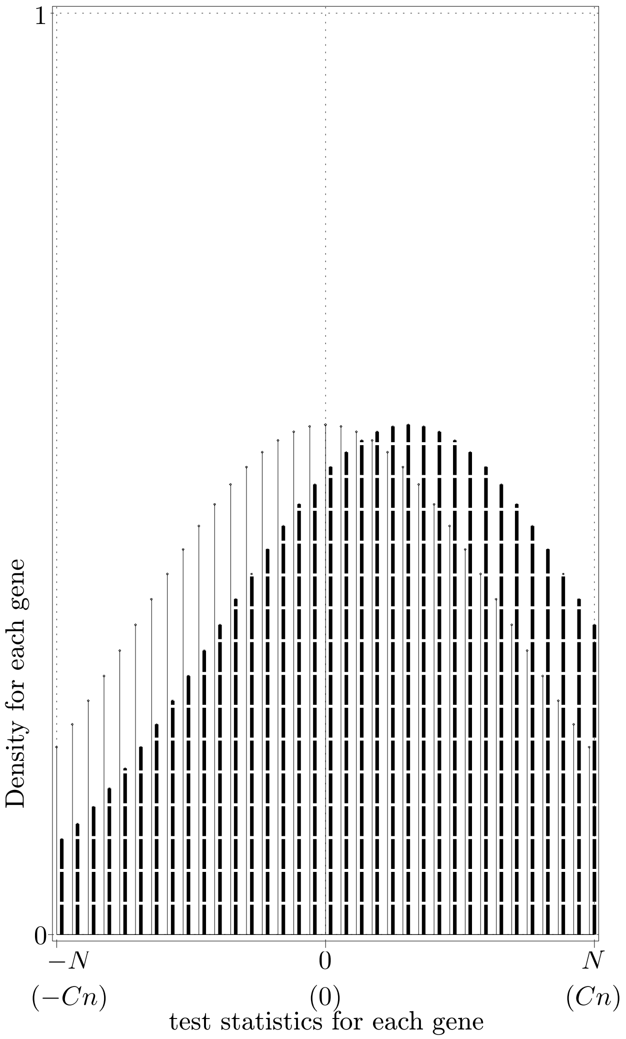

Figure 1. Comparison of the null distribution with the alternative distribution: This picture relates only to Kendall's tau statistics considered in Sen [17]. For general T_{nk}, the range will not be $(-N, N)$, but $\pm C_n$, where $C_n = \max\limits_{i,i'} \sum\limits_{S} [a_n(i') - a_n(i)]$.

Let $t^*_{n,\alpha}$ be the critical level computed from (4.3). Then the following MHT procedure can be adopted. If all $T_{nk} \leqslant t^*_{n,\alpha}$, accept H_0. If $T^*_n > t^*_{n,\alpha}$, reject H_0. Further, in the case of rejection of H_0, at least one T_{nk} will be larger than $t^*_{n,\alpha}$. Thus, reject those H_{0k} in favor of H_{1k} for which $T_{nk} > t^*_{n,\alpha}$.

Based on the perspectives in genomic studies relating to False discovery rate (FDR) and other measures, the above MHT may not be powerful. Therefore, we proceed to consider an alternative way of incorporating the p-value.

5. Chen-Stein theorem and MHT

First, let us present an extended version of the Chen-Stein theorem, considered in detail in Sen [17]; the proof is omitted.

Consider a set of (discrete) mass points $\{\tau_1 < \ldots < \tau_M\}$, the possible realization of T_{n1}, and let $\eta_{n1}, \ldots, \eta_{nM}$ be their respective probability masses under H_0; M, typically, depends on n and $t_1, \ldots, t_n$ (or N). Also, let

$$\nu_{nj} = \sum_{i \leqslant j} \eta_{M-i+1} \quad \text{for} \quad j = 1, \ldots, M. \tag{5.1}$$

Let

$$Y_{kj} = I(T_{n1} \geqslant \tau_{M-j+1}), \quad 1 \leqslant k \leqslant K, \ 1 \leqslant j \leqslant M. \tag{5.2}$$

Thus, Y_{kj} are increasing in j and $E_0(Y_{kj}) = \nu_{nj}$, $1 \leqslant j \leqslant M$. For each $k \in K = \{1, \ldots, K\}$, let J_k be the dependence set of k and let J_k^C be the complementary independence set.

Let then

$$l_{j1} = \sum_{k=1}^{K} \sum_{q \in J_k} E(Y_{kj})E(Y_{qj}), \tag{5.3}$$

$$l_{j2} = \sum_{k=1}^{K} \sum_{q \in J_k} E(Y_{kj}Y_{qj}), \tag{5.4}$$

$$l_{j3} = \sum_{k=1}^{K} E|\{E(Y_{kj} - EY_{kj})|Y_{qj}, \quad \forall\, q \neq k \in J_k\}| \tag{5.5}$$

for $j = 1, \ldots, M$. Let $W_{Kj} = \sum\limits_{k=1}^{K} Y_{kj}$, $j \geqslant 1$ and denote

$$m = \max_{j \leqslant M} \{(l_{j1} + l_{j2} + l_{j3})\nu_{nj}^{-1}(1 - \mathrm{e}^{-K\nu_{nj}})K^{-1}\} \to 0 \tag{5.6}$$

as $k \to \infty$. Let $\mathbf{W}_K = (W_{k1}, \ldots, W_{km})$ and $\mathbf{Z}_K = (Z_{k1}, \ldots, Z_{km})$, where $\mathbf{Z}_K$ is a discrete time parameter Poisson process with $EZ_{Kj} = EW_{Kj} = K\nu_{nj}$, $j \geqslant 1$. Then

$$\|\mathcal{L}(\mathbf{W}_K) - \mathcal{L}(\mathbf{Z}_K)\| \to 0 \quad \text{as} \quad K \to \infty. \tag{5.7}$$

(If $M \to \infty$ (with $n \to \infty$)), then the discrete time parameter processes can be replaced by continuous time parameter processes.)

Corresponding to $\mathbf{W}_K$, let $r_1 \leqslant \ldots \leqslant r_m$ be a set of nonnegative integers such that, by (5.7),

$$P_0\{W_{Kj} > r_j \text{ for some } j \leqslant m\} \sim \alpha \tag{5.8}$$

for some specified α: $0 < \alpha < 1$. Then we consider the following MHT procedure:

Consider the W_{Kj}, $j \leqslant 1$. If $W_{Kj} \leqslant r_j$, $\forall j \geqslant 1$, then accept H_0, i.e., there is no disease gene. If, on the other hand, $W_{Kj} > r_j$ for at least one j, then reject the null hypothesis and proceed to detect those genes as disease gene for which $Y_{kj} = 1$ for some j.

In particular, a two-stage Poisson approximation considered in Kang [10] has better power and FDR prospects than a single stage procedure.

6. Discussion

High dimensional data such as microarray experiments raise large multiplicity problems in which thousands of hypotheses are simultaneously tested. For this multiple hypotheses testing issue, consider the proportion of falsely rejected hypotheses among the number of rejections.

This is introduced as follows.

	Number not rejected	Number rejected	Total
Non-differentially expressed (NDG)	U	V	m_0
Differentially expressed (DG)	T	S	$K - m_0$
	$K - R$	R	K

Table 1. Number of errors committed when testing K tested genes.

In Tab. 1, V represents the number of rejected genes among non-differentially expressed genes and R $(= V+S)$ represents the number of rejected genes. The focus lies in the proportion of false positives V with respect to the number of rejected hypotheses R. In the microarray setting, there is a null hypothesis H_i for each gene i and the rejection of H_i corresponds to declaring that the gene i is differentially expressed. In general, we would like to minimize the number V corresponding to Type I error and the number T corresponding to Type II error. Benjamini and Hochberg [4] introduced the concept of the false discovery rate (FDR). Let the unobserved random variable be $Q = V/(V+S)$, $Q = 0$ when $V+S = 0$. We define the FDR as $E(Q) = E\{V/(V+S) \mid V+S > 0\} \times V+S > 0 = E\{V/R \mid R > 0\} \times P(R > 0)$. Storey [19] suggested an updated version of FDR called pFDR, $E\{V/R \mid R > 0\}$, which is defined as the conditional FDR given that there is at least one rejection. The expected value of the proportion, False Discovery Rate (FDR), is a useful alternative to the familywise error rate (FWER) which is unduly conservative when there are many tests (or genes). However, current FDR controlling procedures fail to incorporate plausible complex dependence structures among tested genes in an appropriate manner.

It is possible to categorize tested genes into two groups: non-differentially expressed genes (NDG) and differentially expressed genes (DG). In general, the NDG have smaller gene expression levels based on the intensity, whereas the DG have bigger gene expression levels. This aspect enables us to conceive of some milder regularity conditions, namely that the correlations between the DG and the NDG is negligible and the correlations between the NDGs are very small as well. As a matter of fact, it is hard to find out some regularity conditions that (classical) central limit theorems may work out. Under the milder conditions discussed above, it may be more captivating to apply the Chen-Sten theorem which is one of the Poisson-limit theorems to allow more general dependence structures among tested genes in Arratia et al. [1]. Utilizing this theorem, it may become more feasible to apply the FDR method to dependent tested genes. For more details, we refer to Sen [17].

For illustration purpose, consider the data in Lobenhofer et al. [12]. This data consists of 1900 genes in rows, and each row (or gene) has 48 observations measured at 6 time points with 8 observations. We can approximate the normal distribution to the null and distribution of T_{nk} (or P_k), since we have sufficiently large sample size. In this dataset, we evaluate the performance of Kendall's tau rank statistics over that of Kendall's tau statistics considered in Sen [17]. We use a single-stage FDR procedure in Kang [10]. Under the null hypothesis, Kendall's tau statistics $T_{nk}^{(1)}$ has $E_0(T_{nk}^{(1)}) = 0$ and $V_0(T_{nk}^{(1)}) = (N + \frac{1}{3}(N_1 - N_2))/N^2 = 0.002256944$, whereas Kendall's tau-type rank statistics has the same expectation $E_0(T_{nk}) = 0$ but $V_0(T_{nk}) = (2 * N + N_1 - N_2) * A_n^2 = 705.3061$. The computations of N_1 and N_2 are given in Appendix.

It is quite feasible to compute the p-value (P_k) for each row (each gene) based on the asymptotic normal distribution of T_{nk}, the assumption of uniform distribution of the p-values under the null being tenable for $n = 48$. One of the concerns is that p-values are not stochastically independent even though each p-value has the same marginal null distribution. The FDR method based on the Chen-Stein method plays a key role in this context.

As explained earlier, under the stochastic ordering (NDG vs DG) we have

$$E[T_{nk}^{(1)} | H_{1k}] \geqslant 0; \quad E[T_{nk} | H_{1k}] \geqslant 0, \quad k = 1, \dots, K.$$

For small sample size, the permutation distribution based on exact permutation theory can be extensively used. It is plausible to construct a suitable test based on the right-hand side of the critical region of the permutation distribution of each T_{nk} (or $T_{nk}^{(1)}$).

This distribution is the same for every gene. Though it is not continuous, the distribution of P_k, under H_0, is discrete over (0,1) and the mass points depend on the scores $a_n(i)$, so that ties among the T_{nk} may not be negligible with probability one. Moreover, the T_{nk}, and hence the P_k may not be independent across all genes tested; the Chen-Stein method may be used to apply the FDR method.

For such a relatively large n $(= 48)$, we have the following asymptotic distributions for each statistics:

$$T_{nk}^{(1)}/\sqrt{V_0(T_{nk}^{(1)})} \to \mathcal{N}(0,1); \quad T_{nk}/\sqrt{V_0(T_{nk})} \to \mathcal{N}(0,1).$$

In Tab. 2, the FDR and Storey's FDR in Storey [19] for Kendall's tau statistics and Kendalls' tau-type rank statistics are presented. We have 1864 genes in Lobenhofer et al. dataset, but we removed every gene having missing gene expression levels. The total number of tested genes is 1818. Using Storey's method, we estimate π_0 (the proportion of non-differentially expressed genes) as 0.83 for p-values generated by Kendall's tau, whereas π_0 as 0.86 for p-values generated by Kendall's tau-type rank statistics. We vary the number of rejected genes r as 4, 8, 12 and 16. Now, controlling the FDR at $\alpha = 0.05$, Poisson distributional approximation leads to finding an appropriate cut-off point, that is,

$$P_0(W > r) = \alpha = \mathrm{e}^{-\lambda} \sum_{k>r} \frac{(\lambda)^k}{k!}, \tag{6.1}$$

where $\lambda = (Kp*)$ represents the expectation of the number of events (declared to be a differentially expressed gene) and K is the number of genes tested. Each threshold c is determined by λ/K. FDR controlling procedure must be controlled at preassigned level $\alpha = 0.05$. Poisson cumulative probability table shows how to determine λ for given r and α. Recently, discrete p-value problem has been a focal issue in multiple testing procedure. Kendall's tau and Kendall's tau-type rank statitics generate discrete p-values, though we use normal approximation. The proposed FDR procedure utilizing the Chen-Stein method accounts for discrete p-values problems well.

The number of rejections (r)	λ	c	FDR (T)	FDR (R)	Storey (T)	Storey (R)
4	1.99	0.001	0.0106	0.0104	0.0071	0.0089
8	4.5	0.002	0.0208	0.0202	0.0136	0.0173
12	7.5	0.004	0.0316	0.0313	0.0214	0.0283
16	11	0.006	0.0428	0.0409	0.0294	0.0378

Table 2. Comparison of Kendall's tau statistics (T) with Kendall's tau-type rank statistics (R).

For each threshold c, all the FDR procedures should be controlled at 0.05. In Tab. 2, Kendall's tau-type rank statistics has smaller FDR than Kendall's tau. However, Storey's FDR has the opposite result to proposed FDR, since this does not take into account the fact that we have discrete p-values. We now turn to other p-values based on classical ANOVA framework under the conventional assumption of Gaussian distribution of gene expression levels. The same number of genes and cut-off points are used as in Tab. 2. A summary of these results of using normal theory under consideration is provided in Tab. 3. In contrast to the result given in Tab. 2, each FDR procedure has slightly larger values. It can be shown that the proposed FDR still has smaller values than the Storey FDR.

The number of rejections (r)	λ	c	Proposed FDR	Storey FDR
4	1.99	0.001	0.0127	0.0141
8	4.5	0.002	0.0235	0.0233
12	7.5	0.004	0.0314	0.0322
16	11	0.006	0.0409	0.0421

Table 3. Comparison of proposed FDR with Storey FDR using ANOVA.

The Chen-Stein method was utilized to produce an appropriate FDR procedure under fairly milder regularity conditions, which adjust for discrete p-values. Utilizing both concepts of Kendall's tau and linear rank statistics, we suggest a better multiple testing procedure based on more effective test statistics to analyze dependent genes with heterogeneity, even in a small sample. Two-stage FDR procedures considered in Kang [10] provide a multiple testing procedure under less stringent regularity conditions. Finally, we have concluded that Kendall's tau-type rank statistics has better performance over other conventional approaches in HDLSS data.

Appendix

Derivation of $V_0(T_{nk})$. We have

$$\begin{aligned} A &= E_0\left[\sum_S (a(R_{i'k}) - a(R_{ik}))\right]^2 \\ &= N \times E_0[(a(R_{i'k}) - a(R_{ik}))^2] \\ &= \frac{N}{n(n-1)} \sum_{1\leqslant i\neq i'\leqslant n} (a(i') - a(i))^2 \\ &= \frac{2N}{n-1} \sum_{i=1}^{n} (a(i') - \bar{a}_n)^2 \\ &= 2N \times A_n^2, \end{aligned}$$

$$
\begin{aligned}
B &= E_0\Big[\sum_{S_1}(a(R_{jk}) - a(R_{ik}))(a(R_{j'k}) - a(R_{i'k}))\Big] \\
&= \sum_{S_1} E_0[(a(R_{jk}) - a(R_{ik}))(a(R_{j'k}) - a(R_{i'k}))] \\
&= N_1 \times E_0[(a(R_{jk}) - a(R_{ik}))(a(R_{jk}) - a(R_{i'k})), \quad 1 \leqslant i \neq i' \neq j \leqslant n \\
&= N_1 \times \frac{1}{n(n-1)(n-2)} \sum_{1\leqslant i\neq j\neq k\leqslant n} (a(i) - a(j))(a(i) - a(k)) \\
&= N_1 \times \frac{1}{n(n-1)(n-2)} \sum_{1\leqslant i\neq j\neq k\leqslant n} (a^2(i) - a(i)a(j) - a(i)a(k) + a(j)a(k)) \\
&= N_1 \times \frac{1}{n(n-1)(n-2)} \sum_{1\leqslant i\neq j\neq k\leqslant n} (a^2(i) - a(i)a(j) - a(i)a(k) + a(j)a(k)) \\
&= N_1 \times \Big[\frac{1}{n}\sum_{i=1}^{n} a^2(i) - \frac{2}{n(n-1)} \sum_{1\leqslant i\neq j\leqslant n} a(i)a(j) \\
&\qquad\qquad + \frac{1}{n(n-1)} \sum_{1\leqslant j\neq k\leqslant n} a(j)a(k)\Big] \\
&= N_1 \times \Big[\frac{1}{n}\sum_{i=1}^{n} a^2(i) - \frac{1}{n(n-1)}\Big[\sum_{1\leqslant i\neq j\leqslant n} a(i)a(j)\Big]\Big] \\
&= N_1 \times \Big[\frac{1}{n}\sum_{i=1}^{n} a^2(i) - \frac{1}{n(n-1)}\Big(\sum_{i=1}^{n} a(i)\Big)^2 + \frac{1}{n(n-1)}\sum_{i=1}^{n} a^2(i)\Big] \\
&= N_1 \times \Big[\frac{1}{n}\Big(1 + \frac{1}{n-1}\Big)\sum_{i=1}^{n} a^2(i) - \frac{1}{n(n-1)}\Big(\sum_{i=1}^{n} a(i)\Big)^2\Big] \\
&= N_1 \times \frac{1}{n-1}\Big[\sum_{i=1} a(i) - \bar{a}_n]^2\Big] \\
&= N_1 \times A_n^2.
\end{aligned}
$$

Using the above result, we obtain

$$
\begin{aligned}
C &= E_0\Big[\sum_{S_2}(a(R_{jk}) - a(R_{ik}))(a(R_{j'k}) - a(R_{i'k}))\Big] \\
&= \sum_{S_2} E_0[(a(R_{jk}) - a(R_{ik}))(a(R_{j'k}) - a(R_{i'k}))] \\
&= N_2 \times E_0[(a(R_{jk}) - a(R_{ik}))(a(R_{i'k}) - a(R_{jk}))], \quad 1 \leqslant i \neq i' \neq j \leqslant n \\
&= -N_2 \times E_0[(a(R_{jk}) - a(R_{ik}))(a(R_{jk}) - a(R_{i'k}))] \\
&= -N_2 \times A_n^2,
\end{aligned}
$$

$$
\begin{aligned}
D &= E_0\left[\sum_{S_3}(a(R_{jk}) - a(R_{ik}))(a(R_{j'k}) - a(R_{i'k}))\right] \\
&= \sum_{S_3} E_0[(a(R_{jk}) - a(R_{ik}))(a(R_{j'k}) - a(R_{i'k}))] \\
&= N_3 \times E_0[(a(R_{jk}) - a(R_{ik}))(a(R_{j'k}) - a(R_{i'k}))], \quad 1 \leqslant i \neq i' \neq j \neq j' \leqslant n \\
&= N_3 \times \left[\frac{1}{n(n-1)(n-2)(n-3)}\right. \\
&\qquad\qquad \left.\times \sum_{1\leqslant i\neq i'\neq j\neq j'\leqslant n} [a(j)a(j') - a(j)a(i') - a(i)a(j') + a(i)a(i')]\right] \\
&= N_3 \times \frac{1}{n(n-1)}\left[\sum_{1\leqslant j\neq j'\leqslant n} a(j)a(j') - \sum_{1\leqslant j\neq i'\leqslant n} a(j)a(i')\right. \\
&\qquad\qquad \left. - \sum_{1\leqslant i\neq j'\leqslant n} a(i)a(j') + \sum_{1\leqslant i\neq i'\leqslant n} a(i)a(i')\right] = 0.
\end{aligned}
$$

Next we consider the computation of N_1 and N_2.

Suppose we have a multisample of DNA microarray data in which there are G (> 2) groups. Each group g with sample size n_g has a common design variate t_g^0, where $\sum_{g=1}^{G} n_g = n$. Now consider the subset of S_1, which is $S_{1s} = \{((i,j),(i',j'))\colon\ t_i < t_j$ and $t_{i'} < t_{j'};\ 1 \leqslant i < j \leqslant n,\ 1 \leqslant i' < j' \leqslant n\,(i = i'$ and $j \neq j')\}$. Note that $t_i^0 = t_i = t_{i'}$, $t_j = t_j^0$ and $t_{j'} = t_{j'}^0$, where $1 \leqslant i, j, j' \leqslant n$. The cardinality of S_1, N_1 should be twice the cardinality of this subset N_{11}, since S_1 consists of two disjoint subsets with the same cardinality. We can rewrite this subset as

$$\{((i,j),(i,j'))\colon\ t_i^0 < t_j^0 \neq t_{j'}^0;\ 1 \leqslant i < j \neq j' \leqslant n\}.$$

In Lobenhofer et al. dataset, we have 6 groups measured at 6 time points with each one having 8 observations:

$$N_{11} = \sum_{k=1}^{5} 8 \times \left\{\binom{6-k}{1} \times \binom{8}{2} + \binom{6-k}{2} \times 8^2\right\} = 13,600 = N_{12}.$$

Hence, the cardinality of S_1 is $27,200$ $(= N_{11} + N_{12})$.

Likewise, consider the subset of S_2, which is $S_{2s} = \{((i,j),(i',j'))\colon\ t_i^0 < t_{i'}^0 < t_{j'}^0;\ 1 \leqslant i < i' = j < j' \leqslant n\}$, with the corresponding cardinality N_{21}

$$N_{21} = \sum_{k=2}^{5} 8 \times \left\{\binom{k-1}{1} \times 8\binom{6-k}{1} \times 8\right\} = 8^3 \times 20 = 10,240 = N_{22}.$$

Hence, the cardinality of S_2 is $20,480$ $(= N_{21} + N_{22})$.

The result is the same as in the case that the t_i have a monotone nonincreasing pattern.

References

[1] *R. Arratia, L. Goldstein, L. Gordon*: Two moments suffice for Poisson approximations: The Chen-Stein method. Ann. Probab. *17* (1989), 9–25.

[2] *R. Arratia, L. Goldstein, L. Gordon*: Poisson approximation and the Chen-Stein method. Stat. Sci. *5* (1990), 403–424.

[3] *R. Arratia, L. Goldstein, L. Gordon*: Poisson approximation and the Chen-Stein method: Rejoinder. Stat. Sci. *5* (1990), 432–434.

[4] *Y. Benjamini, Y. Hochberg*: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc., Ser. B *57* (1995), 289–300.

[5] *L. H. Y. Chen*: Poisson approximation for dependent trials. Ann. Probab. *3* (1975), 534–545.

[6] *L. H. Y. Chen*: Poisson approximation and the Chen-Stein method: Comment. Stat. Sci. *5* (1990), 429–432.

[7] *L. Goldstein, M. Watterman*: Poisson, compound Poisson and process approximations for testing statistical significance in sequence comparisons. Bull. Math. Biol. *54* (1992), 785–812.

[8] *L. Goldstein, A. Xia*: Zero biasing and a discrete central limit theorem. Ann. Probab. *34* (2006), 1782–1806.

[9] *J. Jurečková, P. K. Sen*: Robust Statistical Procedures: Asymptotics and Interrelations. Wiley Series. Wiley & Sons, New York, 1996.

[10] *M. Kang*: Multiple testing in genome-wide studies. UNC Biostatistics Thesis. 2007.

[11] *M. G. Kendall*: A new measure of rank correlation. Biometrika *30* (1938), 81–93.

[12] *E. K. Lobenhofer, L. Bennett, P. L. Cable, L. Li, P. R. Bushel, C. A. Afshari*: Regulation of DNA replication fork genes by 17-estradiol. Molecular Endocrinology *16* (2002), 1215–1229.

[13] *S. N. Roy*: On a heuristic method of test construction and its use in multivariate analysis. J. Ann. Math. Stat. *24* (1953), 220–238.

[14] *S. K. Sarkar, C.-K. Chang*: The Simes method for multiple hypothesis testing with positively dependent test statistics. J. Am. Stat. Assoc. *92* (1997), 1601–1608.

[15] *P. K. Sen*: Estimates of the regression coefficients based on Kendall's tau. J. Am. Stat. Assoc. *63* (1968), 1379–1389.

[16] *P. K. Sen*: Robust statistical inference for high-dimensional data models with application to genomics. Aust. J. Stat. *35* (2006), 197–211.

[17] *P. K. Sen*: Kendall's tau in high-dimension genomics parsimony. IMS Collection *3* (2008), 250–265.

[18] *P. K. Sen, M.-T. Tsai, Y.-S. Jou*: High-dimension low sample size perspectives in constrained statistical inference: The SARSCoV RNA genome in illustration. J. Am. Stat. Assoc. *102* (2007), 686–694.

[19] *J. Storey*: A direct approach to false discovery rates. J. R. Stat. Soc., Ser. B *64* (2002), 479–498.

Authors' address: *M. Kang*, *P. K. Sen*, Department of Biostatistics, Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599–7420, U.S.A., email: moonsukang0223@gmail.com; pksen@bios.unc.edu.