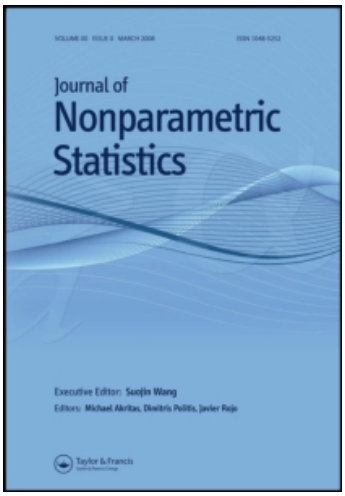

Journal of Nonparametric Statistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713645758

Rank tests in heteroscedastic multi-way HANOVA

Haiyan Wang [a]; Michael G. Akritas [b]
[a] Department of Statistics, Kansas State University, Manhattan, KS, USA [b] Department of Statistics, Pennsylvania State University, University Park, PA, USA

Online Publication Date: 01 August 2009

To cite this Article Wang, Haiyan and Akritas, Michael G.(2009)'Rank tests in heteroscedastic multi-way HANOVA',Journal of Nonparametric Statistics,21:6,663 — 681
To link to this Article: DOI: 10.1080/10485250902971757
URL: http://dx.doi.org/10.1080/10485250902971757

Journal of Nonparametric Statistics
Vol. 21, No. 6, August 2009, 663–681

Taylor & Francis
Taylor & Francis Group

Rank tests in heteroscedastic multi-way HANOVA

Haiyan Wang[a]* and Michael G. Akritas[b]

[a]*Department of Statistics, Kansas State University, 101 Dickens Hall, Manhattan, KS 66503, USA;*
[b]*Department of Statistics, Pennsylvania State University, 316 Thomas Building, University Park, PA 16802, USA*

(*Received 23 July 2008; final version received 14 April 2009*)

This article develops rank tests for the nonparametric main factor effects and interactions in multi-way high-dimensional analysis of variance when the cell distributions are completely unspecified. The design can be balanced or unbalanced with the cell sample sizes fixed or tending to infinity. An arbitrary number of factors and all types of ordinal data are allowed. This extends the use of rank methods to the Neymann–Scott and triangular array problems. The asymptotic distribution of the rank statistics is obtained by showing their asymptotic equivalence to corresponding expressions based on the asymptotic rank transform. Compared with test procedures based on the original observations, the proposed rank procedures are free of moment conditions, converge to their limiting distribution faster, and have better power when the underlying distributions are heavy tailed or skewed. These advantages are demonstrated by simulations and an application to a real data set.

Keywords: Neymann–Scott problem; nonparametric hypotheses; asymptotic distribution theory of quadratic form; projection method; rank tests

AMS Subject Classification: 62E20; 62G30; 62G10; 62J12; 62J10

1. Introduction

Advances in data gathering technologies have produced massive data sets. For example, DNA microarray technology enables the measurement of gene expressions of the entire human genome (∼30,000 genes) on a single glass slide. A particular example we will consider deals with gene expressions of *Arabidopsis thaiana* genes under multiple stress conditions using shoot or root tissue (see Section 4.1). In this application, the gene expression data show a high degree of skewness and variance heterogeneity. There are a large number of factor-level combinations and only two replications.

Such *high dimensional* designs have motivated the development of *high-dimensional analysis of variance* or HANOVA – a term introduced by Fan and Lin [1]. HANOVA test procedures are developed using asymptotic techniques under non-classical settings, which are characterised by the number of variables, p, approaching infinity. In high-dimensional designs, the p variables correspond to the factor-level combinations (also called groups or cells). It is shown in [2]

*Corresponding author. Email: hwang@ksu.edu

ISSN 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/10485250902971757
http://www.informaworld.com

that the F-statistic is not asymptotically valid in the unbalanced one-way design, even under homoscedasticity, unless the group sizes are also large. Thus, HANOVA methods strive to accommodate non-normality and/or heteroscedasticity with possibly small sample sizes. For additional literature dealing with HANOVA, see [3–8], and the references therein. See [8] for a brief discussion regarding the contributions of these papers. For related asymptotic results, see [9–13] for parametric inference and [14–19] for estimation under rectangular array framework when p grows at the same rate or at a power rate of n. See also [20] for an interesting geometric perspective of Neymann–Scott-type asymptotics.

This article develops a general theory of testing, using (mid-)rank statistics, in multi-factor high-dimensional heteroscedastic designs with possibly few unbalanced and non-normal replications. The presentation will be limited to test procedures for main effects and interactions up to the third order. Testing null hypotheses involving a small number of parameters is radically different from that involving a large number of parameters, so the two cases will be presented separately. Moreover, the presence of an arbitrary number of factors entails notational challenges that are met by adopting the formulation in [8]. The aforementioned paper, which develops test statistics based on the original observations, is also used in a fundamental way for establishing the asymptotic results for the proposed statistics. Indeed, the present asymptotic results are established by showing the asymptotic equivalence of the (mid-)rank statistics to the corresponding expressions in [8] that use the asymptotic rank transform [21]. While the paper by Wang and Akritas [8] pertains to the same general context as the present article, it is well known that test statistics based on the original observations are sensitive to outliers, converge slower to their asymptotic distributions, and are less powerful with non-normal distributions than rank statistics.

When dealing with rank statistics, the hypotheses to be tested need to be invariant under monotone transformations and are formulated in terms of the distribution functions [22]. To include both discrete and continuous distributions in the formulation, it is convenient to use the version of the distribution function which is the average of its left and right continuous versions: $F(x) = (1/2)[F^+(x) + F^-(x)]$. Empirical distribution functions are defined accordingly.

The rest of the article is organised as follows. Section 2 reviews the data representation, the nonparametric hypotheses, and presents the test statistics. Section 3 gives the main results about the (mid)-rank test statistics. Real data analysis and simulation results are presented in Section 4. A summary is given in Section 5 and the technical proofs are given in the appendices.

2. Data presentation and hypotheses

To represent data arising from factorial designs, it is customary to use different indices to represent the levels of different factors. Because we place no limit on the number of factors, this convention is not practical. Following the idea of [8] for parsimonious data presentation, the number of indices we use in presenting the proposed test statistics will depend on the hypothesis being tested. Thus, when testing for the main effects of a particular factor, say factor A, we will use i to enumerate the levels of factor A, j to enumerate the levels of all other factors combined, and k to enumerate the replications in cell (i, j). When testing for the interaction of factors A and B, the levels of these factors are indexed by i, j, respectively, k will enumerate the levels of all other factors combined, and l will index the replications in cell (i, j, k). Finally, when testing for the interaction of factors A, B and C, we will use the indices i, j, k, and l for the levels of A, B, C, and all other factors combined, and m for the replications. Lower case letters will be used to denote the number of levels of the different factors. For example, when testing for a three-factor interaction, the letters a, b, c, and d will be used to denote the number of levels of factors A, B, C, and all other factors combined. Cell sample sizes will be denoted by n suitably subscripted, and N will denote the total sample size.

This convention leads to the following notation. When testing for three-factor interactions, we define $\bar{Y}_{ijkl\cdot} = n_{ijkl}^{-1} \sum_{m=1}^{n_{ijkl}} Y_{ijklm}$, $\tilde{Y}_{i\cdots} = (bcd)^{-1} \sum_{j=1}^{b} \sum_{k=1}^{c} \sum_{l=1}^{d} \bar{Y}_{ijkl\cdot}$, $\tilde{Y}_{ij\cdots} = (cd)^{-1} \sum_{k=1}^{c} \sum_{l=1}^{d} \bar{Y}_{ijkl\cdot}$, $\tilde{Y}_{\cdot j\cdots} = (acd)^{-1} \sum_{i=1}^{a} \sum_{k=1}^{c} \sum_{l=1}^{d} \bar{Y}_{ijkl\cdot}$, and $\tilde{Y}_{\cdots\cdots} = (abcd)^{-1} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \sum_{l=1}^{d} \bar{Y}_{ijkl\cdot}$. Also, set $S^2_{ijkl,Y} = (n_{ijkl} - 1)^{-1} \sum_{m=1}^{n_{ijkl}} (Y_{ijklm} - \bar{Y}_{ijkl\cdot})^2$, with similar meaning when Y is replaced by e, R, or Z. Finally, define $n(a, b, c, d) = \min\{n_{ijkl}, i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, c, l = 1, \ldots, d\}$. Analogous versions of this notation are used when testing for two-factor interactions and for main effects.

We will consider testing the nonparametric hypotheses that are defined by decomposing the cumulative distribution functions (CDF) in a way similar to the decomposition of the means [22]. For example, in the case of a two-factor design, the cell CDFs $F_{ij}(x)$ can be decomposed as

$$F_{ij}(x) = M(x) + A_i(x) + B_j(x) + C_{ij}(x), \tag{1}$$

where $\sum_{i=1}^{a} A_i(x) = \sum_{j=1}^{b} B_j(x) = \sum_{i=1}^{a} C_{ij}(x) = \sum_{j=1}^{b} C_{ij}(x) = 0, \ \forall x$. The functions A_i, B_j, C_{ij} in Equation (1) are the fully nonparametric effects, and the fully nonparametric hypotheses specify that the corresponding effects are zero. The null hypothesis $H_0(C)$: all $C_{ij}(x) = 0$ means that F_{ij} is the mixture of a distribution which depends on i and a distribution which depends on j, with the mixing parameter being the same for all i, j. See also [22,23] for further discussion on the interpretation of the fully nonparametric effects and hypotheses.

The rank test statistics are obtained by replacing the original observations by their overall (mid-)ranks in the test statistics given in [8]. Since the aforementioned paper yields the distribution of corresponding expressions where the original observations are substituted by the asymptotic rank transformation [21], the asymptotic distribution of the rank statistics will be established by establishing their asymptotic equivalence to the corresponding expressions with the asymptotic rank transformation. The techniques for doing so are closely related to U-statistics.

3. Main results for rank statistics

3.1. *Testing for no main effects*

Let $X_{ijk} \sim F_{ij}$, for $k = 1, \ldots, n_{ij}$, denote the kth independent observation in cell (i, j), where $i = 1, \ldots, a$ enumerate the levels of factor A, and $j = 1, \ldots, b$ enumerate the levels of all factors other than A. Consider the decomposition given in Equation (1). In this section, we present two procedures for testing

$$H_0(A) : A_i(x) = 0, \quad \text{for all } i. \tag{2}$$

Both procedures require b to be large, but the first applies to the case where a is small and the other to the case where a is large. We remark that the classical case where both a, b are small has been studied in [23], whereas the case where a is large and b is small can be found in [6].

Let $c(x, y) = [I(x \le y) + I(x < y)]/2$, where $I(A)$ is the indicator function for the event A, and set $\widehat{H}(t) = N^{-1} \sum_i \sum_j \sum_k c(X_{ijk}, t)$. Then the (mid)-rank of an observation $X_{i_1 j_1 k_1}$ is $R_{i_1 j_1 k_1} = N\widehat{H}(X_{i_1 j_1 k_1}) + 0.5$. Also, let $H(x) = N^{-1} \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij} F_{ij}(x)$, and set $Y_{ijk} = H(X_{ijk})$. For the case that a is small, the test statistic is

$$Q_R(A) = N\mathbf{W}_R' \mathbf{C}_A' (\mathbf{C}_A \widehat{\mathbf{V}_R} \mathbf{C}_A')^{-1} \mathbf{C}_A \mathbf{W}_R, \tag{3}$$

where $\mathbf{W}_R = (\tilde{R}_{1\cdot\cdot}, \ldots, \tilde{R}_{a\cdot\cdot})'$, $\mathbf{C}_A$ is a contrast matrix with full row rank, and

$$\hat{\mathbf{V}}_R = \text{diag}(\hat{\eta}_1, \ldots, \hat{\eta}_a), \quad \text{with } \hat{\eta}_i = \frac{N}{b^2} \sum_{j=1}^{b} \frac{S^2_{ij,R}}{n_{ij}}.$$

For the case that a is large, the test statistic is

$$F_R(A) = \frac{\text{MST}_A}{\text{MSE}}, \tag{4}$$

where

$$\text{MST}_A = \frac{1}{a-1}\sum_{i=1}^{a}\sum_{j=1}^{b}(\tilde{R}_{i\cdot\cdot} - \tilde{R}_{\cdot\cdot\cdot})^2, \quad \text{MSE} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{S^2_{ij,R}}{n_{ij}}.$$

Note that the MST_A and MSE are different from traditional mean squares because of our use of unweighted means (cf. [24, pp. 220–222]); moreover, the summation over the replications is not included in MST_A.

THEOREM 3.1 *Let* $\sigma^2_{ij} = \text{Var}(Y_{ijk})$.

(1) *When a is small and b is large, let $Q_R(A)$ be the statistic given in Equation* (3)*, and assume that for all i,*

$$\lim_{b\to\infty}\frac{N}{b^2}\sum_{j=1}^{b}\frac{1}{n_{ij}} > 0, \quad \left(\frac{1}{b}\sum_{j=1}^{b}\frac{\sigma^2_{ij}}{n_{ij}}\right)^{-2}\frac{1}{b^2}\sum_{j=1}^{b}\frac{1}{n^3_{ij}} \to 0. \tag{5}$$

Then under $H_0(A)$ given in Equation (2)*,*

$$Q_R(A) \xrightarrow{d} \chi^2_{\text{rank}(\mathbf{C}_A)} \text{ as } b \to \infty.$$

(2) *When both a and b are large, let $F_R(A)$ be the statistic given in Equation* (4)*. Further, define $\tau_{1,A}$ and σ^2_A as*

$$\tau_{1,A} = \frac{2}{ab^2}\sum_{i=1}^{a}\left(\sum_{j=1}^{b}\frac{\sigma^2_{ij}}{n_{ij}}\right)^2, \quad \sigma^2_A = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma^2_{ij}}{n_{ij}}.$$

Then under $H_0(A)$, as $a, b \to \infty$,

$$\frac{\sqrt{a}(F_R(A)-1)}{\tau_A} \xrightarrow{d} N(0,1), \text{ where } \tau_A = \frac{\sqrt{\tau_{1,A}}}{\sigma^2_A},$$

regardless of whether the $n_{ij} \geq 2$ stay fixed, or tend to ∞ with a or b.

Remark 1 The results in Theorem 3.1 are of similar form as the corresponding results in [8], except that the present results hold without any moment conditions. Thus, the present results are of wider applicability in addition to the fact that rank statistics converge faster to their limiting distributions.

Remark 2 A consistent estimate of the asymptotic variance is obtained by replacing the σ^2_{ij} by the cell sample variance divided by N^2, $S^2_{ij,R}/N^2$.

The two remarks above also apply to the tests for interaction effects in the next two subsections.

3.2. *Testing for no two-way interaction effects*

Let $X_{ijkm} \sim F_{ijk}$, for $m = 1, \ldots, n_{ijk}$, denote the mth observation in cell (i, j, k), where $i = 1, \ldots, a$, $j = 1, \ldots, b$ enumerate the levels of factors A and B, respectively, and $k = 1, \ldots, c$ enumerates the levels of all factors other than A and B. Consider the decomposition

$$F_{ijk} = M + A_i + B_j + C_k + (AB)_{ij} + (AC)_{ik} + (BC)_{jk} + (ABC)_{ijk},$$

which is unique under the restrictions $\sum_{i=1}^{a} A_i = \sum_{j=1}^{b} B_j = \sum_{k=1}^{c} C_k = \sum_{i=1}^{a}(AB)_{ij} = \sum_{i=1}^{a}(AC)_{ik} = 0, \sum_{j=1}^{b}(AB)_{ij} = \sum_{j=1}^{b}(BC)_{jk} = \sum_{k=1}^{c}(AC)_{ik} = \sum_{k=1}^{c}(BC)_{jk} = 0$ and $\sum_{i=1}^{a}(ABC)_{ijk} = \sum_{j=1}^{b}(ABC)_{ijk} = \sum_{k=1}^{c}(ABC)_{ijk} = 0$. In this subsection, we present three procedures for testing

$$H_0(AB) : (AB)_{ij} = 0 \quad \text{for all } i, j. \tag{6}$$

The procedures pertain to the cases where

- a and b are small while c is large,
- a is large, b is small, and c can be small or large,
- a and b are large and c can be small or large.

The classical case where all factors have a small number of levels is treated in [23]. Set $\hat{H}(t) = N^{-1}\sum_i \sum_j \sum_k \sum_m c(X_{ijkm}, t)$. Then the (mid)-rank of an observation $X_{i_1 j_1 k_1 m_1}$ is $R_{i_1 j_1 k_1 m_1} = N\hat{H}(X_{i_1 j_1 k_1 m_1}) + 0.5$. Also, let $H(x) = N^{-1}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} n_{ijk} F_{ijk}(x)$, and set $Y_{ijkm} = H(X_{ijkm})$. For the case that both a and b are small and c is large, the test statistic is

$$Q_R(AB) = N\mathbf{W}_R'\mathbf{C}_{AB}'(\mathbf{C}_{AB}\hat{\mathbf{V}}\mathbf{C}_{AB}')^{-1}\mathbf{C}_{AB}\mathbf{W}_R, \tag{7}$$

where $\mathbf{W}_R = (\tilde{R}_{11..}, \ldots, \tilde{R}_{1a..}, \ldots, \tilde{R}_{a1..}, \ldots, \tilde{R}_{ab..})'$, $\hat{\mathbf{V}} = \text{diag}\{\hat{\eta}_{11}, \ldots, \hat{\eta}_{1b}, \ldots, \hat{\eta}_{b1}, \ldots, \hat{\eta}_{ab}\}$ with $\hat{\eta}_{ij} = (N/c^2)\sum_{k=1}^{c} S_{ijk,R}^2/n_{ijk}$ and $S_{ijk,R}^2 = (n_{ijk}-1)^{-1}\sum_{m=1}^{n_{ijk}}(R_{ijkm} - \bar{R}_{ijk.})^2$, and $\mathbf{C}_{AB} = \mathbf{M}_a \otimes \mathbf{M}_b$, with $\mathbf{M}_b = (\mathbf{1}_{b-1} \mid -I_{b-1})$.

When a is large and b is small, regardless of whether c is small or large, the test statistic is $\sqrt{a}(F_R(AB) - 1)$, where

$$\begin{aligned} F_R(AB) &= \frac{\text{MST}_{AB}}{\text{MSE}}, \\ \text{MST}_{AB} &= \frac{c}{(a-1)(b-1)}\sum_{i,j}(\tilde{R}_{ij..} - \tilde{R}_{i...} - \tilde{R}_{.j..} + \tilde{R}_{....})^2, \\ \text{MSE} &= \frac{1}{abc}\sum_{i,j,k}\frac{S_{ijk,R}^2}{n_{ijk}}. \end{aligned} \tag{8}$$

When both a and b are large, regardless of whether c is small or large, the test statistic is $\sqrt{ab}(F_R(AB) - 1)$, where $F_R(AB)$ is given in Equation (8). The asymptotic distribution of the test statistics above are given below.

THEOREM 3.2 *Let* $\sigma_{ijk}^2 = \text{Var}(Y_{ijkm})$.

(1) *When both a and b are small, let* $Q_R(AB)$ *be the statistic given in Equation* (7). *Assume that for all* i, j,

$$\lim_c \frac{N}{c^2}\sum_{k=1}^{c}\frac{1}{n_{ijk}} > 0, \quad \left(\frac{1}{c}\sum_{k=1}^{c}\frac{\sigma_{ijk}^2}{n_{ijk}}\right)^{-2}\frac{1}{c^2}\sum_{k=1}^{c}\frac{1}{n_{ijk}^3} \to 0.$$

Then under $H_0(AB)$: all $(AB)_{ij} = 0$,

$$Q_R(AB) \xrightarrow{d} \chi^2_{(a-1)\times(b-1)}, \quad \text{as } c \to \infty.$$

(2) *When a is large and b is small, let $F_R(AB)$ be the statistic given in Equation* (8). *Let $\tau_{AB} = \sqrt{\tau_1^2 + \tau_2^2 + \tau_3^2}/\sigma_{AB}^2$, where*

$$\tau_1^2 = \frac{2}{b^2c^2a}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\frac{n_{ijk}^{-2}\sigma_{ijk}^4}{(n_{ijk}-1)}, \quad \tau_2^2 = \frac{2(b-2)}{ab(b-1)^2c^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\sum_{k=1}^{c}\frac{\sigma_{ijk}^2}{n_{ijk}}\right)^2, \tag{9}$$

$$\tau_3^2 = \frac{2}{ab^2(b-1)^2c^2}\sum_{i=1}^{a}\left(\sum_{j=1}^{b}\sum_{k=1}^{c}\frac{\sigma_{ijk}^2}{n_{ijk}}\right)^2, \quad \sigma_{AB}^2 = \frac{1}{abc}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\frac{\sigma_{ijk}^2}{n_{ijk}}. \tag{10}$$

Then under $H_0(AB)$, as $a \to \infty$, regardless of whether c and $n_{ijk} \geq 2$ go to ∞ or stay fixed,

$$\frac{\sqrt{a}(F_R(AB)-1)}{\tau_{AB}} \xrightarrow{d} N(0,1).$$

(3) *When both a and b are large, let $F_R(AB)$ be the statistic given in Equation* (8). *Then under $H_0(AB)$, as $a, b \to \infty$, regardless of whether c and $n_{ijk} \geq 2$ tend to ∞ or stay fixed,*

$$\frac{\sqrt{ab}(F_R(AB)-1)}{\tau_{AB2}} \xrightarrow{d} N(0,1), \quad \text{where } \tau_{AB2} = \frac{\sqrt{b\tau_1^2 + (b-1)^2/(b-2)\tau_2^2}}{\sigma_{AB}^2},$$

and τ_1^2, τ_2^2 and σ_{AB}^2 are defined in Equations (9) *and* (10).

3.3. *Testing for no three-way interaction effects*

Let $X_{ijklm} \sim F_{ijkl}(x)$, for $m = 1, \ldots, n_{ijkl}$, denote the mth observation in cell (i, j, k, l), where $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, c$ enumerate the levels of factors A, B, C, respectively, and $l = 1, \ldots, d$ enumerates the levels of all factors other than A, B or C. Consider the decomposition

$$\begin{aligned} F_{ijkl} = {} & M + A_i + B_j + C_k + D_l + (AB)_{ij} + (AC)_{ik} + (AD)_{il} + (BC)_{jk} + (BD)_{jl} \\ & + (CD)_{kl} + (ABC)_{ijk} + (ABD)_{ijl} + (BCD)_{jkl} + (ACD)_{ikl} + (ABCD)_{ijkl}, \\ & i = 1, \ldots, r,\ j = 1, \ldots, c, k = 1, \ldots, c, l = 1, \ldots, d, \end{aligned}$$

which is unique under the restrictions $\sum_{i=1}^{a} A_i = \sum_{j=1}^{b} B_j = \sum_{k=1}^{c} C_k = \sum_{l=1}^{d} D_l = 0$, $\sum_{i=1}^{a}(AB)_{ij} = \sum_{j=1}^{b}(AB)_{ij} = 0$, and similar constraints for all the other two-way interaction effects, $\sum_{i=1}^{a}(ABC)_{ijk} = \sum_{j=1}^{b}(ABC)_{ijk} = \sum_{k=1}^{c}(ABC)_{ijk} = 0$, and similar constraints for all other three-way interaction effects, $\sum_{i=1}^{a}(ABCD)_{ijkl} = \sum_{j=1}^{b}(ABCD)_{ijkl} = \sum_{k=1}^{c}(ABCD)_{ijkl} = \sum_{l=1}^{d}(ABCD)_{ijkl} = 0$.

Set $\hat{H}(t) = N^{-1}\sum_i\sum_j\sum_k\sum_l\sum_m c(X_{ijklm}, t)$. Then the (mid)-rank of $X_{i_1j_1k_1l_1m_1}$ is $R_{i_1j_1k_1l_1m_1} = N\hat{H}(X_{i_1j_1k_1l_1m_1}) + 0.5$. Also, let $H(x) = N^{-1}\sum_{i,j,k,l} n_{ijkl}F_{ijkl}(x)$, and set $Y_{ijklm} = H(X_{ijklm})$.

In this subsection, we present four procedures for testing

$$H_0(ABC) : (ABC)_{ijk} = 0 \quad \text{for all } i, j, k. \tag{11}$$

(1) When a, b, and c are all small and d is large, the test statistic for testing $H_0(ABC)$ is given by

$$Q_R(ABC) = N\mathbf{W}_R'\mathbf{C}_{ABC}'(\mathbf{C}_{ABC}\hat{\mathbf{V}}\mathbf{C}_{ABC}')^{-1}\mathbf{C}_{ABC}\mathbf{W}_R, \tag{12}$$

where $\mathbf{W}_R = (\tilde{R}_{111..}, \ldots, \tilde{R}_{11c..}, \tilde{R}_{121..}, \ldots, \tilde{R}_{12c..}, \ldots, \tilde{R}_{ab1..}, \ldots, \tilde{R}_{abc..})'$, $\mathbf{C}_{ABC} = \mathbf{M}_a \otimes \mathbf{M}_b \otimes \mathbf{M}_c$ with $\mathbf{M}_a = (\mathbf{1}_{a-1} | -I_{a-1})$, and $\hat{\mathbf{V}} = \text{diag}\{\hat{\eta}_{111}, \ldots, \hat{\eta}_{11c}, \hat{\eta}_{121}, \ldots, \hat{\eta}_{12c}, \ldots, \hat{\eta}_{ab1}, \ldots, \hat{\eta}_{abc}\}$ with $\hat{\eta}_{ijk} = (N/d^2)\sum_{l=1}^{d}(S^2_{ijkl,R}/n_{ijkl})$ and $S^2_{ijkl,R} = (n_{ijkl}-1)^{-1}\sum_{m=1}^{n_{ijkl}}(R_{ijklm} - \bar{R}_{ijkl.})^2$.

(2) When two of the three factors have a large number of levels, say a, b are large, and c is small, regardless of whether d is small or large, the test statistic for $H_0(ABC)$ is $\sqrt{ab}(F_R(ABC) - 1)$, where

$$F_R(ABC) = \frac{\text{MST}_{ABC}}{\text{MSE}}, \tag{13}$$

with $\text{MST}_{ABC} = [(a-1)(b-1)(c-1)]^{-1}\sum_{i,j,k,l}(\tilde{R}_{ijk..} - \tilde{R}_{ij...} - \tilde{R}_{i.k.} - \tilde{R}_{.jk.} + \tilde{R}_{i....} + \tilde{R}_{.j...} + \tilde{R}_{..k.} - \tilde{R}_{.....})^2$ and $\text{MSE} = \frac{1}{abcd}\sum_{i,j,k,l}(S^2_{ijkl,R}/n_{ijkl})$.

(3) The test statistic for $H_0(ABC)$ when a is large, b, c small, and d either large or small is $\sqrt{a}(F_R(ABC) - 1)$, where $F_R(ABC)$ is defined in Equation (13).

(4) When a, b, and c are all large, the test statistic for $H_0(ABC)$ is $\sqrt{abc}(F_R(ABC) - 1)$, where $F_R(ABC)$ is defined in Equation (13), regardless of whether d is small or large.

The classical case where all factors have a small number of levels is treated in [23]. Next theorem gives the asymptotic results of above test statistics.

THEOREM 3.3 *Let* $\sigma^2_{ijkl} = \text{Var}(Y_{ijklm})$.

(1) *When a, b, c are all small, let $Q_R(ABC)$ be the statistic given in Equation* (12). *Assume that for all i, j, k,*

$$\lim_d \frac{N}{d^2}\sum_{l=1}^{d}\frac{1}{n_{ijkl}} > 0, \quad \left(\frac{1}{d}\sum_{l=1}^{d}\frac{\sigma^2_{ijkl}}{n_{ijkl}}\right)^{-2}\frac{1}{d^2}\sum_{l=1}^{d}\frac{1}{n^3_{ijkl}} \to 0.$$

Then under $H_0(ABC)$: all $(ABC)_{ijk} = 0$,

$$Q_R(ABC) \xrightarrow{d} \chi^2_{(a-1)\times(b-1)\times(c-1)}, \quad \text{as } d \to \infty.$$

(2) *When a, c are large and b is small, let $F_R(ABC)$ be the statistic given in Equation* (13). *Further let τ_4, τ_5, τ_6, and σ^2_{ABC} be as defined in Equations* (14) *and* (15):

$$\tau_4 = \frac{2}{ac^2b^2d^2}\sum_{i,j,k,l}\frac{n^{-2}_{ijkl}\sigma^4_{ijkl}}{(n_{ijkl}-1)}, \quad \tau_5 = \frac{2(b-2)}{ac^2d^2b(b-1)^2}\sum_{i,j,k}\left(\sum_{l=1}^{d}\frac{\sigma^2_{ijkl}}{n_{ijkl}}\right)^2, \tag{14}$$

$$\tau_6 = \frac{2}{ac^2b^2d^2(b-1)^2}\sum_{i,k}\left(\sum_{j,l}\frac{\sigma^2_{ijkl}}{n_{ijkl}}\right)^2, \quad \sigma^2_{ABC} = \frac{1}{abcd}\sum_{i,j,k,l}\frac{\sigma^2_{ijkl}}{n_{ijkl}}. \tag{15}$$

Then under $H_0(ABC)$, as $a, c \to \infty$ while b remains fixed, regardless of whether d and $n_{ijkl} \geq 2$ tend to ∞ or remain fixed,

$$\frac{\sqrt{ac}(F_R(ABC) - 1)}{\tau_{ABC2}} \xrightarrow{d} N(0,1), \quad \textit{where } \tau_{ABC2} = \frac{\sqrt{c(\tau_4 + \tau_5 + \tau_6)}}{\sigma^2_{ABC}}.$$

(3) *When a is large and b, c are small, let $F_R(ABC)$ be the statistic given in Equation* (13) *and let τ_4, τ_5, τ_6, and σ^2_{ABC} be given in Equations* (14) *and* (15) *and τ_7, τ_8 be as defined below:*

$$\tau_7 = \frac{2n^2(a,b,c,d)}{ac^2d^2(b-1)^2(c-1)^2} \sum_{i,j} \sum_{k\neq k'}^{c} \sum_{l,l'} \frac{\sigma^2_{ijkl}}{n_{ijkl}} \frac{\sigma^2_{ijk'l'}}{n_{ijk'l'}},$$

$$\tau_8 = \frac{2n^2(a,b,c,d)}{ab^2c^2d^2(b-1)^2(c-1)^2} \sum_{i=1}^{a} \sum_{j,j'} \sum_{k\neq k'}^{c} \sum_{l,l'} \frac{\sigma^2_{ijkl}}{n_{ijkl}} \frac{\sigma^2_{ij'k'l'}}{n_{ij'k'l'}}.$$

Then under $H_0(ABC)$, as $a \to \infty$ regardless of whether d and $n_{ijkl} \geq 2$ tend to ∞ or remain fixed,

$$\frac{\sqrt{a}(F_R(ABC)-1)}{\tau_{ABC1}} \xrightarrow{d} N(0,1), \quad \text{where } \tau_{ABC1} = \frac{\sqrt{\tau_4+\tau_5+\tau_6+\tau_7+\tau_8}}{\sigma^2_{ABC}}.$$

(4) *When a, b, c are all large, let $F_R(ABC)$ be the statistic given in Equation* (13). *Then under $H_0(ABC)$, as $a, b, c \to \infty$, regardless of whether d and $n_{ijkl} \geq 2$ tend to ∞ or remain fixed,*

$$\frac{\sqrt{abc}(F_R(ABC)-1)}{\tau_{ABC3}} \xrightarrow{d} N(0,1), \quad \text{where } \tau_{ABC3} = \frac{\sqrt{bc\tau_4 + c(b-1)^2/(b-2)\tau_5}}{\sigma^2_{ABC}},$$

where τ_4, τ_5 and σ^2_{ABC} are defined in Equations (14) *and* (15).

4. Numerical results

A real data application is given in Section 4.1. The rest of this section is devoted to simulations for designs with three factors. Among the available test procedures for such designs are the class of generalised linear models (GLM) for the exponential family, the ANOVA F-tests, and the NP.org tests from [8]. The proposed NP.rank tests are compared with GLM and NP.org for binary and count data in Section 4.2, and with the ANOVA F-test and NP.org for continuous data in Section 4.3.

4.1. *Analysis of stress response gene expression data*

Arabidopsis thaliana is a model species for plant genome research. Its entire genome contains 25,000 genes, a relatively small number compared with other plants. As scientists can better understand how other organisms behave genetically by studying this plant, NSF devoted a tremendous amount of resources ($43.8 million over the period 2001–2005) aiming to identify how each of the plant's 25,000 genes function. Numerous experiments were done under different conditions and gene expression data are available from various websites. Nevertheless, the number of replicates is typically very small due to the large number of experimental conditions (e.g. many biotic, abiotic stresses, and pathogen infections). Therefore, the proposed test procedures provide an effective statistical tool for the identification of genes in biological processes under different stresses or pathogen infections.

Here, we illustrate our methods by analysis of a single gene, $PLD\alpha 3$ (At5g25370). Gene expression data for wild-type *Arabidopsis thaliana* shoot and root tissues under various abiotic stresses (cold/freezing, osmotic, salt, drought, genotoxic, oxidative, UV-B, wounding, heat) and corresponding controls were produced at several time points after the stresses. The raw data are

available at the e-Northerns of Bio-Array Resource for Arabidopsis Functional Genomics database [25]. Due to the small size of the plant, multiple plants were often used to produce the material for a microarray chip. Therefore, the data at different time points were from different plants and they can be treated as independent. For each gene, the effect of three factors and their interactions are to be considered: stress category (nine stresses and corresponding controls), time course (six levels, 0.5, 1, 3, 6, 12, and 24 h), and tissue (two levels, shoot or root). Two replications are available for each factor-level combination.

The overall median expression level at each time point and median expression from the shoot and root tissues at each time point are given in the left panel of Figure 1. Median gene expression levels from multiple stresses for shoot and root tissues are given in the right panel of Figure 1. It can be seen from the plots that the overall median expression level is high at 6 h after the stresses, which suggests a time effect. The median expression for the root tissue reaches the lowest at 12 h after stresses, whereas that for the shoot tissue reaches the highest level at 12 h after stresses. This suggests that there is an interaction effect between tissue and time. Similarly, the right panel of Figure 1 suggests a stress $\times$ tissue interaction effect.

We applied the proposed rank tests (NP.rank), the tests based on original observations of [8] (NP.org), and the tests from traditional linear model or ANOVA based on type III sum of squares. The p-values are given in Table 1. Both NP.rank and NP.org found a significant time effect. The p-value from ANOVA, while being significant at the 0.05 level, is orders of magnitude larger than that from the other two tests. Note that when all the genes are considered for testing on a gene-by-gene basis, false discovery rate control or Bonferroni correction need to be used to adjust for multiple comparisons. Even with the most conservative Bonferroni correction for 25,000 genes ($0.05/25000 = 2 \times 10^{-6}$), both NP.rank and NP.org yield significant time effect but the test using ANOVA would fail. Similar conclusions apply to the test for tissue and time interaction effect. For the stress and tissue interaction effect, NP.rank test is significant at 0.05 level and may still be significant after a false discovery rate control, but ANOVA and NP.org would not identify this significant interaction even without adjusting for multiple comparisons. For the rest of the effects, all three tests produce comparable results.

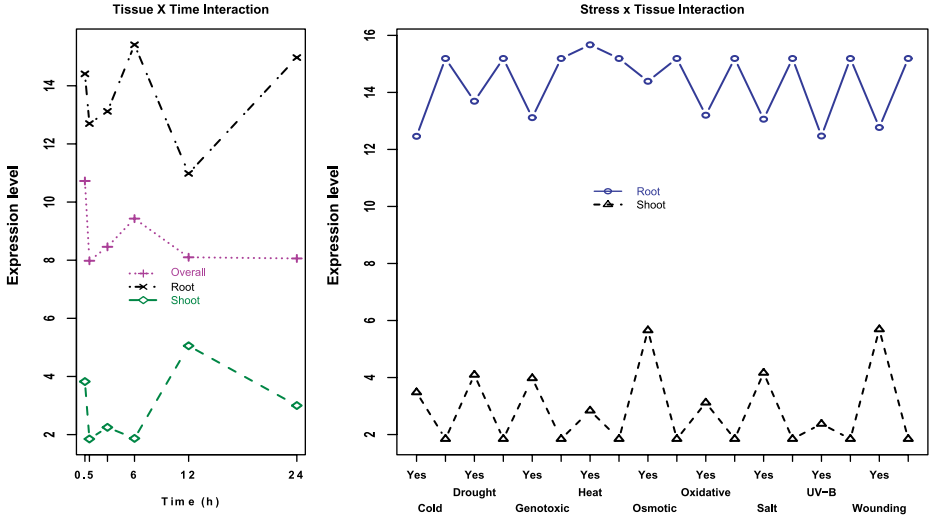

Figure 1. Left panel: Median gene expression level of $PLD\alpha 3$ at different time points for shoot and root tissues. Right panel: Median gene expression level of $PLD\alpha 3$ from multiple abiotic stresses for shoot and root tissues. The ticks labelled with 'yes' corresponds to the stress category listed underneath; the empty ticks beside 'yes' are for the corresponding control for each stress.

Table 1. p-value for all effects for gene $PLD\alpha 3$ (At5g25370).

Effect	ANOVA	NP.org	NP.rank
Stress	0.552	0.888	0.458
Tissue	0.000	0.000	0.000
Time	0.011	1.47×10^{-9}	2.18×10^{-7}
Stress–tissue	0.869	0.095	9.08×10^{-5}
Stress–time	0.322	0.959	0.983
Time–tissue	0.132	0.000	4.04×10^{-7}
Time–tissue–stress	0.988	0.971	0.997

In summary, for gene $PLD\alpha 3$, the proposed rank tests have better or comparable power with NP.org in identifying significant effects. Both are better than the ANOVA based on type III sum of squares. This happens because the number of replicates being two is too small while the traditional linear models require constant variance and large sample sizes for its validity [26]. On the other hand, the proposed tests are developed under the correct asymptotic setting (small sample size, large number of levels, and allow heteroscedasticity) for this data set.

4.2. *Simulations with count and binary data*

In [8], simulations using binary and count data are used for comparing their test (NP.org) with the deviance test from GLM. We performed simulations using exactly the same settings, and included the proposed NP.rank test procedures in the comparisons. Specifically, we consider three factors A, B, and C in a model with 20, 2, and 20 levels, respectively; 77.5% of the cell sample sizes are 4, 15% are 5, and 7.5% are 6.

For binary data the NP.rank tests produced exactly the same results as those based on the original observations. This outcome can also be justified theoretically using the fact that, since there are only two different values, the mid-rank transformation is a linear transformation. Thus, we omit the presentation of these results and refer to [8] for commentaries.

The count data are generated from a Poisson distribution with mean μ_{ijk}, $i = 1, \ldots, 20$, $j = 1, 2, k = 1, \ldots, 20$, where μ_{ijk} are given by (P1): $\mu_{ijk} = 1 + i\tau/a$, or (P2): $\mu_{ijk} = 1 + j\tau/b$, with $\tau = 0, 0.2, 0.4, 0.6, 0.8$. Non-zero values of τ under $P1$ correspond to alternative hypotheses for factor A effect and null hypotheses for all other effects. Under $P2$, non-zero values of τ correspond to alternative hypotheses for factor B effect and null hypotheses for all other effects. The type I error and power estimates reported here are based on 1000 runs.

Table 2 reports the empirical type I error rates at level $\alpha = 0.05$ for testing the hypotheses of no main effect of factor A, B, and no AB, AC, and ABC interaction effects. The type I error rates of the NP.rank tests are close to the nominal α levels, and comparable with those achieved by the NP.org test. However, the deviance tests become increasingly liberal as the number of parameters involved in the null hypothesis increase; in particular, GLM's type I error rate for the hypothesis of no ABC interaction effects is unacceptably high.

The estimated power for the test of no main factor A and factor B effects is given in Table 3. Here, we see that the deviance test has very little or no power to detect alternatives. The achieved powers of NP.rank and NP.org are similar and both increase to 1 as τ increases.

4.3. *Simulations with continuous data*

In this section, we compare the proposed NP.rank tests with the NP.org tests of [8] and the classical ANOVA F-tests based on type III sum of squares. The simulations are based on 2000 runs with data generated from the normal, log-normal and Cauchy distributions. In this section,

Table 2. Type I error estimate at level 0.05 for count data.

	$\tau = 0$			$\tau = 0.2$			$\tau = 0.4$			$\tau = 0.6$		
Effect	GLM	NP.org	NP.rank	GLM	NP.org	NP.rank	GLM	NP.org	NP.rank	GLM	NP.org	NP.rank
$X_{ijkm} \sim \text{Poisson}\left(1 + \frac{i\tau}{a}\right)$												
A	0.099	0.070	0.069									
B	0.076	0.038	0.033	0.077	0.048	0.044	0.068	0.055	0.050	0.074	0.056	0.054
AC	0.238	0.059	0.068	0.210	0.068	0.069	0.161	0.067	0.063	0.163	0.068	0.073
AB	0.098	0.072	0.075	0.096	0.068	0.078	0.082	0.078	0.081	0.081	0.073	0.063
ABC	0.402	0.065	0.063	0.334	0.063	0.064	0.262	0.068	0.063	0.249	0.061	0.068
$X_{ijkm} \sim \text{Poisson}\left(1 + \frac{j\tau}{b}\right)$												
A	0.099	0.070	0.069	0.076	0.046	0.045	0.077	0.067	0.066	0.057	0.077	0.078
B	0.076	0.038	0.033									
AC	0.238	0.059	0.068	0.213	0.061	0.056	0.162	0.053	0.046	0.168	0.067	0.066
AB	0.098	0.072	0.075	0.084	0.052	0.056	0.081	0.078	0.081	0.073	0.072	0.078
ABC	0.402	0.065	0.063	0.320	0.069	0.076	0.245	0.054	0.057	0.221	0.061	0.068

Table 3. Estimated power at level $\alpha = 0.05$ for no main factor A and B effect for count data.

	$H_0(A)$				$H_0(B)$			
	$\text{Poisson}\left(1 + \frac{i\tau}{a}\right)$				$\text{Poisson}\left(1 + \frac{j\tau}{b}\right)$			
	τ	GLM	NP.org	NP.rank	τ	GLM	NP.org	NP.rank
Poisson	0.00	0.099	0.070	0.069	0.00	0.076	0.038	0.033
	0.20	0.069	0.469	0.443	0.10	0.068	0.301	0.280
	0.40	0.096	0.990	0.983	0.20	0.087	0.779	0.760
	0.60	0.111	1.000	1.000	0.40	0.077	1.000	0.999
	0.80	0.161	1.000	1.000	0.60	0.079	1.000	1.000
					0.80	0.103	1.000	1.000

we use normal(c_1, c_2) to denote a normal distribution with mean c_1 and standard deviation c_2, log-normal(c_1, c_2) for a log-normal distribution arising by exponentiating a normal(c_1, c_2) random variable, and Cauchy(c_1, c_2) for a Cauchy distribution with location and scale parameters c_1, c_2, respectively. We only report some typical results in the unbalanced heteroscedastic case. For other cases, we refer to [27].

Three factors are included in the simulations. For comparing the type I error rates, the number of levels of factor A takes values, $a = 20$, 30, and 50, factors B has $b = 2$ levels, and factor C has $c = 20$ levels. The type I error rates, at nominal level $\alpha = 0.05$, for testing $H_0(A)$, $H_0(B)$, $H_0(AC)$, $H_0(AB)$, and $H_0(ABC)$ are reported in Table 4 when the data are generated with $c_1 = 0$ and $c_2 = 4jk/(bc)$ from the aforementioned three distributions. For power comparisons, the number of levels of factors A, B, and C are taken to be $a = 20$, $b = 2$, and $c = 20$, respectively. The power estimates for testing $H_0(B)$ are reported in Table 5 when data are generated from above three distributions with $c_1 = j\tau/(4b)$ and $c_2 = 4jk/(bc)$. The power for $H_0(AC)$ is given in Table 6 when the data are generated with $c_1 = 4ik\tau/(ac)$ and $c_2 = 4j/b$. The sample sizes range from 4 to 6 for $a = 20$ and from 4 to 7 for $a = 30$ or $a = 50$.

The type I error estimate of ANOVA F-test depends on the pattern of heteroscedasticity, the distribution of the data, and whether the number of parameters involved is large or not. For example, when $a = 50$, the ANOVA F-test for no main effect of factor B yielded type I error 0.059 for the

Table 4. Estimated level, unbalanced heteroscedastic case, $\alpha = 0.05$, $b = 2$, $c = 20$.

		Normal(0, $4jk/(bc)$)			Log-normal(0, $4jk$/(bc))			Cauchy(0, $4jk$/(bc)		
a	H_0	F	NP.org	NP.rank	F	NP.org	NP.rank	F	NP.org	NP.rank
20	A	0.050	0.069	0.068	0.080	0.003	0.068	0.020	0.031	0.059
	AB	0.047	0.067	0.070	0.027	0.002	0.058	0.027	0.038	0.057
	AC	0.112	0.073	0.061	0.147	0.005	0.067	0.038	0.091	0.068
	ABC	0.109	0.074	0.072	0.112	0.006	0.055	0.156	0.086	0.058
	B	0.049	0.049	0.058	0.621	0.634	0.057	0.022	0.020	0.057
30	A	0.052	0.066	0.071	0.099	0.003	0.061	0.020	0.031	0.067
	AB	0.060	0.065	0.062	0.035	0.003	0.074	0.036	0.030	0.058
	AC	0.107	0.061	0.058	0.385	0.001	0.058	0.242	0.103	0.061
	ABC	0.138	0.062	0.055	0.439	0.001	0.059	0.435	0.095	0.060
	B	0.053	0.048	0.054	0.688	0.677	0.064	0.026	0.025	0.056
50	A	0.044	0.055	0.056	0.034	0.001	0.062	0.027	0.031	0.060
	AB	0.044	0.058	0.064	0.034	0.001	0.069	0.030	0.024	0.064
	AC	0.112	0.065	0.065	0.451	0.001	0.059	0.456	0.087	0.071
	ABC	0.115	0.059	0.059	0.451	0.001	0.054	0.452	0.087	0.064
	B	0.059	0.063	0.070	0.706	0.710	0.056	0.024	0.023	0.054

Note: For $a = 20$, 620 of the group sizes are 4, 120 of them are 5, and 60 of them are 6. For $a = 30$, 800 of the group sizes are 4, 160 of them are 5, 220 of them are 6, and 20 of them are 7. For $a = 50$, 1400 of the group sizes are 4, 240 of them are 5, 320 of them are 6, and 40 of them are 7.

Table 5. Achieved power for testing $H_0(B)$, unbalanced heteroscedastic case, $\alpha = 0.05$, $a = 20$, $b = 2$, $c = 20$.

	Normal($j\tau/(4b)$, $4jk/(bc)$)			Log-normal($j\tau/(4b)$, $4jk/(bc)$)			Cauchy($j\tau/(4b)$, $4jk/(bc)$)		
τ	F	NP.org	NP.rank	F	NP.org	NP.rank	F	NP.org	NP.rank
0	0.060	0.062	0.059	0.638	0.630	0.050	0.017	0.019	0.052
0.5	0.165	0.163	0.283	0.641	0.642	1.000	0.040	0.041	1.000
1	0.490	0.484	0.758	0.666	0.667	1.000	0.040	0.041	1.000
1.5	0.823	0.819	0.980	0.656	0.651	1.000	0.077	0.074	1.000

Note: The group sizes are $n_{4,2,k} = n_{8,2,k} = n_{10,2,k} = n_{13,1,k} = n_{15,2,k} = n_{20,1,k} = 5$, $n_{6,2,k} = n_{9,1,k} = n_{10,1,k} = 6$, for all k, and the rest of the group sizes are 4.

Table 6. Achieved power for testing $H_0(AC)$, unbalanced heteroscedastic case, $\alpha = 0.05$, $a = 20$, $b = 2$, $c = 20$.

	Normal($4ik\tau/(ac)$, $4j/b$)			Log-normal($4ik\tau/(ac)$, $4j/b$)			Cauchy($4ik\tau/(ac)$, $4j/b$)		
τ	F	NP.org	NP.rank	F	NP.org	NP.rank	F	NP.org	NP.rank
0	0.049	0.062	0.055	0.020	0.109	0.067	0.024	0.106	0.060
0.5	0.090	0.105	0.119	0.020	0.102	0.716	0.021	0.105	0.446
1	0.345	0.369	0.427	0.023	0.105	0.918	0.021	0.124	0.954
1.5	0.843	0.863	0.879	0.023	0.105	0.983	0.029	0.108	1.000
2	0.993	0.995	0.994						

Note: The group sizes are $n_{4,2,k} = n_{8,2,k} = n_{10,2,k} = n_{13,1,k} = n_{15,2,k} = n_{20,1,k} = 5$, $n_{6,2,k} = n_{9,1,k} = n_{10,1,k} = 6$, for all k, and the rest of of the group sizes are 4.

normal(0, $4jk/(bc)$) data, 0.706 for the log-normal(0, $4jk/(bc)$) data, and 0.024 for the Cauchy data. Even though the NP.org tests of [8] are developed for heteroscedastic unbalanced case, the performance of these tests is not good for the log-normal or Cauchy data. This is because the NP.org tests rely on finite moment assumptions and the cell sample variances are used to estimate the cell variances to obtain consistent estimate of the asymptotic variance. For log-normal(c_1, c_2) distribution with $c_2 \neq 1$, the sample variance is a very poor estimator of the variance; for Cauchy distribution, the moments do not exist. The proposed rank tests produced reliable type I error estimates in all cases.

Table 7. Estimated level for $\alpha = 0.05$, unbalanced heteroscedastic case, $b = 2$, $c = 20$, $X_{ijkm} \sim N(0, 4jk/bc)$.

	$a = 20$			$a = 30$			$a = 50$		
H_0	F	NP.org	NP.rank	F	NP.org	NP.rank	F	NP.org	NP.rank
$H_0(A)$	0.096	0.070	0.076	0.098	0.066	0.071	0.106	0.065	0.061
$H_0(AB)$	0.129	0.049	0.053	0.122	0.059	0.060	0.092	0.049	0.052
$H_0(AC)$	0.491	0.079	0.077	0.469	0.061	0.060	0.475	0.067	0.067
$H_0(ABC)$	0.381	0.069	0.070	0.431	0.065	0.063	0.429	0.059	0.058
$H_0(B)$	0.999	0.066	0.066	0.999	0.062	0.070	0.999	0.058	0.056

Note: The group sizes in these simulations are as follows: when $a = 20$, $n_{i1k} = 12$ for $i = 1, \ldots, 10$, and all $k = 1, \ldots, 20$; $n_{i1k} = 10$ for $i = 11$, and $k = 1, \ldots, 20$; and $n_{i1k} = 5$ for $i = 12, \ldots, 20$; $n_{i2k} = 4$, for all i, k. When $a = 30$, $n_{i1k} = 12$ for $i = 1, \ldots, 10$, and all $k = 1, \ldots, 20$; $n_{i1k} = 10$ for $i = 11$, and $k = 1, \ldots, 20$; and $n_{i1k} = 5$ for $i = 12, \ldots, 30$; $n_{i2k} = 4$, for all i, k. When $a = 50$, $n_{i1k} = 12$ for $i = 1, \ldots, 10$, and all $k = 1, \ldots, 20$; $n_{i1k} = 10$ for $i = 11$, and $k = 1, \ldots, 20$; and $n_{i1k} = 5$ for $i = 12, \ldots, 50$; $n_{i2k} = 4$ for all i, k.

The power estimates in Tables 5 and 6 clearly show that the proposed NP.rank tests are much more powerful than both the ANOVA F-test and the NP.org from [8] under non-normality. Under the heteroscedastic normal model used for Table 5, the NP.rank test for $H_0(B)$ has better power than the other two tests. In Table 6, all three tests have comparable power under the normal model. Note that the ANOVA F-test is not optimal under heteroscedasticity. In fact, if the design is unbalanced and heteroscedastic, the numerator and denominator of the ANOVA F-statistics have different expectations even under the null hypotheses. As the unbalancedness increases, the situation gets worse. As an example, we generate data from normal distribution with the same cell variances as are used in Table 5 ($c_2^2 = (4jk/bc)^2$), but cell sample sizes of 12, 10, 5, and 4 (see Table 7 for details). We still use $b = 2$, $c = 20$, and $a = 20$, 30, and 50. The type I error estimates for all three tests are reported in Table 7. The ANOVA F-test has type I error as high as 0.999 for $H_0(B)$ for all three values of a. On the other hand, both the NP.rank and NP.org tests have acceptable type I error when $a = 30$ or 50 (slightly liberal for the smaller value $a = 20$).

5. Summary and recommendations

In this article, we proposed rank transform versions (NP.rank) of the test statistics of [8] (NP.org), obtained their asymptotic distribution, and demonstrated their usefulness with the analysis of a real data set and simulations.

Compared with their NP.org counterparts, the NP.rank tests are free from finite moment assumptions and therefore are robust to outliers. Thus, they can be used for testing hypotheses in multi-way high-dimensional arrays of random variables with arbitrary heteroscedasticity and unbalanced sample sizes.

The simulation results provide strong support to the proposed rank tests. In particular, the NP.rank tests either produce comparable results or outperform available methods in terms of both the achieved α-level and power in the presence of low sample sizes, heteroscedasticity, and a large number of factor levels. On the basis of these results, we recommend the NP.rank tests in all situations where the combined number of factor levels is large. Moreover, the simulations make a compelling case that the ANOVA F-test should not be used in such cases because of the potential of grossly misleading results.

References

[1] J. Fan and S. Lin, *Test of significance when data are curves.* J. Amer. Statist. Assoc. 93 (1998), pp. 1007–1021.
[2] M.G. Akritas and N. Papadatos, *Heteroscedastic one-way ANOVA and lack-of-fit test,* J. Amer. Statist. Assoc. 99 (2004), pp. 368–382.
[3] M. G. Akritas and S. Arnold, *Asymptotics for ANOVA when the number of levels is large*. J. Amer. Statist. Assoc. 95 (2000), pp. 212–226.

[4] A. Bathke, *ANOVA for a large number of treatments*, Math. Meth. Stat. 11 (2002), pp. 118–132.
[5] A. Bathke, *The ANOVA F test can still be used in some balanced designs with unequal variances and nonnormal data*. J. Statist. Plann. Inference 126 (2004), pp. 413–422.
[6] H. Wang and M.G. Akritas, *Rank tests for ANOVA with large number of factor levels*. J. Nonparametr. Stat. 16 (2004), pp. 563–589.
[7] L. Wang and M.G. Akritas, *Two-way heteroscedastic ANOVA when the number of levels is large*, Statist. Sinica. 16 (2006), pp. 1387–1408.
[8] H. Wang and M.G. Akritas, *Asymptotically distribution free tests in heteroscedastic unbalanced high dimensional ANOVA*, manuscript under review, 2008.
[9] J. Neymann and E. Scott, *Consistent estimates based on partially consistent observations*. Econometrica 16 (1948), pp. 1–32.
[10] E. Andersen, *On Fisher's lower bound to asymptotic variances in case of infinitely many nuisance parameters*. Skand. Akt. J. 53 (1970), pp. 78–85.
[11] E. Andersen, *Asymptotic properties of conditional maximum-likelihood estimators*. J. Roy. Statist. Soc. Ser. B 32 (1970), pp. 283–301; correction, 33 (1971), p. 167.
[12] T. Mak, *Estimation in the presence of incidental parameters*. Can. J. Statist. 10 (1982), pp. 121–132.
[13] J. Pfanzagl, *Incidental versus random nuisance parameters*. Ann. Statist. 21 (1993), pp. 1663–1691.
[14] S. Portnoy, *Asymptotic behavior of m-estimators of p regression parameters when p^2/n is large: I, consistency*, Ann. Statist. 12 (1984), pp. 1298–1309.
[15] S. Portnoy, *Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity*, Ann. Statist. 16 (1988), pp. 356–366.
[16] H. Li, B. Lindsay, and R. Waterman, *Efficiency of projected score methods in rectangular array asymptotics*, J. Roy. Statist. Soc. Ser. B 65 (2003), pp. 191–208.
[17] Z. Bai and H. Sarandasa, *Effect of high dimension: By an example of a two sample problem*, Statist. Sinica. 6 (1996), pp. 311–329.
[18] H. Sarandasa and S. Altan, *The analysis of small-sample multivariate data*, J. Biopharm. Statist. 8 (1998), pp. 163–186.
[19] I.M. Johnstone, *On the distribution of the largest eigenvalue in principal components analysis*, Ann. Statist. 29 (2001), pp. 295–327.
[20] P. Hall, J.S. Marron, and A. Neeman, *Geometric representation of high dimension, low sample size data*. J. Roy. Statist. Soc. Ser. B 67 (2005), pp. 427–444.
[21] M.G. Akritas, *The rank transform method in some two-factor designs*. J. Amer. Statist. Assoc. 85 (1990), pp. 73–78.
[22] M. G. Akritas and S. Arnold, *Fully nonparametric hypotheses for factorial designs I: Multivariate repeated measures designs*, J. Amer. Statist. Assoc. 89 (1994), pp. 336–343.
[23] M.G. Akritas, S. Arnold and E. Brunner, *Nonparametric hypotheses and rank statistics for unbalanced factorial designs*, J. Amer. Statist. Assoc. 92 (1997), pp. 258–265.
[24] H. Sahai and M. Ageel, *The Analysis of Variance: Fixed, Random, and Mixed Models*, Springer Verlag, New York, 2000.
[25] K. Toufighi, S.M. Brady, R. Austin, E. Ly, and N.J. Provart, *The botany array resource: e-Northerns, expression angling, and promoter analyses*. Plant J. 43 (2005), 153–163, doi:10.1111/j.1365–313X.2005.02437.x.
[26] S. Arnold, *The Theory of Linear Models and Multivariate Analysis*, Wiley, New York, 1981.
[27] H. Wang, *Testing in multifactor heteroscedastic ANOVA and repeated measures designs with large number of levels*, PhD. thesis, Pennsylvania State University, 2004.

Appendix 1. Major proofs

For simplicity in presentation, we only give some typical proofs. Specifically, we give only the proofs of parts (a) and (b) of Theorem 3.1, part (b) of Theorem 3.2, and part (b) of Theorem 3.3. Proofs of the remaining parts follow by similar arguments. The reader may refer to [27] for all omitted proofs.

A.1. *Proof of Theorem* 3.1

(a) Let $Q_H(A)$, $Q_{\hat{H}}(A)$ be the $Q_R(A)$ statistic evaluated at $H(X_{ijk})$, $\hat{H}(X_{ijk})$, respectively, and similarly let $\mathbf{W}_H$, $\hat{V}_H$, $\mathbf{W}_{\hat{H}}$, $\hat{V}_{\hat{H}}$, be the $\mathbf{W}$, $\hat{V}$, as given below (3), again with R_{ijk} replaced by $H(X_{ijk})$, $\hat{H}(X_{ijk})$, respectively. By part (a) of Theorem 2.1 of [8], we know that the $Q_H(A)$ converges in distribution to χ^2_{a-1}. Thus, since $Q_R(A) = Q_{\hat{H}}(A)$, it suffices to establish that

$$\begin{aligned} Q_{\hat{H}}(A) - Q_H(A) &= \sqrt{N}(\mathbf{W}'_{\hat{H}} - \mathbf{W}'_H)\mathbf{C}'_A(\mathbf{C}_A\hat{\mathbf{V}}_H\mathbf{C}'_A)^{-1}\mathbf{C}_A\sqrt{N}\mathbf{W}_{\hat{H}} \\ &\quad + \sqrt{N}\mathbf{W}'_H\mathbf{C}'_A(\mathbf{C}_A\hat{\mathbf{V}}_H\mathbf{C}'_A)^{-1}\mathbf{C}_A\sqrt{N}(\mathbf{W}_{\hat{H}} - \mathbf{W}_H) \\ &\quad + \sqrt{N}\mathbf{C}'_A\mathbf{W}'_{\hat{H}}[(\mathbf{C}_A\hat{\mathbf{V}}_{\hat{H}}\mathbf{C}'_A)^{-1} - (\mathbf{C}_A\hat{\mathbf{V}}_H\mathbf{C}'_A)^{-1}]\mathbf{C}_A\sqrt{N}\mathbf{W}_{\hat{H}} \to 0. \end{aligned} \tag{A1}$$

Given that the elements of $\hat{\mathbf{V}}_{\hat{H}}$ and $\hat{\mathbf{V}}_H$ stay bounded away from zero and infinity, the first two expressions on the right-hand side of (A1) will be shown to converge to zero under $H_0(A)$ if

$$\mathbf{C}_A(\mathbf{W}_{\hat{H}} - \mathbf{W}_H) = \mathbf{C}_A \int (\hat{H} - H)\, \mathrm{d}\hat{\mathbf{V}} = \mathbf{C}_A \int (\hat{H} - H) \mathrm{d}(\hat{\mathbf{F}} - \mathbf{F}) = o_p\left(\frac{1}{\sqrt{N}}\right), \tag{A2}$$

where $\mathbf{F} = (\bar{F}_{1.}, \ldots, \bar{F}_{a.})'$ and $\hat{\mathbf{V}} = (\hat{\bar{F}}_{1.}, \ldots, \hat{\bar{F}}_{a.})'$, with $\bar{F}_{i.}(x) = b^{-1}\sum_{j=1}^{b} F_{ij}(x)$ and $\hat{\bar{F}}_{i.}(x) = b^{-1}\sum_{j=1}^{b} \hat{F}_{ij}(x)$, where $\hat{F}_{ij}(x) = n_{ij}^{-1}\sum_{k=1}^{n_{ij}} c(X_{ijk}, x)$. Note that the second equality in Equation (A2) holds only under $H_0(A)$. For Equation (A2) we will show that for each $i = 1, \ldots, a$,

$$\sqrt{N} \int (\hat{H} - H)\, \mathrm{d}(\hat{\bar{F}}_{i.} - \bar{F}_{i.}) \xrightarrow{p} 0. \tag{A3}$$

A similar result is shown in [22] but under fixed number of factor levels and group sizes tending to infinity, which is not directly applicable to our situation. The proof of Equation (A3) is given in Lemma A.1. Finally, the last expression on the right-hand side of Equation (A1) converges to zero in probability by $\hat{\mathbf{V}}_{\hat{H}} - \hat{\mathbf{V}}_H \to 0$ and the fact that $N^{1/2}\mathbf{C}_A\mathbf{W}_{\hat{H}}$ is bounded in probability, which follows from Equation (A2), and the result of part (a) of Theorem 2.1 in [8].

(b) In this proof, we will keep the notations $Y_{ijm} = H(X_{ijm})$, $e_{ijm} = Y_{ijm} - E(Y_{ijm})$, and $R_{ijm} =$ (mid-)rank of X_{ijm}, and further will denote $Z_{ijm} = \hat{H}(X_{ijm})$. Set $\mathbf{Y} = (Y_{111}, \ldots, Y_{abn_{ab}})$ and let $\mathbf{Z}$, $\mathbf{R}$ be similarly defined. To be clear, we will use $\mathrm{MST}_A(\mathbf{Y})$, $\mathrm{MST}_A(\mathbf{Z})$, and $\mathrm{MST}_A(\mathbf{R})$ to denote the MST_A statistic defined in connection with Equation (4) evaluated on $\mathbf{Y}$, $\mathbf{Z}$ and $\mathbf{R}$, respectively. $\mathrm{MSE}(\mathbf{R})$, $\mathrm{MSE}(\mathbf{Z})$, and $\mathrm{MSE}(\mathbf{Y})$ are defined similarly. Note that $\mathrm{MST}_A(\mathbf{R})/N^2 = \mathrm{MST}_A(\mathbf{Z})$. By Lemmas A.2 and A.4, it suffices to establish the asymptotic distribution of $T(\mathbf{Z} - E(\mathbf{Y})) = n(a,b)\sqrt{a}[P_A(\mathbf{Z} - E(\mathbf{Y})) - \mathrm{MSE}(\mathbf{Z}))]$. We will do so by showing $T(\mathbf{Z} - E(\mathbf{Y})) - T(\mathbf{Y} - E(\mathbf{Y})) = o_p(1)$, using the fact that by applying part (b) of Theorem 2.1 of [8] on $\mathbf{Y}$, we have $T(\mathbf{Y} - E(\mathbf{Y}))/\sqrt{\tau_{1A}} \xrightarrow{d} N(0,1)$. Write $T(\mathbf{Z} - E(\mathbf{Y})) = T_{1A}(\mathbf{Z} - E(\mathbf{Y})) + T_{3A}(\mathbf{Z} - E(\mathbf{Y}))$, where T_{1A} and T_{3A} are defined as

$$T_{1A}(\mathbf{e}) = \frac{n(a,b)}{b\sqrt{a}} \sum_{i=1}^{a}\sum_{j\neq j'}^{b} \bar{e}_{ij.}\bar{e}_{ij'.}, \quad T_{3A}(\mathbf{e}) = \frac{n(a,b)}{b\sqrt{a}} \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{m\neq m'}^{n_{ij}} \frac{e_{ijm}e_{ijm'}}{n_{ij}(n_{ij}-1)}.$$

Using a similar decomposition for $T(\mathbf{Y} - E(\mathbf{Y}))$, it follows that to show $T(\mathbf{Z} - E(\mathbf{Y})) - T(\mathbf{Y} - E(\mathbf{Y})) = o_p(1)$, we only need to show $T_{sA}(\mathbf{Z} - E(\mathbf{Y})) - T_{sA}(\mathbf{Y} - E(\mathbf{Y})) = o_p(1)$, for $s = 1, 3$. These proofs are similar to that of $\sqrt{b}[D_6(\mathbf{Z} - E(\mathbf{Y})) - D_6(\mathbf{Y} - E(\mathbf{Y}))] = o_p(1)$ in the proof of Lemma A.3 and thus are omitted.

A.2. *Proof of part (b) of Theorem* 3.2

In the following proof, we will keep the notations $Y_{ijkm} = H(X_{ijkm})$, $p_{ijk} = E(Y_{ijkm})$, $e_{ijkm} = Y_{ijkm} - p_{ijk}$, and $R_{ijkm} =$ (mid-)rank of X_{ijkm}, and further will denote $Z_{ijkm} = \hat{H}(X_{ijkm})$. Note $R_{ijkm} = NZ_{ijkm} + 0.5$. Also, denote $\mathbf{Y} = (Y_{1111}, \ldots, Y_{abcn_{abc}})$ and let $\mathbf{Z}$, $\mathbf{R}$ be similarly defined. To be clear, we will use $\mathrm{MST}_{AB}(\mathbf{Y})$, $\mathrm{MST}_{AB}(\mathbf{Z})$ and $\mathrm{MST}_{AB}(\mathbf{R})$ to denote the MST_{AB} statistic defined in connection with Equation (8) evaluated on $\mathbf{Y}$, $\mathbf{Z}$, and $\mathbf{R}$, respectively. $\mathrm{MSE}(\mathbf{R})$, $\mathrm{MSE}(\mathbf{Z})$, and $\mathrm{MSE}(\mathbf{Y})$ are defined similarly. Note that $\mathrm{MST}_{ABC}(\mathbf{R})/N^2 = \mathrm{MST}_{ABC}(\mathbf{Z})$.

By Lemmas A.5 and A.6, it suffices to establish the asymptotic distribution of $T_2(\mathbf{Z} - E(\mathbf{Y})) = n(a,b,c)\sqrt{a}[P_{2,AB}(\mathbf{Z} - E(\mathbf{Y})) - \mathrm{MSE}(\mathbf{Z})]$. We will do so by showing the asymptotic equivalence of $T_2(\mathbf{Z} - E(\mathbf{Y}))$ and $T_2(\mathbf{Y} - E(\mathbf{Y}))$, and by the result of part (b) of Theorem 2.2 in [8], we have $T_2(\mathbf{Y} - E(\mathbf{Y}))/\sqrt{\tau_1^2 + \tau_2^2 + \tau_3^2} \xrightarrow{d} N(0,1)$. Write $T_2(\mathbf{Z} - E(\mathbf{Y})) = T_{1AB}(\mathbf{Z} - E(\mathbf{Y})) + T_{2AB}(\mathbf{Z} - E(\mathbf{Y})) + T_{3AB}(\mathbf{Z} - E(\mathbf{Y}))$, where

$$T_{1AB}(\mathbf{Z} - E(\mathbf{Y})) = -\frac{n(a,b,c)c}{b(b-1)\sqrt{a}} \sum_{i=1}^{a}\sum_{j\neq j'}^{b} (\tilde{Z}_{ij..} - \bar{p}_{ij.})(\tilde{Z}_{ij'..} - \bar{p}_{ij'.}),$$

$$T_{2AB}(\mathbf{Z} - E(\mathbf{Y})) = \frac{n(a,b,c)}{bc\sqrt{a}} \sum_{i,j}\sum_{k\neq k'}^{c} (\tilde{Z}_{ijk.} - p_{ijk})(\tilde{Z}_{ijk'.} - p_{ijk'}), \tag{A4}$$

$$T_{3AB}(\mathbf{Z} - E(\mathbf{Y})) = \frac{n(a,b,c)}{bc\sqrt{a}} \sum_{i,j,k}\sum_{m\neq m'}^{n_{ijk}} \frac{(Z_{ijkm} - p_{ijk})(Z_{ijkm'} - p_{ijk})}{n_{ijk}(n_{ijk}-1)}. \tag{A5}$$

Using a similar decomposition for $T_2(\mathbf{Y} - E(\mathbf{Y}))$, it follows that to show $T_2(\mathbf{Z} - E(\mathbf{Y})) - T_2(\mathbf{Y} - E(\mathbf{Y})) = o_p(1)$, we only need to show $T_{sAB}(\mathbf{Z} - E(\mathbf{Y})) - T_{sAB}(\mathbf{Y} - E(\mathbf{Y})) = o_p(1)$, for $s = 1, 2, 3$. These proofs are similar to that of $\sqrt{b}[D_6(\mathbf{Z} - E(\mathbf{Y})) - D_6(\mathbf{Y} - E(\mathbf{Y}))] = o_p(1)$ in the proof of Lemma A.3 and thus are omitted.

A.3. *Proof of part (b) of Theorem* 3.3

For the remainder part of this section, we will keep the notations $Y_{ijklm} = H(X_{ijklm})$, $e_{ijklm} = Y_{ijklm} - E(Y_{ijklm})$, R_{ijklm} =(mid-)rank of X_{ijklm}, and further will denote $Z_{ijklm} = \hat{H}(X_{ijklm})$. Note $R_{ijklm} = NZ_{ijklm} + 0.5$. Also, denote $\mathbf{Y} = (Y_{11111}, \ldots, Y_{abcdn_{abcd}})$ and let $\mathbf{Z}$, $\mathbf{R}$ be similarly defined. To be clear, we will use $\mathrm{MST}_{ABC}(\mathbf{Y})$, $\mathrm{MST}_{ABC}(\mathbf{Z})$, and $\mathrm{MST}_{ABC}(\mathbf{R})$ to denote the MST_{ABC} statistic defined in connection with Equation (13) evaluated on $\mathbf{Y}$, $\mathbf{Z}$, and $\mathbf{R}$, respectively. $\mathrm{MSE}(\mathbf{R})$, $\mathrm{MSE}(\mathbf{Z})$, and $\mathrm{MSE}(\mathbf{Y})$ are defined similarly. Note that $\mathrm{MST}_{ABC}(\mathbf{R})/N^2 = \mathrm{MST}_{ABC}(\mathbf{Z})$.

By Lemmas A.7 and A.8, it suffices to establish the asymptotic distribution of $T_1(\mathbf{Z} - E(\mathbf{Y})) = n(a,b,c,d)\sqrt{ac}[P_{1,ABC}(\mathbf{Z} - E(\mathbf{Y})) - \mathrm{MSE}(\mathbf{Z})]$. We will do so by showing the asymptotic equivalence of $T_2(\mathbf{Z} - E(\mathbf{Y}))$ and $T_2(\mathbf{Y} - E(\mathbf{Y}))$. Applying the result of part (b) of Theorem 2.3 in [8], we have $T_1(\mathbf{Y} - E(\mathbf{Y}))/\sqrt{c(\tau_4 + \tau_5 + \tau_6)} \xrightarrow{d} N(0,1)$. Thus, it suffices to show $T_1(\mathbf{Z} - E(\mathbf{Y})) - T_1(\mathbf{Y} - E(\mathbf{Y})) = o_p(1)$ as $a, c \to \infty$ while b remains fixed regardless of whether d and n_{ijkl} are large or small. Write

$$T_1(\mathbf{Y} - E(\mathbf{Y})) = T_{11}(\mathbf{Y} - E(\mathbf{Y})) + T_{12}(\mathbf{Y} - E(\mathbf{Y})) - \frac{T_{13}(\mathbf{Y} - E(\mathbf{Y}))}{b-1},$$

where

$$T_{11}(\mathbf{Y} - E(\mathbf{Y})) = \frac{n(a,b,c,d)}{bd\sqrt{ac}} \sum_{i,j,k,l} \sum_{m \neq m'}^{n_{ijkl}} \frac{(Y_{ijklm} - p_{ijkl})(Y_{ij'klm'} - p_{ij'kl})}{n_{ijkl}(n_{ijkl} - 1)}, \tag{A6}$$

$$T_{12}(\mathbf{Y} - E(\mathbf{Y})) = \frac{n(a,b,c,d)}{bd\sqrt{ac}} \sum_{i,j,k} \sum_{l \neq l'}^{d} (\bar{Y}_{ijkl.} - p_{ijkl})(\bar{Y}_{ijkl'.} - p_{ijkl'}), \tag{A7}$$

$$T_{13}(\mathbf{Y} - E(\mathbf{Y})) = \frac{dn(a,b,c,d)}{b\sqrt{ac}} \sum_{i,k} \sum_{j \neq j'}^{b} (\bar{Y}_{ijk..} - \bar{p}_{ijk.})(\bar{Y}_{ij'k..} - \bar{p}_{ij'k.}). \tag{A8}$$

Using a similar decomposition for $T_1(\mathbf{Z} - E(\mathbf{Y}))$, it follows that to show $T_1(\mathbf{Z} - E(\mathbf{Y})) - T_1(\mathbf{Y} - E(\mathbf{Y})) = o_p(1)$, we only need to show $T_{1s}(\mathbf{Z} - E(\mathbf{Y})) - T_{1s}(\mathbf{Y} - E(\mathbf{Y})) = o_p(1)$, for $s = 1, 2, 3$ (note that b is fixed). These proofs are similar to that of $\sqrt{b}[D_6(\mathbf{Z} - E(\mathbf{Y})) - D_6(\mathbf{Y} - E(\mathbf{Y}))] = o_p(1)$ in the proof of Lemma A.3 and thus are omitted.

Appendix 2. Some lemmas used in proofs above

LEMMA A.1 *Let $j = 1, \ldots, b$ index the levels of a factor whose number of levels tends to infinity, and let i index the levels of all other factors. Let $\widehat{H}$ denote the (average of the left- and right-continuous versions of the) empirical distribution function of all observations and $H = E(\hat{H})$. Then, regardless of whether or not the cell sample sizes tend to ∞,*

$$\sqrt{N} \int (\hat{H} - H)\, \mathrm{d}(\hat{\bar{F}}_{i.} - \bar{F}_{i.}) \xrightarrow{p} 0, \quad \text{as } b \to \infty. \tag{A9}$$

Proof The left-hand side of (A9) is

$$\frac{1}{b\sqrt{N}} \sum_{i',j} n_{i'j} \sum_{j_1=1}^{b} \int (\hat{F}_{i'j} - F_{i'j})\, \mathrm{d}(\hat{F}_{ij_1} - F_{ij_1}) = \frac{1}{b\sqrt{N}} \sum_{j_1=1}^{b} \sum_{k_1=1}^{n_{ij_1}} \sum_{i' \neq i,j} \sum_{k_2=1}^{n_{i'j}} \frac{h(X_{i'jk_2}, X_{ij_1k_1})}{n_{ij_1}},$$

where

$$h(X_{i'jk_2}, X_{ij_1k_1}) = c(X_{i'jk_2}, X_{ij_1k_1}) - F_{i'j}(X_{ij_1k_1}) - \left[1 - F_{ij_1}(X_{i'jk_2}) - \int F_{i'j}\, \mathrm{d}F_{ij_1}\right].$$

Thus,

$$\begin{aligned}
&E\left[\sqrt{N} \int (\hat{H} - H)\, \mathrm{d}(\hat{\bar{F}}_{i.} - \bar{F}_{i.})\right]^2 \\
&= \frac{1}{b^2 N} \sum_{j_1=1}^{b} \sum_{k_1=1}^{n_{ij_1}} \sum_{j_3=1}^{b} \sum_{k_3=1}^{n_{ij_3}} \frac{1}{n_{ij_1} n_{ij_3}} \sum_{i',j} \sum_{k_2=1}^{n_{i'j}} \sum_{i_4, j_4} \sum_{k_4=1}^{n_{i_4 j_4}} E[h(X_{i'jk_2}, X_{ij_1k_1}) h(X_{i_4 j_4 k_4}, X_{ij_3k_3})] \\
&= \frac{1}{b^2 N} \sum_{j_1=1}^{b} \sum_{k_1=1}^{n_{ij_1}} \sum_{i',j} \sum_{k_2=1}^{n_{i'j}} \frac{1}{n_{ij_1} n_{ij_1}} E[h^2(X_{i'jk_2}, X_{ij_1k_1})] I(j \neq j_1) I(k_1 \neq k_2) \\
&\quad + \frac{1}{b^2 N} \sum_{j_1=1}^{b} \sum_{k_1=1}^{n_{ij_1}} \sum_{j=1}^{b} \sum_{k_2=1}^{n_{ij}} \frac{1}{n_{ij_1} n_{ij}} E[h(X_{ijk_2}, X_{ij_1k_1}) h(X_{ij_1k_1}, X_{ijk_2})],
\end{aligned}$$

where the second equality holds because $E[h(X_{i_1}, X_{i_2})h(X_{i_3}, X_{i_4})] = 0$ if the number of different elements in $\{i_1, i_2, i_3, i_4\}$ is four or three due to independence and

$$E[h(X_{i'jk_2}, X_{ij_1k_1})] = E[h(X_{i'jk_2}, X_{ij_1k_1})|X_{i'jk_2}] = E[h(X_{i'jk_2}, X_{ij_1k_1})|X_{ij_1k_1}] = 0.$$

Since $|h(\cdot,\cdot)|$ is uniformly bounded by 4, we have

$$E\left[\sqrt{N}\int(\hat{H}-H)\,\mathrm{d}(\hat{\bar{F}}_{i\cdot}-\bar{F}_{i\cdot})\right]^2 \le \frac{1}{b^2}\sum_{j_1=1}^{b}\frac{16}{n_{ij_1}}+\frac{16}{N}\to 0,$$

as $b\to\infty$, regardless of whether n_{ij} are large or small, which implies that Equation (A3) holds. ■

LEMMA A.2 *Under the settings and assumptions of part* (*b*) *of Theorem* 3.1,

$$n(a,b)[\mathrm{MSE}/N^2-\sigma_A^2]\xrightarrow{p}0 \quad \text{as } \max\{a,b\}\to\infty,$$

where $n(a,b)=\min_{i,j}\{n_{ij}\}$, *regardless of whether* n_{ij} *remain fixed or tend to* ∞.

Proof By Lemma 4.4 of [8], we only need to show that, in both cases, $\mathrm{MSE}/N^2-\mathrm{MSE}_Y=o_p(n(a,b)^{-1})$, where MSE_Y is similarly defined as MSE with R_{ijk} replaced by $Y_{ijk}=H(X_{ijk})$. Write

$$\begin{aligned}\mathrm{MSE}/N^2-\mathrm{MSE}_Y &= \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{1}{n_{ij}(n_{ij}-1)}\sum_{k=1}^{n_{ij}}[(Z_{ijk}-\bar{Z}_{ij\cdot})^2-(Y_{ijk}-\bar{Y}_{ij\cdot})^2]\\ &= \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{1}{n_{ij}(n_{ij}-1)}\sum_{k=1}^{n_{ij}}(Z_{ijk}-\bar{Z}_{ij\cdot}-Y_{ijk}+\bar{Y}_{ij\cdot})^2\\ &\quad+\frac{2}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{1}{n_{ij}(n_{ij}-1)}\sum_{k=1}^{n_{ij}}(Z_{ijk}-\bar{Z}_{ij\cdot}-Y_{ijk}+\bar{Y}_{ij\cdot})(Y_{ijk}-\bar{Y}_{ij\cdot}).\end{aligned}$$

The first summation is $O_p(N^{-1}n(a,b)^{-1})$ since $\sup_x|\hat{H}(x)-H(x)|=O_p(N^{-1/2})$. Because $Y_{ijk}-\bar{Y}_{ij\cdot}$ is uniformly bounded by 1, the second summation is bounded by

$$\frac{2}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{1}{n_{ij}(n_{ij}-1)}\sum_{k=1}^{n_{ij}}|(Z_{ijk}-\bar{Z}_{ij\cdot}-Y_{ijk}+\bar{Y}_{ij\cdot})|=O_p\left(\frac{1}{\sqrt{N}n(a,b)}\right).$$

Therefore, $\mathrm{MSE}/N^2-\mathrm{MSE}_Y=O_p(N^{-1/2}n(a,b)^{-1})=o_p(n(a,b)^{-1})$, as $\max\{a,b\}\to\infty$, whether n_{ij} is fixed or not. ■

LEMMA A.3 *Let* $P_{1,ABC}(\mathbf{Z}-E(\mathbf{Y}))$, $P_{2,ABC}(\mathbf{Z}-E(\mathbf{Y}))$, *and* $P_{3,ABC}(\mathbf{Z}-E(\mathbf{Y}))$ *be similarly defined as* $P_{1,ABC}(\mathbf{e})$, $P_{2,ABC}(\mathbf{e})$, *and* $P_{3,ABC}(\mathbf{e})$ *in Proposition* 4.1 *in* [8] *with* e_{ijklm} *replaced by* $Y_{ijklm}-E(Y_{ijklm})$. *Also,* $\mathrm{MST}_{ABC}(\mathbf{Z})$ *is similarly defined as* MST_{ABC} *in Equation* (13) *with* $R_{ijklm}=Z_{ijklm}$. *Then under* $H_0(ABC)$,

(1) *under the settings and assumptions of part* (*c*) *of Theorem* 3.3, *as* $a,c\to\infty$,

$$T_1^*(\mathbf{Z}-E(\mathbf{Y}))=n(a,b,c,d)\sqrt{ac}(\mathrm{MST}_{ABC}(\mathbf{Z})-P_{1,ABC}(\mathbf{Z}-E(\mathbf{Y})))\xrightarrow{p}0;$$

(2) *under the settings and assumptions of part* (*b*) *of Theorem* 3.3, *as* $a\to\infty$,

$$T_2^*(\mathbf{Z}-E(\mathbf{Y}))=n(a,b,c,d)\sqrt{a}(\mathrm{MST}_{ABC}(\mathbf{Z})-P_{2,ABC}(\mathbf{Z}-E(\mathbf{Y})))\xrightarrow{p}0;$$

(3) *under the settings and assumptions of part* (*d*) *of Theorem* 3.3, *as* $a,b,c\to\infty$,

$$T_3^*(\mathbf{Z}-E(\mathbf{Y}))=n(a,b,c,d)\sqrt{abc}(\mathrm{MST}_{ABC}(\mathbf{Z})-P_{3,ABC}(\mathbf{Z}-E(\mathbf{Y})))\xrightarrow{p}0;$$

regardless of whether d *and* n_{ijkl} *tend to* ∞ *or stay fixed as long as* $n_{ijkl}\ge 2$.

Proof By Proposition 4.1 in [8], $T_t^*(\mathbf{Y}-E(\mathbf{Y}))=o_p(1)$, for $t=1,2,3$, under $H_0(ABC)$, where $T_t^*(\mathbf{Y}-E(\mathbf{Y}))$ are similarly defined as $T_t^*(\mathbf{Z}-E(\mathbf{Y}))$. Therefore, we only need to show the difference is $o_p(1)$ under $H_0(ABC)$, i.e. $D_{t,ZY}=T_t^*(\mathbf{Z}-E(\mathbf{Y}))-T_t^*(\mathbf{Y}-E(\mathbf{Y}))=o_p(1)$. In fact under $H_0(ABC)$, $T_t^*(\mathbf{Y}-E(\mathbf{Y}))$ have the same decompositions as

$T_t^*(\mathbf{e})$ in the proof of Proposition 4.1 in [8] with $\mathbf{e}$ replaced by $\mathbf{Y} - E(\mathbf{Y})$, for $t = 1, 2, 3$. Similar decompositions apply to $T_t^*(\mathbf{Z} - E(\mathbf{Y}))$ with $\mathbf{Z} - E(\mathbf{Y})$ as argument. That is,

$$T_1^*(\mathbf{Y} - E(\mathbf{Y})) = D_4(\mathbf{Y} - E(\mathbf{Y})) + D_5(\mathbf{Y} - E(\mathbf{Y})) + D_6(\mathbf{Y} - E(\mathbf{Y})) - \frac{D_1(\mathbf{Y} - E(\mathbf{Y}))}{b-1} - \frac{D_2(\mathbf{Y} - E(\mathbf{Y}))}{b-1} - \frac{D_3(\mathbf{Y} - E(\mathbf{Y}))}{b-1},$$

$$T_2^*(\mathbf{Y} - E(\mathbf{Y})) = \frac{D_5(\mathbf{Y} - E(\mathbf{Y})) + D_6(\mathbf{Y} - E(\mathbf{Y}))}{\sqrt{c}} - \frac{D_2(\mathbf{Y} - E(\mathbf{Y})) + D_3(\mathbf{Y} - E(\mathbf{Y}))}{\sqrt{c}(b-1)},$$

$$T_3^*(\mathbf{Y}) = \sqrt{b}T_1^*(\mathbf{Y} - E(\mathbf{Y})) + D_7(\mathbf{Y} - E(\mathbf{Y})),$$

where $D_t(\mathbf{Y} - E(\mathbf{Y}))$ is similarly defined as $D_t(\mathbf{e})$ below with e_{ijklm} replaced by $Y_{ijklm} - E(Y_{ijklm})$ and $Z_{ijklm} - E(Y_{ijklm})$, respectively:

$$D_1(\mathbf{e}) = -\frac{bd\,n(a,b,c,d)}{(c-1)\sqrt{ac}} \sum_{i=1}^{a} \sum_{k \neq k'}^{c} \tilde{e}_{i.k..}\tilde{e}_{i.k'..},$$

$$D_2(\mathbf{e}) = -\frac{bd\,n(a,b,c,d)}{(a-1)\sqrt{ac}} \sum_{i \neq i'}^{a} \sum_{k=1}^{c} \tilde{e}_{i.k..}\tilde{e}_{i'.k..},$$

$$D_3(\mathbf{e}) = \frac{bd\,n(a,b,c,d)}{(a-1)(c-1)\sqrt{ac}} \sum_{i \neq i'}^{a} \sum_{k \neq k'}^{c} \tilde{e}_{i.k..}\tilde{e}_{i'.k'..},$$

$$D_4(\mathbf{e}) = -\frac{d\,n(a,b,c,d)}{(b-1)(c-1)\sqrt{ac}} \sum_{i,j} \sum_{k \neq k'}^{c} \tilde{e}_{ijk..}\tilde{e}_{ijk'..},$$

$$D_5(\mathbf{e}) = \frac{d\,n(a,b,c,d)}{(a-1)(b-1)(c-1)\sqrt{ac}} \sum_{i \neq i'}^{a} \sum_{j=1}^{b} \sum_{k \neq k'}^{c} \tilde{e}_{ijk..}\tilde{e}_{i'jk'..},$$

$$D_6(\mathbf{e}) = -\frac{d\,n(a,b,c,d)}{(a-1)(b-1)\sqrt{ac}} \sum_{i \neq i'}^{a} \sum_{j,k} \tilde{e}_{ijk..}\tilde{e}_{i'jk..},$$

$$D_7(\mathbf{e}) = \frac{dn(a,b,c,d)}{(b-1)\sqrt{abc}} \sum_{i,k} \sum_{j \neq j'}^{b} \tilde{e}_{ijk..}\tilde{e}_{ij'k..}.$$

Define $T_2^*(\mathbf{Z} - E(\mathbf{Y}))$ similarly as $T_2^*(\mathbf{Y} - E(\mathbf{Y}))$ when the argument is $\mathbf{Z} - E(\mathbf{Y})$ instead of $\mathbf{Y} - E(\mathbf{Y})$.

To show $D_{t,ZY} = o_p(1)$, $t = 1, 2, 3$, for all three cases, it suffices to prove $D_t(\mathbf{Z} - E(\mathbf{Y})) - D_t(\mathbf{Y} - E(\mathbf{Y})) = o_p(1)$, for $t = 1, 2, 3, 7$, and $\sqrt{b}[D_s(\mathbf{Z} - E(\mathbf{Y})) - D_s(\mathbf{Y} - E(\mathbf{Y}))] = o_p(1)$, for $s = 4, 5, 6$. The proofs are similar and we only give the last one. Write $\sqrt{b}[D_6(\mathbf{Z} - E(\mathbf{Y})) - D_6(\mathbf{Y} - E(\mathbf{Y}))] = D_{31}^* + D_{32}^*$, where

$$D_{31}^* = -\frac{\sqrt{b}d\,n(a,b,c,d)}{(a-1)(b-1)\sqrt{ac}} \sum_{i \neq i'}^{a} \sum_{j,k} (\tilde{Z}_{ijk..} - \tilde{Y}_{ijk..})(\tilde{Z}_{i'jk..} - \tilde{Y}_{i'jk..}),$$

$$D_{32}^* = -\frac{2\sqrt{b}d\,n(a,b,c,d)}{(a-1)(b-1)\sqrt{ac}} \sum_{i \neq i'}^{a} \sum_{j,k} (\tilde{Z}_{ijk..} - \tilde{Y}_{ijk..})(\tilde{Y}_{i'jk..} - \bar{p}_{i'jk.}).$$

Because $\sup_x(\hat{H}(x) - H(x)) = O_p(N^{-1/2})$, we have $D_{31}^* = O_p(d\sqrt{ab}n(a,b,c,d)/N) = o_p(1)$.

$$D_{32}^* = -\frac{2\sqrt{b}n(a,b,c,d)}{(a-1)(b-1)d\sqrt{ac}} \sum_{i \neq i'}^{a} \sum_{j,k} \sum_{l,m} \sum_{l',m'} \frac{(Z_{ijklm} - Y_{ijklm})(Y_{i'jkl'm'} - p_{i'jkl'})}{n_{ijkl}n_{i'jkl'}}.$$

Write $Z_{ijklm} - Y_{ijklm} = N^{-1}\sum_{i_4,j_4,k_4,l_4,m_4}(c(X_{i_4,j_4,k_4,l_4,m_4}, X_{ijklm}) - F_{i_4,j_4,k_4,l_4,m_4}(X_{ijklm}))$, then

$$E(D_{32}^*)^2 = \frac{4bn^2(a,b,c,d)}{(a-1)^2(b-1)^2d^2ac} \sum_{i \neq i'}^{a} \sum_{j,k,l,m} \sum_{l',m'} \sum_{i_1 \neq i_1'}^{a} \sum_{j_1,k_1,l_1,m_1} \sum_{l_1',m_1'}$$
$$\times \frac{1}{n_{ijkl}n_{i'jkl'}n_{i_1j_1k_1l_1}n_{i_1'j_1k_1l_1'}} \frac{1}{N^2} \sum_{i_4,j_4,k_4,l_4,m_4} \sum_{i_5,j_5,k_5,l_5,m_5} E[(Y_{i'jkl'm'} - p_{i'jkl'm'})$$
$$\times (Y_{i_1'j_1k_1l_1'm_1'} - p_{i_1'j_1k_1l_1'm_1'})(c(X_{i_4,j_4,k_4,l_4,m_4}, X_{ijklm}) - F_{i_4,j_4,k_4,l_4,m_4}(X_{ijklm}))$$
$$\times (c(X_{i_5,j_5,k_5,l_5,m_5}, X_{i_1j_1k_1l_1m_1}) - F_{i_5,j_5,k_5,l_5,m_5}(X_{i_1j_1k_1l_1m_1}))].$$

By independence, the expectation under the summation is zero if the number of different elements in $\{i, i', i_1, i_1', i_4, i_5\}$ is five or six, or the number of different elements in $\{l, l', l_1, l_1', l_4, l_5\}$ is five or six, or the number of different elements in $\{m, m', m_1, m_1', m_4, m_5\}$ is five or six, or the number of different pairs in $\{(j, k), (j_1, k_1), (j_4, k_4), (j_5, k_5)\}$ is three or four. Hence, by the fact that Y_{ijklm} and $c(X_1, X_2)$ are uniformly bounded, $E(D_{32}^*)^2 = O(abd^2n^2(a, b, c, d)/N^2) = o(1)$. Therefore, $D_{32}^* = o_p(1)$.

The following lemmas are stated without proof since the proof of Lemmas A.4 and A.6 follows similar argument as that in Lemma A.3, and the proof of Lemmas A.7 and A.5 is similar to that of Lemma A.2 and thus are omitted (refer to [27] for details). ■

LEMMA A.4 *Let* $P_A(\boldsymbol{Z} - E(\boldsymbol{Y}))$ *be defined as in* $P_A(\mathbf{e}) = (b/a)\sum_{i=1}^{a} \tilde{e}_{i..}^2$, *with* e_{ijm} *replaced by* $Y_{ijm} - E(Y_{ijm})$, *and* $\mathrm{MST}_A(\boldsymbol{Z})$ *be defined as* MST_A *in Equation* (4) *with* $R_{ijm} = Z_{ijm}$. *Then under* $H_0(A)$, *under the settings and assumptions of part (b) of Theorem* 3.1, *as* $a, b \to \infty$, $T_A^*(\boldsymbol{Z} - E(\boldsymbol{Y})) = n(a, b)\sqrt{a}(\mathrm{MST}_A(\boldsymbol{Z}) - P_A(\boldsymbol{Z} - E(\boldsymbol{Y}))) \xrightarrow{p} 0$, *regardless of whether* $n_{ij} \geq 2$ *tend to* ∞ *or stay fixed.*

LEMMA A.5 *Let* $n(a, b, c) = \min_{i,j,k}\{n_{ijk}\}$. *Under the settings and assumptions of Theorem* 3.2,

$$n(a, b, c)[\mathrm{MSE}/N^2 - \sigma_{AB}^2] \xrightarrow{p} 0 \quad \text{as } \max\{a, b, c\} \to \infty,$$

regardless of whether n_{ijk} *remain fixed or tend to infinity.*

LEMMA A.6 *Let* $P_{1,AB}(\boldsymbol{Z} - E(\boldsymbol{Y}))$ *and* $P_{2,AB}(\boldsymbol{Z} - E(\boldsymbol{Y}))$ *be defined as* $P_{1,AB}(\boldsymbol{e}) = (c/ab)\sum_{i=1}^{a}\sum_{j=1}^{b} \tilde{e}_{ij..}^2$, *where* $e_{ijkm} = X_{ijkm} - E(X_{ijkm})$, *and* $P_{2,AB}(\boldsymbol{e}) = P_{1,AB}(\boldsymbol{e}) - (c/ab(b-1))\sum_{i=1}^{a}\sum_{j\neq j'}^{b} \tilde{e}_{ij..}\tilde{e}_{ij'..}$. *with* e_{ijkm} *replaced by* $Y_{ijkm} - E(Y_{ijkm})$, *and* $\mathrm{MST}_{AB}(\boldsymbol{Z})$ *is defined as* MST_{AB} *in Equation* (8) *with* $R_{ijkm} = Z_{ijkm}$. *Then under* $H_0(AB)$,

(1) *under the settings and assumptions of part* (*c*) *of Theorem* 3.2, *as* $a, b \to \infty$,

$$T_1^*(\boldsymbol{Z} - E(\boldsymbol{Y})) = n(a, b, c)\sqrt{ab}(\mathrm{MST}_{AB}(\boldsymbol{Z}) - P_{1,AB}(\boldsymbol{Z} - E(\boldsymbol{Y}))) \xrightarrow{p} 0;$$

(2) *under the settings and assumptions of part* (*b*) *Theorem* 3.2, *as* $a \to \infty$, *with b fixed,*

$$T_2^*(\boldsymbol{Z} - E(\boldsymbol{Y})) = n(a, b, c)\sqrt{a}(\mathrm{MST}_{AB}(\boldsymbol{Z}) - P_{2,AB}(\boldsymbol{Z} - E(\boldsymbol{Y}))) \xrightarrow{p} 0,$$

regardless of whether c and $n_{ijk} \geq 2$ *tend to* ∞ *or stay fixed.*

LEMMA A.7 *Under the settings and assumptions of Theorem* 3.3,

$$n(a, b, c, d)[\mathrm{MSE}/N^2 - \sigma_{ABC}^2] \xrightarrow{p} 0 \quad \text{as } \max\{a, b, c, d\} \to \infty,$$

where $n(a, b, c, d) = \min_{i,j,k,l}\{n_{ijkl}\}$, *regardless of whether* n_{ijkl} *remain fixed or tend to infinity.*