

DISCRETE CONVOLUTION STATISTIC FOR HYPOTHESIS TESTING

Giulio Prevedello, p.giulio@hotmail.it

Ken R. Duffy, ken.duffy@mu.ie

Hamilton Institute

Maynooth University

Maynooth, Ireland

Key Words: discrete convolution; sum of discrete random variables; statistical hypothesis testing; nonparametric maximum-likelihood estimation; sub-independence

Mathematics Subject Classification: 62G05; 62G10; 62G20; 62P10; 62P20

ABSTRACT

The question of testing for equality in distribution between two linear models, each consisting of sums of distinct discrete independent random variables with unequal numbers of observations, has emerged from the biological research. In this case, the computation of classical χ^2 statistics, which would not include all observations, results in loss of power, especially when sample sizes are small. Here, as an alternative that uses all data, the nonparametric maximum likelihood estimator for the distribution of sum of discrete and independent random variables, which we call the convolution statistic, is proposed and its limiting normal covariance matrix determined. To challenge null hypotheses about the distribution of this sum, the generalized Wald's method is applied to define a testing statistic whose distribution is asymptotic to a χ^2 with as many degrees of freedom as the rank of such covariance matrix. Rank analysis also reveals a connection with the roots of the probability generating functions associated to the addend variables of the linear models. A simulation study is performed to compare the convolution test with Pearson's χ^2 , and to provide usage guidelines.

1. INTRODUCTION

In this paper we examine the problem of testing the null hypothesis of equality in distri-

bution, denoted \sim , for two linear models with distinct observables, that is

$$\mathbf{H}_0: a_0 + \sum_{i=1}^k a_i A_i \sim b_0 + \sum_{i=1}^h b_i B_i. \quad (1)$$

We assume that the random variables $A_1, \dots, A_k, B_1, \dots, B_h$ are bounded, independent, of possibly different distribution and all take values in a real lattice $\Lambda(\zeta) = \{\zeta u: u \in \mathbb{Z}\}$ for some $\zeta \in \mathbb{R}$ and that $a_0, b_0 \in \Lambda(\zeta)$, $a_1, \dots, a_k, b_1, \dots, b_h \in \mathbb{Z}$, the set of integers.

Equality in distribution between random variables can be tested using statistics such as Pearson's χ^2 (Pearson, 1900) or the more general power-divergence family (Cressie and Read, 1984). The computation of these statistics, however, assumes that the number of observations of each of A_1, \dots, A_k (and B_1, \dots, B_h) are equal. Otherwise, it would seem that the data sets must be truncated for application of those methods, which could prove wasteful if samples come in unequal counts and their collection is costly or laborious.

For example, consider a problem in meta-analysis, where two studies are described by linear models with distinct independent variables, and we wish to test for equality in distribution between these models as in (1). In the simplest case, for $k = 2$, $h = 1$ with $\mathbf{H}_0: A_1 + A_2 \sim B_1$, the independent variables observed are n_1 distributed as A_1 , n_2 as A_2 and n_3 as B_1 , respectively noted $\{A_{11}, \dots, A_{1n_1}\}$, $\{A_{21}, \dots, A_{2n_2}\}$ and $\{B_{11}, \dots, B_{1n_3}\}$, with $n_1, n_2, n_3 \in \mathbb{N}$. This scenario may arise because the independent variables are grouped differently in the studies (e.g., A_1 occurrences of event E_1 , A_2 occurrences of event E_2 , B_1 occurrences of any event E_1 or E_2) or because the model choice is different (e.g., model one is $A_1 + A_2$ and model two is $B_1 \sim f(A_1, A_2)$ for a given function f). Then, to test \mathbf{H}_0 , Pearson's χ^2 could be computed using B_1, \dots, B_{n_3} and the data from A_1 and A_2 paired as, for example, $\{A_{11} + A_{21}, \dots, A_{1m} + A_{2m}\}$ with $m = n_1 = n_2$, so that these m variables are independent and identically distributed as $A_1 + A_2$ to comply with Pearson's statistic assumptions. If observation sizes are unequal, e.g. $n_1 > n_2$, then $n_1 - n_2 > 0$ variables from $\{A_{11}, \dots, A_{1n_1}\}$ could be excluded from the calculation of $\{A_{11} + A_{21}, \dots, A_{1m} + A_{2m}\}$, now with $m = \min(n_1, n_2)$. But any pairing or variables exclusion are two choices that, either arbitrarily or randomly determined, influences the outcome of the test.

We were personally motivated to address this question during a study on how stimulatory signals are integrated by the immune system (Marchingo et al., 2016). We wished to statistically test the hypothesis that the expansion impetus of two stimuli were integrated independently by cells when the signals were provided together, which had been hypothesized in a previously published study (Marchingo et al., 2014). The experimental data obtained for that work was costly to produce, both in terms of manpower and reagents, and inherently came with distinct numbers of observations of all variables. Thus we sought to develop a statistical procedure that utilized all available data. The resulting test may prove useful in other fields, such as medicine for efficacy evaluation of combination therapies, (e.g. Wolchok et al., 2013), which is a topic of growing interest (Editorial, 2017).

We first show that the null hypothesis in (1) is equivalent to one without the scalar multipliers, $\sum_{i=1}^k X_i \sim \sum_{i=1}^h Y_i$, which simplifies notation (Lemma 1). To obtain a test statistic that utilizes all data and therefore outperforms methods that require equal sized data sets, in Section 2 we study the maximum likelihood estimator (MLE) for the probability mass vector (PMV) of $\sum_{i=1}^k X_i$. This transpires to be the discrete convolution of the empirical probability mass vector (EPMV) of each variable X_1, \dots, X_k and so we refer to it as the “convolution statistic” (Proposition 1).

We then derive the asymptotic distribution of the convolution statistic and build a testing procedure for both goodness-of-fit and equality in distribution versions (Proposition 2), leveraging the generalized Wald’s method. This technique was introduced in Moore’s work (Mihalko and Moore, 1980; Moore, 1977, 1982; Moore and Spruill, 1975), as an extension of Wald’s method (Wald, 1943), to build χ^2 tests for statistics that are asymptotically normal distributed with a singular covariance matrix. It was subsequently adjusted in (Hadi and Wells, 1990), whose version we employ. Such methodology found applications in the fields of econometrics (Andrews, 1987, 1988; Vuong, 1987; Wilson and Koehler, 1991), biology (Marchingo et al., 2016; Zhang, 1999), and statistical theory (Drost, 1989; Tyler, 1981; Voinov et al., 2008).

In Section 3, we investigate the covariance matrix rank asymptotic of the convolution

statistic (Theorem 1, Corollary 1 and Corollary 2), which is the central problem for the derivation of a testing procedure through the generalized Wald's framework. Interestingly, such rank is related to the roots of the probability generating functions of X_1, \dots, X_k and Y_1, \dots, Y_h (Lemma 2). Finally, we show how a test for sub-independence (Hamedani, 2013; Schennach, 2019) can be built from the results achieved above (Corollary 3).

To conclude, in Section 4 we provide simulated performance analysis for the convolution statistic against Person's χ^2 , and in Section 5 we discuss the guidelines for its application. We remark that all the necessary proofs are reported in Appendix.

2. CONVOLUTION STATISTIC

To derive a statistic for the testing of $\mathbf{H}_0: a_0 + \sum_{i=1}^k a_i A_i \sim b_0 + \sum_{i=1}^h b_i B_i$, we begin by showing that this null hypothesis is equivalent to another in which the variables have finite, positive integer support and no parameters $a_0, \dots, a_k, b_0, \dots, b_h$ are present. As a consequence, we will work in this new setting as it facilitates the definition of the convolution statistic for the testing of \mathbf{H}_0 , especially in regard to a simpler notation.

Lemma 1 (Null hypothesis simplification). *Let $A_1, \dots, A_k, B_1, \dots, B_h$ be a sequence of finite and independent random variables that map the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into a lattice $\Lambda(\zeta)$ with $\zeta \in \mathbb{R}$. Given the null hypothesis*

$$\mathbf{H}_0: a_0 + \sum_{i=1}^k a_i A_i \sim b_0 + \sum_{i=1}^h b_i B_i, \quad (2)$$

with $a_0, b_0 \in \Lambda(\zeta)$, $a_1, \dots, a_k, b_1, \dots, b_h \in \mathbb{Z}$, there exists a sequence of positive, finite and independent random variables $X_1, \dots, X_k, Y_1, \dots, Y_h$ from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into $\{0, \dots, r_l\}$, respectively, with $r_l \in \mathbb{N}$ for $l = 1, \dots, k+h$, $\mathbb{P}(X_i = 0) > 0$ for $i = 1, \dots, k$, $\mathbb{P}(Y_j = 0) > 0$ for $j = 1, \dots, h$, and such that

$$\mathbf{H}_0: \sum_{i=1}^k X_i \sim \sum_{i=1}^h Y_i, \quad (3)$$

is equivalent to (2).

As a result of Lemma 1, we need only to consider \mathbf{H}_0 stated in equation (3). Thus given a sequence of $k \geq 2$ integer and independent random variables X_1, \dots, X_k we write, for fixed $i \in \{1, \dots, k\}$, that $X_i \sim \mathbf{x}_i \in \Delta^{r_i}$ with

$$\Delta^{r_i} = \left\{ \mathbf{v} = (v_0, \dots, v_{r_i}) \in \mathbb{R}^{r_i+1}: v_0, v_{r_i} \in (0, 1); v_j \geq 0, j = 1, \dots, r_i - 1; \sum_{j=1}^{r_i} v_j = 1 \right\}$$

to indicate that X_i takes values in $\{0, \dots, r_i\} \subseteq \mathbb{N} \cup \{0\}$, with $r_i > 0$, and is distributed with PMV $\mathbf{x}_i = (x_{i0}, \dots, x_{ir_i})$, that is $\mathbb{P}(X_i = j) = x_{ij}$ for $j = 0, \dots, r_i$.

We remark that $x_{i0}, x_{ir_i} \in (0, 1)$, for every $i = 1, \dots, k$, are assumed to avoid degenerate cases, without loss of generality. In fact, $x_{i0} > 0$ descends from Lemma 1. Moreover, given any $t \leq k$ such that $x_{tr_t} = \mathbb{P}(X_t = r_t) = 0$, there exists $\tau = \max\{j: x_{tj} > 0\} < r_t$, so that X_t can be replaced by $\tilde{X}_t \sim \tilde{\mathbf{x}}_t = (\tilde{x}_{t0}, \dots, \tilde{x}_{t\tau}) \in \Delta^\tau$, with $\tilde{x}_{tj} = x_{tj}$ for $j = 0, \dots, \tau$. Lastly, the constraint $x_{i0}, x_{ir_i} < 1$ for every $i = 1, \dots, k$ ensures that X_i is not a constant value.

From now on, with these assumptions and notation, for the null hypothesis of goodness-of-fit test we consider

$$\mathbf{H}_0: \sum_{i=1}^k X_i \sim \mathbf{z} \quad (4)$$

with $s = \sum_{i=1}^k r_i$ and $\mathbf{z} \in \Delta^s$. By independence, the sum of X_1, \dots, X_k is distributed as the discrete convolution, denoted $*$, of their PMVs, that is

$$\sum_{i=1}^k X_i \sim \mathbf{x}_1 * \dots * \mathbf{x}_k,$$

where, for any two vectors $\mathbf{v} = (v_0, \dots, v_a) \in \mathbb{R}^{a+1}$, $\mathbf{w} = (w_0, \dots, w_b) \in \mathbb{R}^{b+1}$ with $a, b > 0$, we have $\mathbf{v} * \mathbf{w} \in \mathbb{R}^{a+b+1}$ and $(\mathbf{v} * \mathbf{w})_i = \sum_{j=0}^a \sum_{l=0}^b v_j w_l \delta_{j+l, i}$ with $\delta_{i,j} = 1$ if $i = j$ and being null otherwise. Hence, the null hypothesis for the goodness-of-fit-test (4) is equivalent to

$$\mathbf{H}_0: \mathbf{x}_1 * \dots * \mathbf{x}_k = \mathbf{z}.$$

Our first goal is to determine a statistic to test (4), which will subsequently be extended to assess the equality in distribution between $\sum_{i=1}^k X_i$ and $\sum_{i=1}^h Y_i$, where $Y_1 \sim \mathbf{y}_1 \in \Delta^{r_{k+1}}, \dots, Y_h \sim \mathbf{y}_h \in \Delta^{r_{k+h}}$ are other $h \geq 1$ independent random variables, with $s =$

$\sum_{i=1}^k r_i = \sum_{i=1}^h r_{k+i}$, namely

$$\mathbf{H}_0: \sum_{i=1}^k X_i \sim \sum_{i=1}^h Y_i, \quad (5)$$

or, equivalently,

$$\mathbf{H}_0: \mathbf{x}_1 * \dots * \mathbf{x}_k = \mathbf{y}_1 * \dots * \mathbf{y}_h.$$

When the PMVs \mathbf{x}_i for $i = 1, \dots, k$ and \mathbf{y}_j for $j = 1, \dots, h$ are unknown, care must be taken to define the test statistics for (4) and (5) based only on available information. In regard, the data consist of the observation of n_i independent random variables $\{X_{i1}, \dots, X_{in_i}\}$ identically distributed as X_i for $i = 1, \dots, k$ and n_{k+i} independent random variables $\{Y_{i1}, \dots, Y_{in_{k+i}}\}$ identically distributed as Y_i for $i = 1, \dots, h$. Following a nonparametric approach, we fix $i \in \{1, \dots, k\}$ and define $\hat{\mathbf{x}}_{in_i} \in \Delta^{r_i}$ the MLE of \mathbf{x}_i , that is

$$(\hat{\mathbf{x}}_{in_i})_u = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbb{1}_{\{X_{ij}=u\}}$$

for $u = 0, \dots, r_i$, with $\mathbb{1}_A$ being the indicator function of the event A . In particular, by the multivariate central limit theorem (Serfling, 1980), $\sqrt{n_i}(\hat{\mathbf{x}}_{in_i} - \mathbf{x}_i)$ is asymptotically distributed as a centered normal random variable with covariance $\Sigma(\mathbf{x}_i)$ for n_i large, namely $\sqrt{n_i}(\hat{\mathbf{x}}_{in_i} - \mathbf{x}_i) \sim_{n_i \rightarrow \infty} \mathcal{N}(\Sigma(\mathbf{x}_i))$, where, for any PMV $\mathbf{v} = (v_0, \dots, v_a) \in \Delta^a$ and $a \geq 0$, we define $\Sigma(\mathbf{v}) \in \mathbb{R}^{a+1} \times \mathbb{R}^{a+1}$ such that $(\Sigma(\mathbf{v}))_{ij} = v_i \delta_{i,j} - v_i v_j$ for $i, j = 0, \dots, a$. With this notation, we derive the MLE for the distribution of $\sum_{i=1}^k X_i$.

Proposition 1 (MLE for a sum of independent random variables). *Given $k \geq 2$, let $\{X_{i1}, \dots, X_{in_i}\}$ be $n_i \in \mathbb{N}$ random variables independent and identically distributed as $X_i \sim \mathbf{x}_i \in \Delta^{r_i}$ with $r_i \in \mathbb{N}$, for $i = 1, \dots, k$. Set $s = \sum_{i=1}^k r_i$. The MLE for the PMV $\mathbf{x}_1 * \dots * \mathbf{x}_k \in \Delta^s$ of $\sum_{i=1}^k X_i$ is $\hat{\mathbf{x}}_{1n_1} * \dots * \hat{\mathbf{x}}_{kn_k} \in \Delta^s$, defined as*

$$(\hat{\mathbf{x}}_{1n_1} * \dots * \hat{\mathbf{x}}_{kn_k})_u = \left(\prod_{j=1}^k n_j \right)^{-1} \sum_{i_1=1}^{n_1} \dots \sum_{i_k=1}^{n_k} \mathbb{1}_{\{\sum_{j=1}^k X_{ji_j}=u\}},$$

for every $u = 0, \dots, s$.

Proposition 1 shows that the MLE for $\mathbf{x}_1 * \dots * \mathbf{x}_k$ is calculated using all observables X_{ij} for $j = 1, \dots, n_i$ and $i = 1, \dots, k$, which may not be the case for Person's statistic, as explained in Section 1, if at least one of the sample sizes n_1, \dots, n_k differs from another.

We introduce additional notation for what follows: \mathbf{A}^+ , \mathbf{A}' , $\text{Ker}(\mathbf{A})$, $\text{nul}(\mathbf{A})$, $\text{rk}(\mathbf{A})$ for the Moore-Penrose inverse, transpose, kernel, nullity and rank of a matrix \mathbf{A} , respectively (Hagen et al., 2000; Horn and Johnson, 1986); $\chi^2(s)$ to indicate the χ^2 distribution with $s > 0$ degrees of freedom; $\mathbf{T}^{b+1}(v) \in \mathbb{R}^{b+1} \times \mathbb{R}^{a+b+1}$ for the matrix of the discrete convolution between $\mathbf{v} \in \mathbb{R}^{a+1}$ and any $b+1$ -dimensional vector, i.e. $\mathbf{T}^{b+1}(\mathbf{v})\mathbf{w} = \mathbf{v} * \mathbf{w} \in \mathbb{R}^{a+b+1}$ with $\mathbf{w} \in \mathbb{R}^{b+1}$, given $a, b \geq 0$. We write $\mathbf{T}(\mathbf{v})$ without explicit domain dimension if this is clear from the context.

We are now ready to determine the asymptotic behavior of $\hat{\mathbf{x}}_{1n_1} * \dots * \hat{\mathbf{x}}_{kn_k}$, which follows from an application of the delta method as well as properties of quadratic transformation of asymptotically multivariate normal vectors (Serfling, 1980). In order for the MLE $\hat{\mathbf{x}}_{1n_1} * \dots * \hat{\mathbf{x}}_{kn_k}$ to converge to $\mathbf{x}_1 * \dots * \mathbf{x}_k$ it is necessary that the sample sizes n_1, \dots, n_k grow with proportional rates. For this reason, from now on, we set

$$m = \min(n_1, \dots, n_{k+h}) \text{ and assume } c_i = \lim_{m \rightarrow \infty} \frac{m}{n_i}$$

is finite and positive for every $i = 1, \dots, k + h$.

Proposition 2 (Asymptotic normality of convolutions). *Under the null hypothesis (4), that $\mathbf{x}_1 * \dots * \mathbf{x}_k = \mathbf{z}$, it holds that*

$$\mathbf{V}_m = \sqrt{m} (\hat{\mathbf{x}}_{1n_1} * \dots * \hat{\mathbf{x}}_{kn_k} - \mathbf{z}) \underset{m \rightarrow \infty}{\sim} \mathcal{N}(\Psi) \quad (6)$$

and

$$\mathbf{V}_m' \Psi^+ \mathbf{V}_m \underset{m \rightarrow \infty}{\sim} \chi^2(\text{rk}(\Psi)) \quad (7)$$

where $\Psi = \sum_{i=1}^k c_i \mathbf{T}(\mathbf{x}_{(i)}) \Sigma(\mathbf{x}_{(i)}) \mathbf{T}(\mathbf{x}_{(i)})'$ and $\mathbf{x}_{(i)} = \mathbf{x}_1 * \dots * \mathbf{x}_{i-1} * \mathbf{x}_{i+1} * \dots * \mathbf{x}_k$ for $i = 1, \dots, k$. Alternatively, under the null hypothesis (5), that $\mathbf{x}_1 * \dots * \mathbf{x}_k = \mathbf{y}_1 * \dots * \mathbf{y}_h$, it holds that

$$\mathbf{W}_m = \sqrt{m} (\hat{\mathbf{x}}_{1n_1} * \dots * \hat{\mathbf{x}}_{kn_k} - \hat{\mathbf{y}}_{1n_{k+1}} * \dots * \hat{\mathbf{y}}_{hn_{k+h}}) \underset{m \rightarrow \infty}{\sim} \mathcal{N}(\Psi + \Xi) \quad (8)$$

and

$$\mathbf{W}_m'(\boldsymbol{\Psi} + \boldsymbol{\Xi})^+ \mathbf{W}_m \underset{m \rightarrow \infty}{\sim} \chi^2(\text{rk}(\boldsymbol{\Psi} + \boldsymbol{\Xi})) \quad (9)$$

where $\boldsymbol{\Xi} = \sum_{i=1}^h c_{k+i} \mathbf{T}(\mathbf{y}_{(i)}) \boldsymbol{\Sigma}(\mathbf{y}_{(i)}) \mathbf{T}(\mathbf{y}_{(i)})'$ and $\mathbf{y}_{(i)} = \mathbf{y}_1 * \dots * \mathbf{y}_{i-1} * \mathbf{y}_{i+1} * \dots * \mathbf{y}_k$ for $i = 1, \dots, h$.

We remark that expressions (7) and (9) require the knowledge of $\boldsymbol{\Psi}$ and $\boldsymbol{\Psi} + \boldsymbol{\Xi}$, but these may, in general, be unknown. Thus we take advantage of the generalized Wald's method (Moore, 1977), which shows how to construct χ^2 tests from consistent estimators of the covariance variance matrices such as $\boldsymbol{\Psi}$ and $\boldsymbol{\Psi} + \boldsymbol{\Xi}$. We recall here Moore (1977, Theorem 2) which will serve as backbone for the subsequent results.

Proposition 3 (Generalized Wald's method; Moore (1977), Theorem 2). *Suppose a sequence of estimators $\{\hat{\boldsymbol{\theta}}_m\}_{m \geq 1}$ of a parameter $\boldsymbol{\theta}_0 \in \mathbb{R}^d$, with $d > 0$, is such that*

$$\sqrt{m} \left(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0 \right) \underset{m \rightarrow \infty}{\sim} \mathcal{N}(\boldsymbol{\Sigma})$$

with $\text{rk}(\boldsymbol{\Sigma}) \leq d$. Noted $\{\mathbf{B}_m\}_{m \geq 1}$ a sequence of d -dimensional square matrices such that $\mathbf{B}_m \underset{m \rightarrow \infty}{\sim} \mathbf{B}$ with \mathbf{B} generalized-inverse of $\boldsymbol{\Sigma}$, then

$$m \left(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0 \right)' \mathbf{B}_m \left(\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0 \right) \underset{m \rightarrow \infty}{\sim} \chi^2(\text{rk}(\boldsymbol{\Sigma})).$$

Since the entries of

$$\hat{\boldsymbol{\Psi}}_m = \sum_{i=1}^k c_i \mathbf{T}(\hat{\mathbf{x}}_{(i)n_i}) \boldsymbol{\Sigma}(\hat{\mathbf{x}}_{in_i}) \mathbf{T}(\hat{\mathbf{x}}_{(i)n_i})'$$

are continuous functions of $\mathbf{x}_1, \dots, \mathbf{x}_k$, whose consistent estimator are $\hat{\mathbf{x}}_{1n_1}, \dots, \hat{\mathbf{x}}_{kn_k}$ respectively, then $\hat{\boldsymbol{\Psi}}_m$ is a consistent estimator of $\boldsymbol{\Psi}$ and similarly

$$\hat{\boldsymbol{\Xi}}_m = \sum_{i=1}^h c_{k+i} \mathbf{T}(\hat{\mathbf{y}}_{(i)n_{k+i}}) \boldsymbol{\Sigma}(\hat{\mathbf{y}}_{in_{k+i}}) \mathbf{T}(\hat{\mathbf{y}}_{(i)n_{k+i}})'$$

for $\boldsymbol{\Xi}$.

Note that Proposition 3 cannot be directly applied to (6) by setting $\mathbf{B}_m = \hat{\boldsymbol{\Psi}}_m^+$, as $\hat{\boldsymbol{\Psi}}_m^+$ may not be a consistent estimator of $\boldsymbol{\Psi}^+$. Given a sequence of consistent estimators

$\{\mathbf{A}_m\}_{m \geq 1}$ for a matrix \mathbf{A} of finite dimensions, then $\{\mathbf{A}_m^+\}_{m \geq 1}$ is a sequence of consistent estimators for \mathbf{A}^+ if and only if $\text{rk}(\mathbf{A}_m) = \text{rk}(\mathbf{A})$ for m large (Nashed, 1976). In particular, as the rank is a lower-semicontinuous operator on the space of finite dimensional matrices, then only $\text{rk}(\hat{\Psi}_m) \geq \text{rk}(\Psi)$ is guaranteed as m tends to infinity.

If the limiting rank is known, consistency is ensured by Eckart-Young-Mirsky's theorem (Eckart and Young, 1936), which is the solution to the basic low rank approximation of a finite dimensional matrix (Markovsky, 2012). To this end, given any d -dimensional symmetric matrix $\mathbf{A} \in \mathbb{R}^d \times \mathbb{R}^d$ with $d \in \mathbb{N}$ and its eigendecomposition $\mathbf{A} = \mathbf{P}'\mathbf{\Lambda}\mathbf{P}$, with $0 < \text{rk}(\mathbf{A}) \leq d$, $\mathbf{\Lambda}$ diagonal matrix of the decreasing eigenvalues and \mathbf{P} orthogonal matrix, then for any $0 < r \leq \text{rk}(\mathbf{A})$ we define a rank- r matrix that approximates \mathbf{A} (in light of the Eckart-Young-Mirsky theorem) as $\mathbf{A}^r = (\mathbf{D}^r\mathbf{P})'\mathbf{\Lambda}^r\mathbf{D}^r\mathbf{P} \in \mathbb{R}^d \times \mathbb{R}^d$, where \mathbf{D}^r is a $\mathbb{R}^r \times \mathbb{R}^d$ matrix with 1 at the diagonal and 0 elsewhere and $\mathbf{\Lambda}^r$ is the $\mathbb{R}^r \times \mathbb{R}^r$ diagonal matrix of the largest r eigenvalues of Σ . In particular, \mathbf{A}^r may not be unique, as for the case when the r^{th} and $r+1^{\text{th}}$ eigenvalues are equal. The following result is found in Hadi and Wells (1990, Theorem 2.3) for the generalized inverses and we report it here for the case of Moore-Penrose inverses.

Proposition 4 (Rank approximation; Hadi and Wells (1990), Theorem 2.3). *Suppose a sequence of centered random variables $\{\mathbf{U}_m\}_{m \geq 1} \in \mathbb{R}^d$ is asymptotically distributed as $\mathcal{N}(\Sigma)$ for m large, with $0 < \text{rk}(\Sigma) \leq d$ where $d > 0$. Let $\{\hat{\Sigma}_m\}_{m \geq 1}$ be a sequence of square matrices that are consistent estimators of Σ , then for every $0 < r \leq \text{rk}(\Sigma)$*

$$\mathbf{U}_m'(\hat{\Sigma}_m^r)^+\mathbf{U}_m \underset{m \rightarrow \infty}{\sim} \chi^2(r),$$

where $\hat{\Sigma}_m^r$ is a rank- r approximation of $\hat{\Sigma}_m$.

Proposition 4 highlights the central role of the rank of Σ , which will be derived in the next section for $\Sigma = \Psi$ and $\Sigma = \Psi + \Xi$. In general, the determination of $\text{rk}(\Sigma)$ may be a difficult problem that depends on the structure of the Σ under consideration, and this limitation may explain why an otherwise flexible tool such as the generalized Wald's method from Proposition 3 is not more widely employed. But, if the rank is known, Proposition 4 provides a method for statistical testing null hypotheses, such as $\mathbf{H}_0: \mathbf{x}_1 * \dots * \mathbf{x}_k = \mathbf{z}$

and $\mathbf{H}_0: \mathbf{x}_1 * \dots * \mathbf{x}_k = \mathbf{y}_1 * \dots * \mathbf{y}_h$, under which $\mathbf{\Sigma}$ is not invertible. This result also assures a solution if only a lower bound of the rank is given, at the cost of statistical power. Furthermore, the exclusion of smaller eigenvalues may still be necessary to achieve numerical stability when calculating the pseudo-inverse of $\hat{\mathbf{\Sigma}}_m$, as due to Proposition 4, the effect of that truncation can be accounted for in the statistic formulation.

3. DETERMINING THE COVARIANCE MATRIX RANK

In this section, we investigate the rank of $\mathbf{\Psi}$ and $\mathbf{\Psi} + \mathbf{\Xi}$, the covariance matrices from (6) and (8) of Proposition 2, in order to derive the number of degrees of freedom from the limiting statistics for the goodness-of-fit (7) and equality in distribution (9) tests. Focusing on $\mathbf{\Psi}$, we begin by showing that $c_i \mathbf{T}(\mathbf{x}_{(i)}) \mathbf{\Sigma}(\mathbf{x}_i) \mathbf{T}(\mathbf{x}_{(i)})'$ is a positive semidefinite matrix for any fixed $i \in \{1, \dots, k\}$. In fact, since $\mathbf{\Sigma}(\mathbf{x}_i)$ is positive semidefinite, for every $\mathbf{v} \in \mathbb{R}^{s+1}$

$$c_i \mathbf{v}' \mathbf{T}(\mathbf{x}_{(i)}) \mathbf{\Sigma}(\mathbf{x}_i) \mathbf{T}(\mathbf{x}_{(i)})' \mathbf{v} = c_i \mathbf{w}' \mathbf{\Sigma}(\mathbf{x}_i) \mathbf{w} \geq 0, \quad (10)$$

where $\mathbf{w} = \mathbf{T}(\mathbf{x}_{(i)})' \mathbf{v}$. Additionally, we deduce from Horn and Johnson (1986, Observation 7.1.3) that given \mathbf{A} and \mathbf{B} two positive semidefinite matrices with the same dimensions, then

$$\text{Ker}(\mathbf{A} + \mathbf{B}) = \text{Ker}(\mathbf{A}) \cap \text{Ker}(\mathbf{B}). \quad (11)$$

Taken together, (10) and (11) imply

$$\text{Ker}(\mathbf{\Psi}) = \bigcap_{i=1}^k \text{Ker}(\mathbf{T}(\mathbf{x}_{(i)}) \mathbf{\Sigma}(\mathbf{x}_i) \mathbf{T}(\mathbf{x}_{(i)})'). \quad (12)$$

Using kernel properties, we write

$$\text{Ker}(\mathbf{T}(\mathbf{x}_{(i)}) \mathbf{\Sigma}(\mathbf{x}_i) \mathbf{T}(\mathbf{x}_{(i)})') = \text{Ker}(\mathbf{T}(\mathbf{x}_{(i)})') \oplus \{\mathbf{v} \in \mathbb{R}^{s+1} : \mathbf{T}(\mathbf{x}_{(i)})' \mathbf{v} \in \text{Ker}(\mathbf{\Sigma}(\mathbf{x}_i))\} \quad (13)$$

where \oplus represents the direct sum operation.

Let $i \in \{1, \dots, k\}$ be fixed and let $L_i = \{l : x_{il} = 0\}$ be the set of indexes of the null entries of $\mathbf{x}_i \in \Delta^{r_i}$. In general, the kernel of $\mathbf{\Sigma}(\mathbf{x}_i)$ is generated by the $r_i + 1$ -dimensional all-ones vector $\mathbf{1}_{r_i}$ and the canonical vectors $\mathbf{e}_l^i = (e_{l0}^i, \dots, e_{lr_i}^i) \in \mathbb{R}^{r_i+1}$ for every $l \in L_i$, where $e_{lu}^i = \delta_{l,u}$ for $u = 0, \dots, r_i$, that is $\text{Ker}(\mathbf{\Sigma}(\mathbf{x}_i)) = \langle \mathbf{1}_{r_i} \rangle \oplus E_i$, where $E_i = \langle \{\mathbf{e}_l^i : l \in L_i\} \rangle$.

Denoting $r_{(i)} = \sum_{j \neq i}^k r_j$, we can expand $\mathbf{T}(\mathbf{x}_{(i)})$ into

$$\mathbf{T}(\mathbf{x}_{(i)}) = \begin{bmatrix} x_{(i)0} & 0 & \dots & 0 \\ x_{(i)1} & x_{(i)0} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ x_{(i)r_{(i)}} & x_{(i)r_{(i)}-1} & \ddots & x_{(i)0} \\ 0 & x_{(i)r_{(i)}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & x_{(i)r_{(i)}} \end{bmatrix},$$

from which we deduce $\mathbf{T}(\mathbf{x}_{(i)})'\mathbf{1}_s = \mathbf{1}_{r_i}$, for every $\mathbf{x}_{(i)} \in \Delta^{r_{(i)}}$, and in particular $\mathbf{1}_s \notin \text{Ker}(\mathbf{T}(\mathbf{x}_{(i)}'))$.

To achieve an explicit formulation for the rank of Ψ (and analogously for $\Psi + \Xi$), in Theorem 1 we will assume that $\mathbf{x}_i \in \Delta_{\text{Int}}^{r_i} \subset \Delta^{r_i}$, defined as

$$\Delta_{\text{Int}}^{r_i} = \{\mathbf{v} = (v_0, \dots, v_{r_i}) \in \mathbb{R}^{r_i+1} : \sum_{j=1}^{r_i} v_j = 1; 0 < v_j < 1, j = 0, \dots, r_i\}$$

for every $i = 1, \dots, k$. This ensures that $E_i = \emptyset$, so that $\text{Ker}(\Sigma(\mathbf{x}_i)) = \langle \mathbf{1}_{r_i} \rangle$. Under this hypothesis, from (12) and (13) we deduce that

$$\text{Ker}(\Psi) = \bigcap_{i=1}^k \text{Ker}(\mathbf{T}(\mathbf{x}_{(i)})\Sigma(\mathbf{x}_i)\mathbf{T}(\mathbf{x}_{(i)}')) = \langle \mathbf{1}_s \rangle \oplus \bigcap_{i=1}^k \text{Ker}(\mathbf{T}(\mathbf{x}_{(i)}')), \quad (14)$$

and, with the same reasoning applied to $\Psi + \Xi$, it follows that

$$\text{Ker}(\Psi + \Xi) = \langle \mathbf{1}_s \rangle \oplus \left(\left(\bigcap_{i=1}^k \text{Ker}(\mathbf{T}(\mathbf{x}_{(i)}')) \right) \cap \left(\bigcap_{j=1}^h \text{Ker}(\mathbf{T}(\mathbf{y}_{(j)}')) \right) \right). \quad (15)$$

In the following Lemma we show how

$$\bigcap_{i=1}^k \text{Ker}(\mathbf{T}(\mathbf{x}_{(i)}')) \quad (16)$$

and

$$\left(\bigcap_{i=1}^k \text{Ker}(\mathbf{T}(\mathbf{x}_{(i)}')) \right) \cap \left(\bigcap_{j=1}^h \text{Ker}(\mathbf{T}(\mathbf{y}_{(j)}')) \right) \quad (17)$$

depend on the roots in common between the probability generating functions of the random variables $X_1, \dots, X_k, Y_1, \dots, Y_h$. This result will be achieved in full generality without restrictions for the PMVs, that is with $\mathbf{x}_1 \in \Delta^{r_1}, \dots, \mathbf{x}_k \in \Delta^{r_k}, \mathbf{y}_1 \in \Delta^{r_{k+1}}, \dots, \mathbf{y}_h \in \Delta^{r_{k+h}}$. We first provide some insight into this connection by considering the case $k = 2$

$$\text{Ker}(\mathbf{T}(\mathbf{x}_{(1)})') \cap \text{Ker}(\mathbf{T}(\mathbf{x}_{(2)})') = \text{Ker}\left(\left[\mathbf{T}(\mathbf{x}_2) \quad \mathbf{T}(\mathbf{x}_1)\right]'\right).$$

In fact, $\left[\mathbf{T}(\mathbf{x}_2) \quad \mathbf{T}(\mathbf{x}_1)\right] \in \mathbb{R}^{r_1+r_2+2} \times \mathbb{R}^{r_1+r_2+1}$ has the same structure, with different dimensions, of a Sylvester matrix (Markovsky, 2012), whose nullity is the degree of the polynomial from the greatest common divisor of the probability generating functions associated to \mathbf{x}_1 and \mathbf{x}_2 .

To formalize the connection with PMVs and polynomials, we introduce the bijection $\varphi: \cup_{a \geq 0} \{\mathbf{u} = (u_0, \dots, u_a) \in \mathbb{R}^{a+1}: \sum_{i=0}^a u_i = 1; u_a \neq 0\} \rightarrow \{u(t) \in \mathbb{R}[t]: u(1) = 1\}$ that, for any $a \geq 0$, maps a vector $\mathbf{v} = (v_0, \dots, v_a) \in \{\mathbf{u} = (u_0, \dots, u_a) \in \mathbb{R}^{a+1}: \sum_{i=0}^a u_i = 1; u_a \neq 0\}$ to the polynomial $\varphi(\mathbf{v})(t) = \sum_{i=0}^a v_i t^i \in \mathbb{R}[t]$ of degree $\deg \varphi(\mathbf{v}) = a$ with coefficients \mathbf{v} . In particular, this map transforms the convolution of vectors into the product of polynomials: given $\mathbf{w} \in \{\mathbf{u} = (u_0, \dots, u_b) \in \mathbb{R}^{b+1}: \sum_{i=0}^b u_i = 1; u_b \neq 0\}$, for any $b \geq 0$, we have

$$\varphi(\mathbf{v} * \mathbf{w})(t) = \sum_{i=0}^a (\mathbf{v} * \mathbf{w})_i t^i = \varphi(\mathbf{v})(t) \varphi(\mathbf{w})(t).$$

The map φ allows the extension of the notion of greatest common divisor between any two polynomials $\gcd(\varphi(\mathbf{v}), \varphi(\mathbf{w}))$ to their related vectors \mathbf{v}, \mathbf{w} . This is achieved by establishing $\gcd(u_1(t), u_2(t)) \in \{u(t) \in \mathbb{R}[t]: u(1) = 1\}$ for any $u_1(t), u_2(t) \in \{u(t) \in \mathbb{R}[t]: u(1) = 1\}$, so that the greatest common divisor is uniquely defined, and by setting, for any $\mathbf{v}, \mathbf{w} \in \cup_{a \geq 0} \{\mathbf{u} = (u_0, \dots, u_a) \in \mathbb{R}^{a+1}: \sum_{i=0}^a u_i = 1; u_a \neq 0\}$,

$$\gcd(\mathbf{v}, \mathbf{w}) = \varphi^{-1}(\gcd(\varphi(\mathbf{v})(t), \varphi(\mathbf{w})(t))) \in \{\mathbf{u} = (u_0, \dots, u_{r_g}) \in \mathbb{R}^{r_g+1}: \sum_{i=0}^{r_g} u_i = 1; u_{r_g} \neq 0\}$$

with $r_g = \deg \gcd(\varphi(\mathbf{v})(t), \varphi(\mathbf{w})(t)) \geq 0$. In particular, we say the vectors \mathbf{v}, \mathbf{w} are coprime if and only if $r_g = 0$. Following the same logic, we import the concept of least common multiple between \mathbf{v} and \mathbf{w} , denoted $\text{lcm}(\mathbf{v}, \mathbf{w})$, and the property of divisibility between vectors. Of

note, with the notation above, the probability generating functions of $\mathbf{x}_1, \dots, \mathbf{x}_k, \mathbf{y}_1, \dots, \mathbf{y}_h$ are, respectively, $\varphi(\mathbf{x}_1), \dots, \varphi(\mathbf{x}_k), \varphi(\mathbf{y}_1), \dots, \varphi(\mathbf{y}_h)$.

We now establish the relation between the kernels of (16), (17) and the greatest common divisors $\mathbf{g}_k = \text{gcd}(\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(k)})$ and $\bar{\mathbf{g}}_h = \text{gcd}(\mathbf{y}_{(1)}, \dots, \mathbf{y}_{(h)})$.

Lemma 2 (Kernels from gcd of PMVs). *Let $k \geq 2$ and $\mathbf{x}_1 \in \Delta^{r_1}, \dots, \mathbf{x}_k \in \Delta^{r_k}$. Given $\mathbf{g}_k = \text{gcd}(\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(k)}) \in \mathbb{R}^{r_{g_k}+1}$, it holds that $\mathbf{T}(\mathbf{g}_k)' \in \mathbb{R}^{\sum_{i=1}^k r_i - r_{g_k} + 1} \times \mathbb{R}^{\sum_{i=1}^k r_i + 1}$ and*

$$\bigcap_{i=1}^k \text{Ker}(\mathbf{T}(\mathbf{x}_{(i)}))' = \text{Ker}(\mathbf{T}(\mathbf{g}_k))'. \quad (18)$$

Additionally, let $h \geq 1$ and $\mathbf{y}_1 \in \Delta^{r_{k+1}}, \dots, \mathbf{y}_h \in \Delta^{r_{k+h}}$. Given $\bar{\mathbf{g}}_h = \text{gcd}(\mathbf{y}_{(1)}, \dots, \mathbf{y}_{(h)}) \in \mathbb{R}^{r_{\bar{g}_h}+1}$ and $\tilde{\mathbf{g}} = \text{gcd}(\mathbf{g}_k, \bar{\mathbf{g}}_h) \in \mathbb{R}^{r_{\tilde{g}}+1}$, it holds that $\mathbf{T}(\bar{\mathbf{g}}_h)' \in \mathbb{R}^{\sum_{i=1}^k r_i - r_{\bar{g}_h} + 1} \times \mathbb{R}^{\sum_{i=1}^k r_i + 1}$, $\mathbf{T}(\tilde{\mathbf{g}})' \in \mathbb{R}^{\sum_{i=1}^k r_i - r_{\tilde{g}} + 1} \times \mathbb{R}^{\sum_{i=1}^k r_i + 1}$ and

$$\bigcap_{i=1}^k \text{Ker}(\mathbf{T}(\mathbf{x}_{(i)}))' \cap \bigcap_{j=1}^h \text{Ker}(\mathbf{T}(\mathbf{y}_{(j)}))' = \text{Ker} \left(\begin{bmatrix} \mathbf{T}(\mathbf{g}_k)' \\ \mathbf{T}(\bar{\mathbf{g}}_h)' \end{bmatrix} \right) = \text{Ker}(\mathbf{T}(\tilde{\mathbf{g}}))' \quad (19)$$

As consequence of Lemma 2, we can finally determine the rank of the covariance matrices Ψ and $\Psi + \Xi$ assuming $\mathbf{x}_1 \in \Delta_{\text{Int}}^{r_1}, \dots, \mathbf{x}_k \in \Delta_{\text{Int}}^{r_k}, \mathbf{y}_1 \in \Delta_{\text{Int}}^{r_{k+1}}, \dots, \mathbf{y}_h \in \Delta_{\text{Int}}^{r_{k+h}}$, in order to calculate the number of degrees of freedom of the limiting χ^2 distribution in (7) and (9) from Proposition 2 for this case.

Theorem 1 (Covariance matrix rank). *Under the assumptions of Proposition 2 and the notations of Lemma 2, and given $\mathbf{x}_1 \in \Delta_{\text{Int}}^{r_1}, \dots, \mathbf{x}_k \in \Delta_{\text{Int}}^{r_k}, \mathbf{y}_1 \in \Delta_{\text{Int}}^{r_{k+1}}, \dots, \mathbf{y}_h \in \Delta_{\text{Int}}^{r_{k+h}}$, it follows that*

$$\text{Ker}(\Psi) = \langle \mathbf{1}_s \rangle \oplus \text{Ker}(\mathbf{T}(\mathbf{g}_k))' \quad (20)$$

and

$$\text{Ker}(\Psi + \Xi) = \langle \mathbf{1}_s \rangle \oplus \text{Ker}(\mathbf{T}(\tilde{\mathbf{g}}))'. \quad (21)$$

In particular, with s defined immediately prior to equation (5),

$$\text{rk}(\Psi) = s - r_{g_k} \quad (22)$$

and

$$\text{rk}(\Psi + \Xi) = s - r_{\tilde{g}}. \quad (23)$$

In the case where $\mathbf{x}_1, \dots, \mathbf{x}_k, \mathbf{y}_1, \dots, \mathbf{y}_h$ are coprime, that is when their probability generating functions $\varphi(\mathbf{x}_1), \dots, \varphi(\mathbf{x}_k), \varphi(\mathbf{y}_1), \dots, \varphi(\mathbf{y}_h)$ have no root in common, we can simplify Theorem 1 as follows.

Corollary 1 (Rank from the coprime case). *Under the assumptions of Theorem 1, if $\mathbf{x}_1, \dots, \mathbf{x}_k, \mathbf{y}_1, \dots, \mathbf{y}_h$ are coprime vectors, then $\text{rk}(\Psi) = \text{rk}(\Psi + \Xi) = s$, where s is defined immediately prior to equation (5).*

A statistic for the goodness-of-fit (4) and equality in distribution (5) tests can also be built for the general case where $\mathbf{x}_i \in \Delta^{r_i}$ and $\mathbf{y}_i \in \Delta^{r_{i+k}}$, leveraging Proposition 4 and the following lower bound for the covariance matrix rank.

Corollary 2 (Rank lower bound for the general case). *Under the assumptions of Proposition 2 and the notations of Lemma 2, and given $\mathbf{x}_1 \in \Delta^{r_1}, \dots, \mathbf{x}_k \in \Delta^{r_k}, \mathbf{y}_1 \in \Delta^{r_{k+1}}, \dots, \mathbf{y}_h \in \Delta^{r_{k+h}}$, it follows that*

$$\text{rk}(\Psi) \geq s - r_{gk} - \sum_{i=1}^k |L_i| \quad (24)$$

and

$$\text{rk}(\Psi + \Xi) \geq s - r_{\tilde{g}} - \sum_{i=1}^{k+h} |L_i|, \quad (25)$$

where $L_i = \{l: x_{il} = 0\}$ for $i = 1, \dots, k$, $L_{i+k} = \{l: y_{il} = 0\}$ for $i = 1, \dots, h$, with $|L_i|, |L_{i+k}|$ indicating their respective cardinality, and s is defined immediately prior to equation (5).

Taken together, Lemma 2 and Theorem 1 serve as example of how to determine the covariance matrix rank upon application of the generalized Wald's framework. For example, assuming the variables $\{X_{1j}, \dots, X_{kj}\}$ for $j = 1, \dots, m$ are grouped in m k -tuples (thus with $m = n_1 = \dots = n_k$), each drawn independently from (X_1, \dots, X_k) , a statistic can be built to test the null hypothesis that the random variables X_1, \dots, X_k are sub-independent (Hamedani, 2013), that is

$$\mathbf{H}_0: \psi_{\sum_{i=1}^k X_i}(t) = \prod_{i=1}^k \psi_{X_i}(t), \quad \text{for all } t \in \mathbb{R}, \quad (26)$$

where ψ_X represents the characteristic function of the random variable X . Sub-independence is a property less stringent than independence, and both these properties imply uncorrelatedness. Moreover the assumption of sub-independence can replace that of independence in several limit theorems (Hamedani, 2013; Schennach, 2019).

Corollary 3 (Test for sub-independence). *Under the assumptions of Theorem 1 and the null hypothesis of sub-independence (26), such that $\mathbf{x}_1 * \dots * \mathbf{x}_k = \mathbf{z}$, and given $\hat{\mathbf{z}}_m$ the MLE for the PMV $\mathbf{z} \in \Delta_{Int}^s$ such that $(\hat{\mathbf{z}}_m)_u = \frac{1}{m} \sum_{j=1}^m \mathbb{1}_{\{\sum_{i=1}^k X_{ij}=u\}}$ for $u = 0, \dots, s$, it holds that*

$$\mathbf{S}_m = \sqrt{m} (\hat{\mathbf{x}}_{1m} * \dots * \hat{\mathbf{x}}_{km} - \hat{\mathbf{z}}) \underset{m \rightarrow \infty}{\sim} \mathcal{N}(\mathbf{\Upsilon}), \quad (27)$$

where $\mathbf{\Upsilon} = \Sigma(\mathbf{z}) - \mathbf{\Psi}$ and $\text{rk}(\mathbf{\Upsilon}) = s$.

In particular, given $\hat{\mathbf{\Upsilon}}_m = \Sigma(\hat{\mathbf{z}}_m) - \hat{\mathbf{\Psi}}_m$, then

$$\mathbf{S}'_m (\hat{\mathbf{\Upsilon}}_m^s)^+ \mathbf{S}_m \underset{m \rightarrow \infty}{\sim} \chi^2(s). \quad (28)$$

As the independence of random variables implies their sub-independence, Corollary 3 also provides a routine to test for complete independence in k -way contingency tables (Andersen, 1974; Bishop et al., 2007) (one dimension per random variable X_i , $i = 1, \dots, k$), as an alternative to the power divergence family of statistics (Cressie and Read, 1984). In this instance, the null hypothesis of sub-independence, weaker than complete independence, leads to a reduction of free parameters to be estimated, from $\prod_{i=1}^k r_i$ (in the model from power divergence statistics) to $\sum_{i=1}^k r_i$ (in the framework of the convolution statistics), which is desirable when k is large.

4. POWER COMPARISON

We evaluate the performances of the convolution test in terms of type I error and power (1 minus type II error), which are the proportion of rejections with significance level $\alpha = 0.05$ under the null and alternative hypothesis respectively, setting Pearson's χ^2 test as the benchmark. To do so, we simulate the smallest parametrized model that enables the investigation of how samples size, degrees of freedom reduction, and observables distribution

affect the convolution test, using different parameters choices. It also allows the transition from the null to alternative hypotheses by modulating a single parameter.

We consider $k = 2$, $h = 1$ and X_1, X_2 are two Bernoulli random variables with parameters $p, q \in (0, 1)$ so that $\mathbf{x}_1 = (1 - p, p)$, $\mathbf{x}_2 = (1 - q, q)$ and $\mathbf{x}_1, \mathbf{x}_2 \in \Delta_{\text{Int}}^1$. Then we define, for $\rho \in [0, 1]$,

$$\mathbf{z}(\rho) = (1 - \rho)\mathbf{x}_1 * \mathbf{x}_2 + \rho(1 - a, 0, a), \quad (29)$$

where $a = pq + \sqrt{pq(1-p)(1-q)}$ is defined so that $\mathbf{z}(\rho) = (z(\rho)_0, z(\rho)_1, z(\rho)_2) \in \Delta_{\text{Int}}^2$ is the PMV for the distribution of $Z_1 + Z_2$ where Z_1 and Z_2 are two Bernoulli random variables with parameter p and q , respectively, and ρ is their correlation. The null hypothesis for the goodness-of-fit (GF) test is $\mathbf{H}_0 : X_1 + X_2 \sim \mathbf{z}(0)$ and we set a family of alternative hypotheses parametrized over $\rho \in (0, 1]$ as $\mathbf{H}_1^\rho : X_1 + X_2 \approx \mathbf{z}(\rho)$. Similarly, the null and the alternative hypotheses of the test for equality in distribution (ED) are defined as $\mathbf{H}_0 : X_1 + X_2 \sim Y_1$ with $Y_1 \sim \mathbf{z}(0)$ and $\mathbf{H}_1^\rho : X_1 + X_2 \approx Y_1$ with $Y_1 \sim \mathbf{z}(\rho)$, $\rho \in (0, 1]$.

To facilitate the following discussion, we set the sample sizes n_1, n_2 and n_3 for X_1, X_2 and Y_1 , respectively, so that $n_1, n_2 \leq n_3$ and $m = \min(n_1, n_2, n_3) = \min(n_1, n_2)$. Since we are interested in the comparison between the convolution and Pearson's χ^2 statistics, we need to calculate the latter even in the case of unequal sample size, i.e. when $n_1 \neq n_2$. Thus, we define

$$P_m^{\text{GF}} = \sum_{j=0}^2 \frac{(\sum_{i=1}^m \mathbb{1}_{\{X_{1i}+X_{2i}=j\}} - mz(\rho)_j)^2}{mz(\rho)_j} \quad \text{and}$$

$$P_m^{\text{ED}} = \sum_{j=0}^2 \left(\frac{(\frac{n_3}{m+n_3} \sum_{i=1}^m \mathbb{1}_{\{X_{1i}+X_{2i}=j\}} - \frac{m}{m+n_3} \sum_{i=1}^{n_3} \mathbb{1}_{\{Y_i=j\}})^2}{\frac{m}{m+n_3} (\sum_{i=1}^m \mathbb{1}_{\{X_{1i}+X_{2i}=j\}} + \sum_{i=1}^{n_3} \mathbb{1}_{\{Y_i=j\}})} + \frac{(\frac{m}{m+n_3} \sum_{i=1}^{n_3} \mathbb{1}_{\{Y_i=j\}} - \frac{n_3}{m+n_3} \sum_{i=1}^m \mathbb{1}_{\{X_{1i}+X_{2i}=j\}})^2}{\frac{n_3}{m+n_3} (\sum_{i=1}^m \mathbb{1}_{\{X_{1i}+X_{2i}=j\}} + \sum_{i=1}^{n_3} \mathbb{1}_{\{Y_i=j\}})} \right)$$

for Pearson's goodness-of-fit and equality in distribution testing statistic, respectively. In particular, $n_1 + n_2 - 2m$ observations will not be used in the computation of P_m^{GF} and P_m^{ED} .

We define the convolution statistic with fixed rank $r = 1, 2$ from the notation in Proposition 2 and Proposition 4 as $\mathbf{V}'_m(\hat{\Psi}_m^r)^+ \mathbf{V}_m$ and $\mathbf{W}'_m((\hat{\Psi}_m + \hat{\Xi}_m)^r)^+ \mathbf{W}_m$. In the case where the n_1 and n_2 observations from the random variables X_1 and X_2 , respectively, are all equal,

the sample covariance matrix is null, i.e. $\hat{\Psi}_m = \mathbf{0}$, and $(\hat{\Psi}_m^r)^+$ is not well defined. In this scenario, Pearson's P_m^{GF} can still be calculated. As we aim to compare the power gain over Pearson's procedures, we calculate the convolution statistics for goodness-of-fit test as $\mathbf{V}_m'(\hat{\Psi}_m^r)^+\mathbf{V}_m$, where well defined, otherwise we set it to P_m^{GF} . With the same reasoning for the equality in distribution case, for the following simulations we define the convolution statistic as

$$\begin{aligned} C_{rm}^{\text{GF}} &= \mathbf{V}_m'(\hat{\Psi}_m^r)^+\mathbf{V}_m(1 - \mathbb{1}_{\{\hat{\Psi}_m=\mathbf{0}\}}) + P_m^{\text{GF}}\mathbb{1}_{\{\hat{\Psi}_m=\mathbf{0}\}} \quad \text{and} \\ C_{rm}^{\text{ED}} &= \mathbf{W}_m'((\hat{\Psi}_m + \hat{\Xi}_m)^r)^+\mathbf{W}_m(1 - \mathbb{1}_{\{\hat{\Psi}_m+\hat{\Xi}_m=\mathbf{0}\}}) + P_m^{\text{ED}}\mathbb{1}_{\{\hat{\Psi}_m+\hat{\Xi}_m=\mathbf{0}\}}, \end{aligned}$$

for goodness-of-fit and equality in distribution tests, respectively. Note that $\lim_{m \rightarrow \infty} C_{rm}^{\text{GF}} \sim \chi^2(r)$, since $\hat{\Psi}_m = \mathbf{0}$ if and only if $\hat{\mathbf{x}}_{1n_1}, \hat{\mathbf{x}}_{2n_2} \in \{(1, 0), (0, 1)\}$, but, for $j = 1, 2$, $\lim_{m \rightarrow \infty} \mathbb{P}(\hat{\mathbf{x}}_{jn_j} \in \{(1, 0), (0, 1)\}) = 0$. Analogously, also $\lim_{m \rightarrow \infty} C_{rm}^{\text{ED}} \sim \chi^2(r)$ holds true.

Moreover, in order not to confound the comparative analysis, we do not reduce the limiting χ^2 degrees of freedom in the case where a positive eigenvalue of $\hat{\Psi}_m$, or $\hat{\Psi}_m + \hat{\Xi}_m$, is set to 0 for being smaller than $10^{-\epsilon}$ (here $\epsilon = 15$ is the machine precision from Python's floating point number in Numpy 1.13.1).

Finally, to assess whether deviations from the limit of the convolution statistic are due to the estimate of the covariance matrix pseudo-inverse, we introduce

$$Z_{rm}^{\text{GF}} = \mathbf{V}_m' \Psi^+ \mathbf{V}_m \quad \text{and} \quad Z_{rm}^{\text{ED}} = \mathbf{W}_m' (\Psi + \Xi)^+ \mathbf{W}_m$$

for $r = 1, 2$, which are the convolution statistics calculated with the true covariance matrices.

For all these statistics, we evaluate the proportion of hypothesis rejection for the significance level $\alpha = 0.05$ by Monte Carlo approximation over $L = 100,000$ independent instances of the data. That is, the proportion of rejections for U (assumed to be one between P_m^{GF} , C_{rm}^{GF} , Z_{rm}^{GF} , P_m^{ED} , C_{rm}^{ED} , Z_{rm}^{ED}) is calculated as

$$\frac{1}{L} \sum_{l=1}^L \mathbb{1}_{\{S_r(U_l) < \alpha\}}, \quad (30)$$

given $\{U_1, \dots, U_L\}$ are independent and identically distributed from U , and $S_r(t) = \mathbb{P}(\chi^2(r) \geq t)$ equal to the survival function for a χ^2 distribution with r degrees of freedom.

In Fig. 1, we report the statistical power under \mathbf{H}_0 and a range of alternative hypotheses \mathbf{H}_1^p . We implement these comparisons for small and large samples with respect to $m = \min(n_1, n_2, n_3)$, and with equal and unequal sizes.

To comply with the rule-of-thumb recommendation for Pearson's χ^2 statistic application (Cressie and Read, 1984), the requirement for the expected frequencies $m(\mathbf{x}_1 * \mathbf{x}_2)_u \geq 1$ for the categories $u = 0, 1, 2$, must be met when m observations are sampled from the distribution of $X_1 + X_2$. Thus, we select two cases for the parameters (p, q) so that, under small samples $m = 10$, all three constraints from the rule-of-thumb are satisfied when $(p, q) = (0.3, 0.8)$, while only one, i.e. $m(\mathbf{x}_1 * \mathbf{x}_2)_1 \geq 1$, holds for $(p, q) = (0.1, 0.9)$. These parameters were selected to check whether the convolution statistic offers a better alternative over Pearson's χ^2 , under cases favorable to P_m^{GF} and P_m^{ED} , when $(p, q) = (0.3, 0.8)$ and sample sizes are equal, or unfavorable, when $n_1 + n_2 - 2m > 0$ observations are excluded and the rule-of-thumb is violated.

For the small sample cases when $(p, q) = (0.3, 0.8)$, C_{2m}^{GF} provides better power over P_m^{GF} , and the latter over C_{1m}^{GF} , but C_{2m}^{GF} shows a proportion of rejections that is above α under \mathbf{H}_0 (Fig. 1a, top left and middle left panels). When $(p, q) = (0.1, 0.9)$, C_{1m}^{GF} and C_{2m}^{GF} have similar behavior which outperforms P_m^{GF} (Fig. 1a, top right and middle right panels). Equivalent conclusions are inferred for the equality in distribution testing statistic counterparts (Fig. 1b, top and middle panels). For large samples, under \mathbf{H}_0 , the proportion of rejections becomes closer to α for C_{2m}^{GF} (Fig. 1a, bottom panels) and it coincides for C_{2m}^{ED} (Fig. 1b, bottom panels); in terms of power, convolution statistics outperform Pearson's χ^2 , with the exception of C_{1m}^{GF} and C_{1m}^{ED} when $(p, q) = (0.3, 0.8)$ (Fig. 1a,b, bottom left panels).

In Fig. 2 we illustrate the convergence of the rejection rate for m large to the significance level α under \mathbf{H}_0 and the power convergence under the alternative hypothesis $\mathbf{H}_1^{0.25}$. When testing for goodness-of-fit, C_{2m}^{GF} shows the highest rejection proportion which leads to good power under the alternative hypothesis (Fig. 2b), but a slow convergence to α under \mathbf{H}_0 , reaching peaks of rejection up to 2α (Fig. 2a). C_{1m}^{GF} and P_m^{GF} , instead, have similar behaviors better than C_{2m}^{GF} , with P_m^{GF} outperforming C_{1m}^{GF} in its most favorable case ($(p, q) = (0.3, 0.8)$),

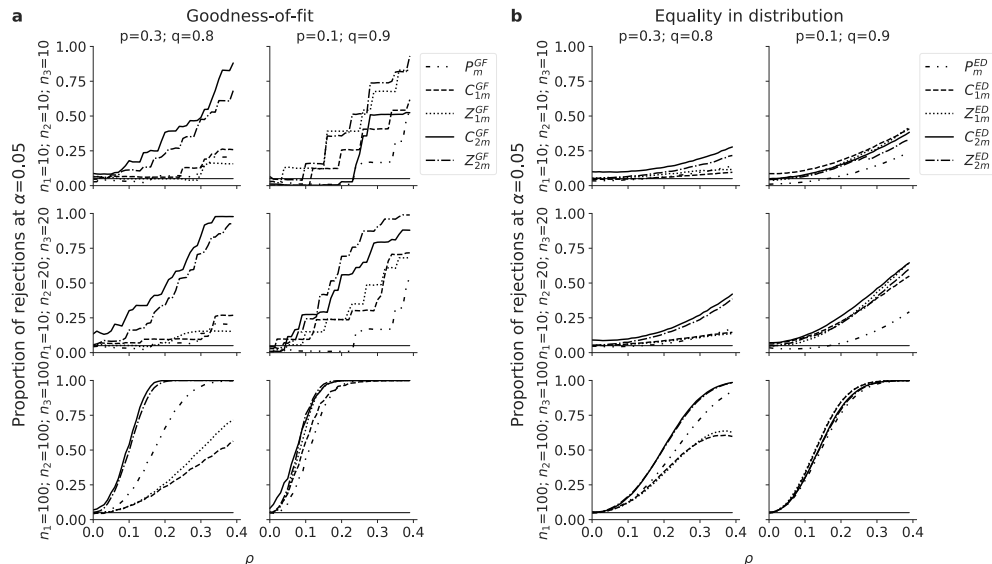


Figure 1: Power comparison of convolution statistics against those calculated with the true covariance matrix and Pearson's χ^2 statistics, under hypotheses \mathbf{H}_0 (for $\rho = 0$) and \mathbf{H}_1^ρ (for $\rho > 0$) when testing for goodness-of-fit (a) and equality in distribution (b). A full horizontal line depicts the nominal rejection level $\alpha = 0.05$.

bottom left panels from Fig. 2a,b) while the converse holds in the other cases. When testing for equality in distribution, results for C_{1m}^{ED} and P_m^{ED} are similar (Fig. 2c,d), while C_{2m}^{ED} presents a much faster convergence under \mathbf{H}_0 than its goodness-of-fit counterpart, as it approaches α already at $m = 100$ (Fig. 2c).

Together, Figs. 1 and 2 suggest a tendency of the convolution statistics to attain a more anti-conservative behavior (type I error higher than α), while for Person's χ^2 statistic this is more conservative (type I error lower than α).

Lastly, in Fig. 3 we analyze the convolution statistic in the case of covariance matrix that is near a reduced rank form, when its smallest positive eigenvalue approaches zero. Equivalently, this situation occurs if the roots of the PMVs for X_1 and X_2 are close, i.e. $(p-1)/p - (q-1)/q$ becomes null. To this end, we fix $q \in (0, 1)$, so that $\text{rk}(\Psi) = 1$ if $p = q$ (but still $\text{rk}(\Psi + \Xi) = 2$), and we compare the proportion of rejections when the roots are well separated or close each other, under both \mathbf{H}_0 and $\mathbf{H}_1^{0.25}$.

For small samples, C_{1m}^{GF} and P_m^{GF} present similar performance under \mathbf{H}_0 (Fig. 3a, left

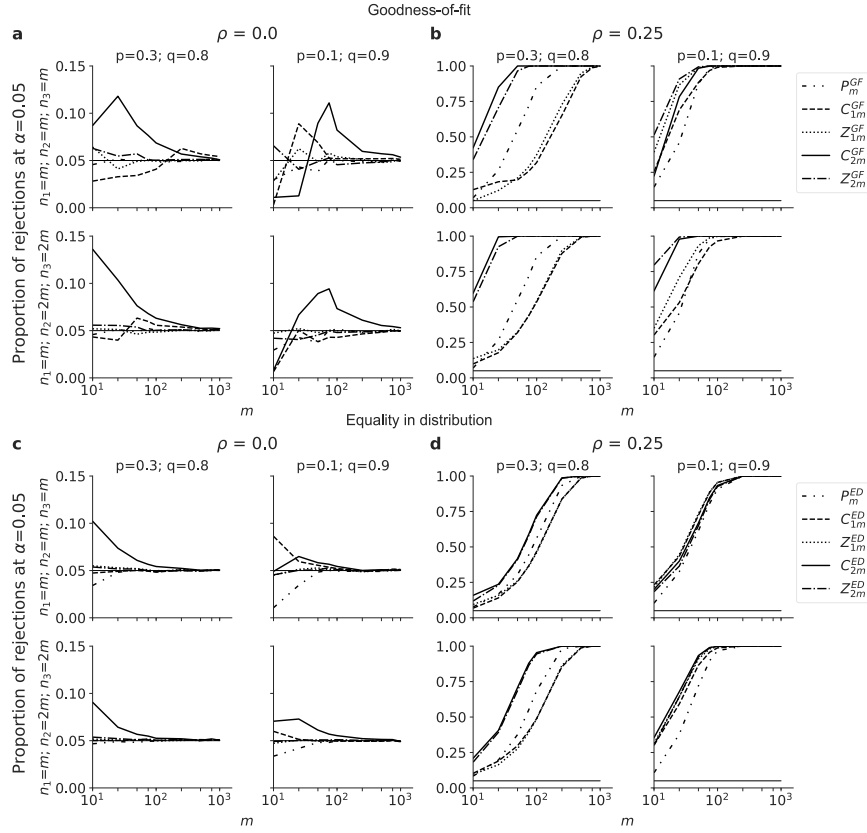


Figure 2: Comparison of convergence speed of convolution statistics against those calculated with the true covariance matrix and Pearson's χ^2 statistics, as samples size becomes large with m . A full horizontal line depicts the nominal rejection level $\alpha = 0.05$.

panels), while P_m^{ED} and C_{1m}^{ED} are, respectively, conservative and anti-conservative (Fig. 3b, left panels). The power under $\mathbf{H}_1^{0.25}$ favors the use of C_{1m}^{GF} over P_m^{GF} for small samples, but, for large samples, P_m^{GF} outperforms C_{1m}^{GF} when p is near q (Fig. 3a, right panels). As the spread between p and q widens, eventually C_{1m}^{GF} performs better than P_m^{GF} . This commentary is also true for the equality in distribution statistic counterparts (Fig. 3b, right panels).

As predicted from Figs. 1 and 2 analyses, the behavior of C_{2m}^{GF} shows higher proportion of rejections. Furthermore, due to $\text{rk}(\Psi) = 1$, as p approaches q , C_{2m}^{GF} undertakes a dramatic deviation from the nominal rejection proportion α , indicative of the consistency failure of $(\hat{\Psi}_m^2)^+$ in the estimation of Ψ^+ when $p = q$ (Fig. 3a, bottom left panels). For large samples, we see that C_{2m}^{ED} is more powerful than C_{1m}^{ED} , since $\text{rk}(\Psi + \Xi) = 2$, and also than P_m^{ED}

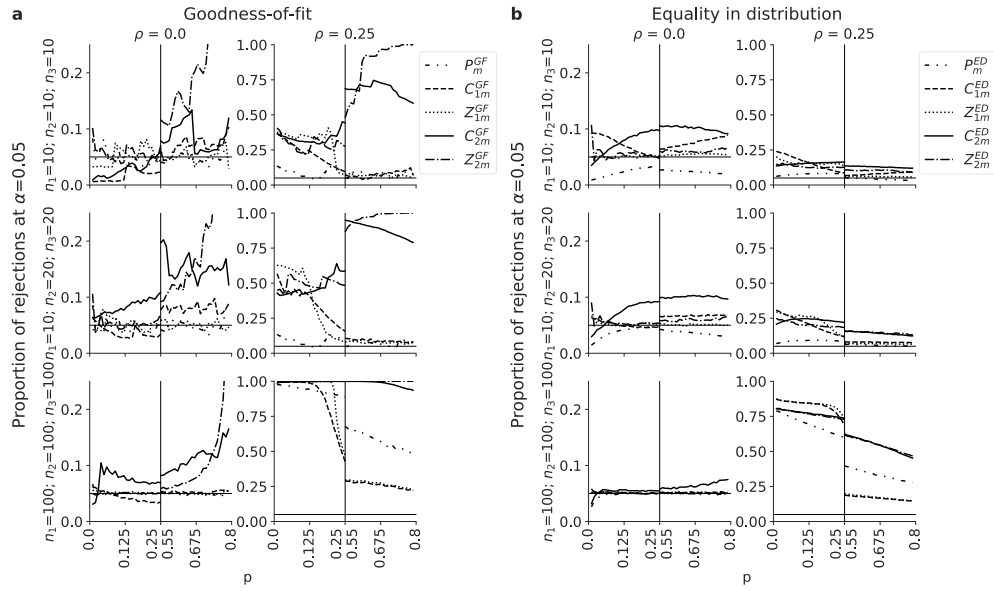


Figure 3: Proportion of rejections when the smallest positive eigenvalue of Ψ tends to zero (when p approaches $q = 0.8$, under the model with $X_1 \sim (1 - p, p)$, $X_2 \sim (1 - q, q)$) and when the rule-of-thumb for Pearson's χ^2 statistic is violated (when p approaches 0). Convolution statistics are compared against those calculated with the true covariance matrix and Pearson's χ^2 statistics. A full horizontal line depicts the nominal rejection level $\alpha = 0.05$.

when the roots are close (Fig. 3b, bottom panels). The discrepancy between C_{2m}^{GF} and C_{2m}^{ED} behaviors highlights the merit for the rank analysis of Section 3 and shows the danger of setting the same degrees of freedom for the convolution as for Pearson's χ^2 statistics without careful consideration.

To summarize the evidence above, we propose the following as general usage guidelines: when the number of samples is large or unbalanced, and PMV's roots are distinct, the convolution statistic with maximum rank possible provides the best power; as PMV's roots become closer or low sample size hinders the estimation of the covariance matrix pseudo-inverse, the convolution statistic with reduced rank should be employed to ensure a good compromise between type I and type II errors control; Pearson's χ^2 is still recommended over the convolution statistics, but only for small sample data and only if the rule-of-thumb is not violated, in case the type I error must be conservatively controlled and type II error is considered

of secondary importance. Ultimately, it is advisable to run comparative simulations over specific scenarios of interest that largely differ from the one considered above.

5. DISCUSSION

In this work, we have shown how to test hypotheses about the sum of discrete random variables using the operation of discrete convolution, from possibly unbalanced datasets and without restrictions to specific distribution classes. Asymptotic properties of the convolution were combined with the generalized Wald’s method to solve the testing problems of: goodness-of-fit, equality in distribution and sub-independence. These results have been achieved studying the rank of the limiting covariance matrix, from the convolution of PMVs. Of note, such rank was determined as a function of the number of roots shared between the probability generating functions of the random variables considered. Benchmarking simulations were conducted on a simple parametric model covering several situations of interest and, in most of these, the convolution statistic performed better than Pearson’s.

Perspective studies are still required to complete the description of the convolution statistics, including the cases where PMVs’ support is not connected, i.e. in $\Delta^r \setminus \Delta_{\text{Int}}^r$ for $r > 0$. Finally, improved strategies for the estimation of the pseudo-inverse matrix, that are ill-conditioned, should be considered to ameliorate type I and II errors.

5. ACKNOWLEDGMENTS

We thank Julia Marchingo, Andrey Kan, Susanne Heinzl and Phil Hodgkin at the Walter and Eliza Hall Institute of Medical Research for collaborating with us on the scientific problem that motivated the development of the convolution statistic and associated tests. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/20072013) under grant agreement 317040 (QuanTI) and by Science Foundation Ireland Grant 12IP1263. On behalf of all authors, the corresponding author states that there is no conflict of interest.

6. CODE AVAILABILITY

Python 3 code for simulation and testing is publicly available at “<https://github.com/>

References

- Andersen, A. H. (1974). Multidimensional contingency tables. *Scandinavian Journal of Statistics*, 1(3):115–127.
- Andrews, D. W. K. (1987). Asymptotic results for generalized Wald tests. *Econ Theory*, 3(3):348–358.
- Andrews, D. W. K. (1988). Chi-square diagnostic tests for econometric models: Introduction and applications. *J Econom*, 37(1):135–156.
- Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007). *Discrete Multivariate Analysis: Theory and Practice*. Springer, New York, NY, USA, 1 edition.
- Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *J R Stat Soc Series B Stat Methodol*, 46(3):440–464.
- Drost, F. C. (1989). Generalized chi-square goodness-of-fit tests for location-scale models when the number of classes tends to infinity. *Ann Stat*, 17(3):1285–1300.
- Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.
- Editorial (2017). Rationalizing combination therapies. *Nat Med*, 23:1113.
- Hadi, A. S. and Wells, M. T. (1990). A note on generalized Wald’s method. *Metrika*, 37(1):309–315.
- Hagen, R., Roch, S., and Silbermann, B. (2000). *C* - Algebras and Numerical Analysis*. CRC Press, New York, NY, USA.

- Hamedani, G. G. (2013). Sub-independence: An expository perspective. *Communications in Statistics – Theory and Methods*, 42(3):3615–3638.
- Horn, R. A. and Johnson, C. R. (1986). *Matrix Analysis*. Cambridge University Press, New York, NY, USA, 1 edition.
- Marchingo, J. M., Kan, A., Sutherland, R. M., Duffy, K. R., Wellard, C. J., Belz, G. T., Lew, A. M., Dowling, M. R., Heinzl, S., and Hodgkin, P. D. (2014). Antigen affinity, costimulation, and cytokine inputs sum linearly to amplify T cell expansion. *Science*, 346:1123–1127.
- Marchingo, J. M., Prevedello, G., Kan, A., Heinzl, S., Hodgkin, P. D., and Duffy, K. R. (2016). T-cell stimuli independently sum to regulate an inherited clonal division fate. *Nat Commun*, 7:13540.
- Markovsky, I. (2012). *Low Rank Approximation: Algorithms, Implementation, Applications*. Springer, Cambridge, UK, 1 edition.
- Mihalko, D. P. and Moore, D. S. (1980). Chi-square tests of fit for type II censored data. *Ann Stat*, 8(3):625–644.
- Moore, D. S. (1977). Generalized inverses, Walds method, and the construction of chi-squared tests of fit. *J Am Stat Assoc*, 72(357):131–137.
- Moore, D. S. (1982). The effect of dependence on chi squared tests of fit. *Ann Stat*, 10(4):1163–1171.
- Moore, D. S. and Spruill, M. C. (1975). Unified large-sample theory of general chi-squared statistics for tests of fit. *Ann Stat*, 3(3):599–616.
- Nashed, M. Z. (1976). Perturbations and approximations for generalized inverses and linear operator equations. In Nashed, M. Z., editor, *Generalized Inverses and Applications*, pages 325–396. Academic Press, New York, NY, USA, 1 edition.

- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.
- Schennach, S. M. (2019). Convolution without independence. *Journal of Econometrics*, 211(1):308–318.
- Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York, NY, USA, 1 edition.
- Tyler, D. E. (1981). Asymptotic inference for eigenvectors. *Ann Stat*, 9(4):725–736.
- Voinov, V., Roza, A., and Pya, N. (2008). Recent achievements in modified chi-squared goodness-of-fit testing. In Vonta, F., Nikulin, M., Limnios, N., and Huber-Carol, C., editors, *Statistical Models and Methods for Biomedical and Technical Systems*.
- Vuong, Q. H. (1987). Generalized inverses and asymptotic properties of Wald tests. *Econ Lett*, 24(4):343–347.
- Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans Am Math Soc*, 54(3):426–482.
- Wilson, J. R. and Koehler, K. J. (1991). Hierarchical models for cross-classified overdispersed multinomial data. *J Bus Econ Stat*, 9(1):103–110.
- Wolchok, J. D., Kluger, H., Callahan, M. K., Postow, M. A., Rizvi, N. A., Lesokhin, A. M., Segal, N. H., Ariyan, C. E., Gordon, R., Reed, K., Burke, M. M., Caldwell, A., Kronenberg, S. A., Agunwamba, B. U., Zhang, X., Lowy, I., Inzunza, H. D., Feely, W., Horak, C. E., Hong, Q., Korman, A. J., Wigginton, J. M., Gupta, A., and Sznol, M. (2013). Nivolumab plus Ipilimumab in advanced melanoma. *N Engl J Med*, 369(2):122–133.
- Zhang, B. (1999). A chi-squared goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 86(3):531–539.

APPENDIX

Proof of Lemma 1. Without loss of generality, it is possible to shift from the lattice $\Lambda(\zeta)$ to the set of integers \mathbb{Z} through the natural isomorphism $\phi: \Lambda(\zeta) \rightarrow \mathbb{Z}$, $\phi(\zeta u) = u$ for every $u \in \mathbb{Z}$. Using this function, we define $A'_i = a_i \phi(A_i)$ and $B'_j = b_j \phi(B_j)$ for $i = 1, \dots, k$, $j = 1, \dots, h$, so to account for the multiplicative constants $a_1, \dots, a_k, b_1, \dots, b_h$ in the variables $A'_1, \dots, A'_k, B'_1, \dots, B'_h$ mapping Ω to $\mathbb{N} \cup \{0\}$, and reduce (2) into

$$\mathbf{H}_0: \phi(a_0) + \sum_{i=1}^k A'_i \sim \phi(b_0) + \sum_{i=1}^h B'_i. \quad (31)$$

Given $\tau_i = \min\{j: \mathbb{P}(A'_i = j) > 0\}$ for $i = 1, \dots, k$ and $\tau_{k+i} = \min\{j: \mathbb{P}(B'_i = j) > 0\}$ for $i = 1, \dots, h$, that are well defined since the variables $A'_1, \dots, A'_k, B'_1, \dots, B'_h$ are assumed finite, we rewrite (31) as

$$\mathbf{H}_0: a_0 + \sum_{i=1}^k \tau_i + \sum_{i=1}^k (A'_i - \tau_i) \sim b_0 + \sum_{i=1}^h \tau_{k+i} + \sum_{i=1}^h (B'_i - \tau_{k+i}). \quad (32)$$

For the null hypothesis (32) to be true, $a_0 + \sum_{i=1}^k \tau_i = b_0 + \sum_{i=1}^h \tau_{k+i}$ must hold. Otherwise, for example, if $a_0 + \sum_{i=1}^k \tau_i < b_0 + \sum_{i=1}^h \tau_{k+i}$, by definition of $\tau_1, \dots, \tau_{k+h}$ we would have

$$\begin{aligned} 0 &= \mathbb{P}\left(b_0 + \sum_{i=1}^h \tau_{k+i} + \sum_{i=1}^h (B'_i - \tau_{k+i}) = a_0 + \sum_{i=1}^k \tau_i\right) \\ &= \mathbb{P}\left(a_0 + \sum_{i=1}^k \tau_i + \sum_{i=1}^k (A'_i - \tau_i) = a_0 + \sum_{i=1}^k \tau_i\right) \geq \prod_{i=1}^k \mathbb{P}(A'_i = \tau_i) > 0, \end{aligned}$$

that is impossible. Therefore (32) is equivalent to

$$\mathbf{H}_0: \sum_{i=1}^k (A'_i - \tau_i) \sim \sum_{i=1}^h (B'_i - \tau_{k+i}),$$

which, in turn, can be reduced to the form (3) by defining $X_i = A'_i - \tau_i$ for $i = 1, \dots, k$ and $Y_i = B'_i - \tau_{k+i}$ for $i = 1, \dots, h$ thus accounting for the subtraction the constants in the distribution of X_i and Y_i . As a consequence, the support of X_i is $\{0, \dots, r_i\}$ for some positive integer $r_i \in \mathbb{N}$ and

$$\mathbb{P}(X_i = 0) > 0 \quad (33)$$

for every $i = 1, \dots, k$. The same applies to Y_1, \dots, Y_h . \square

Proof of Proposition 1. Given $a_{ij} \in \{0, \dots, r_i\}$, the independent sample from X_{ij} for $i = 1, \dots, k$ and $j = 1, \dots, n_i$, the MLE of $\mathbf{x}_1 * \dots * \mathbf{x}_k$ is the element $\boldsymbol{\theta} \in \Delta^s$ that maximizes

$$\begin{aligned} \mathbb{P}(X_{11} = a_{11}, \dots, X_{1n_1} = a_{1n_1}, \dots, X_{k1} = a_{k1}, \dots, X_{kn_k} = a_{kn_k} | \boldsymbol{\theta}) \\ = \prod_{i=1}^k \mathbb{P}(X_{i1} = a_{i1}, \dots, X_{in_i} = a_{in_i} | \boldsymbol{\theta}) \end{aligned}$$

On the right hand side, for fixed $i \in \{1, \dots, k\}$, $\mathbb{P}(X_{i1} = a_{i1}, \dots, X_{in_i} = a_{in_i} | \boldsymbol{\theta})$ achieves maximum value for any $\boldsymbol{\theta} = \boldsymbol{\theta}_1 * \dots * \boldsymbol{\theta}_k$ such that $\boldsymbol{\theta}_i = \hat{\mathbf{x}}_{in_i}$, with $\boldsymbol{\theta}_j \in \Delta^{r_j}$ for any j . In particular $\boldsymbol{\theta} = \hat{\mathbf{x}}_{1n_1} * \dots * \hat{\mathbf{x}}_{kn_k}$ maximizes all factors, hence the whole product. \square

Proof of Lemma 2. We first prove (18) by induction on k . For $k = 2$ we have to show that

$$\text{Ker}(\mathbf{T}(\mathbf{x}_{(1)})') \cap \text{Ker}(\mathbf{T}(\mathbf{x}_{(2)})') = \text{Ker} \left(\begin{bmatrix} \mathbf{T}(\mathbf{x}_{(1)})' \\ \mathbf{T}(\mathbf{x}_{(2)})' \end{bmatrix} \right) = \text{Ker}(\mathbf{T}(\mathbf{g}_2)'), \quad (34)$$

where $\mathbf{g}_2 = \text{gcd}(\mathbf{x}_{(1)}, \mathbf{x}_{(2)}) = \text{gcd}(\mathbf{x}_2, \mathbf{x}_1) \in \mathbb{R}^{r_{g_2}+1}$ and $r_{g_2} \geq 0$. By definition of \mathbf{g}_2 , there exist two coprime vectors $\mathbf{z}_1 \in \mathbb{R}^{r_1-r_{g_2}+1}$, $\mathbf{z}_2 \in \mathbb{R}^{r_2-r_{g_2}+1}$ such that $\mathbf{x}_{(1)} = \mathbf{z}_1 * \mathbf{g}_2$ and $\mathbf{x}_{(2)} = \mathbf{z}_2 * \mathbf{g}_2$ so that, by composition of discrete convolution, we can write

$$\begin{bmatrix} \mathbf{T}(\mathbf{x}_{(1)})' \\ \mathbf{T}(\mathbf{x}_{(2)})' \end{bmatrix} = \begin{bmatrix} \mathbf{T}(\mathbf{z}_1)' \\ \mathbf{T}(\mathbf{z}_2)' \end{bmatrix} \mathbf{T}(\mathbf{g}_2)',$$

where $\mathbf{T}(\mathbf{g}_2)' \in \mathbb{R}^{r_1+r_2-r_{g_2}+1} \times \mathbb{R}^{r_1+r_2+1}$, $\mathbf{T}(\mathbf{z}_1)' \in \mathbb{R}^{r_2+1} \times \mathbb{R}^{r_1+r_2-r_{g_2}+1}$ and $\mathbf{T}(\mathbf{z}_2)' \in \mathbb{R}^{r_1+1} \times \mathbb{R}^{r_1+r_2-r_{g_2}+1}$. As the number of columns of $\begin{bmatrix} \mathbf{T}(\mathbf{z}_1) & \mathbf{T}(\mathbf{z}_2) \end{bmatrix}'$ is lesser than the number of rows, since $r_1 + r_2 - r_{g_2} + 1 \leq r_1 + r_2 + 2 \Leftrightarrow r_{g_2} + 1 \geq 0$, to prove (34), it suffices to show that $\text{rk} \left(\begin{bmatrix} \mathbf{T}(\mathbf{z}_1) & \mathbf{T}(\mathbf{z}_2) \end{bmatrix}' \right)$ is of full rank $r_1 + r_2 - r_{g_2} + 1$. By rank-nullity theorem, this is the case if and only if $\text{nul} \left(\begin{bmatrix} \mathbf{T}(\mathbf{z}_1) & \mathbf{T}(\mathbf{z}_2) \end{bmatrix} \right) = r_{g_2} + 1$. The latter is true by coprimeness between \mathbf{z}_1 and \mathbf{z}_2 and by matrix dimensionality, since the solutions (\mathbf{a}, \mathbf{b}) with $\mathbf{a} \in \mathbb{R}^{r_1+1}$, $\mathbf{b} \in \mathbb{R}^{r_2+1}$, to the homogeneous system of equations $\mathbf{T}(\mathbf{z}_1)\mathbf{a} + \mathbf{T}(\mathbf{z}_2)\mathbf{b} = \mathbf{z}_1 * \mathbf{a} + \mathbf{z}_2 * \mathbf{b} = \mathbf{0}$, are characterized by $\mathbf{a} = \mathbf{z}_2 * \mathbf{u}$, $\mathbf{b} = -\mathbf{z}_1 * \mathbf{u}$ with $\mathbf{u} \in \mathbb{R}^{r_{g_2}+1}$ vector of free parameters. Assuming (18) for k , we now prove the case $k + 1$ to conclude. By inductive step and associative property of

convolution, we can write

$$\bigcap_{i=1}^{k+1} \text{Ker}(\mathbf{T}(\mathbf{x}_{(i)}))' = \text{Ker} \left(\begin{bmatrix} \mathbf{T}(\mathbf{g}_k)' \mathbf{T}(\mathbf{x}_{k+1})' \\ \mathbf{T}(\mathbf{x}_{(k+1)})' \end{bmatrix} \right) = \text{Ker} \left(\begin{bmatrix} \mathbf{T}(\mathbf{g}_k * \mathbf{x}_{k+1})' \\ \mathbf{T}(\mathbf{x}_{(k+1)})' \end{bmatrix} \right). \quad (35)$$

Thus, by the definition of $\mathbf{g}_{k+1} = \text{gcd}(\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(k+1)}) \in \mathbb{R}^{r_{g_{k+1}}+1}$, the properties of gcd lead to $\mathbf{g}_{k+1} = \text{gcd}(\mathbf{g}_k * \mathbf{x}_{k+1}, \mathbf{x}_1 * \dots * \mathbf{x}_k) = \mathbf{g}_k \text{gcd}(\mathbf{x}_{k+1}, \mathbf{u}_{k+1})$, where $\mathbf{u}_{k+1} \in \mathbb{R}^{\sum_{i=1}^k r_i - r_{g_k} + 1}$ is such that $\mathbf{u}_{k+1} * \mathbf{g}_{k+1} = \mathbf{x}_1 * \dots * \mathbf{x}_k$. In particular, we deduce

$$r_{g_{k+1}} \geq r_{g_k}. \quad (36)$$

As in the previous step, we introduce $\mathbf{z}_1 \in \mathbb{R}^{r_{g_k} + r_{k+1} - r_{g_{k+1}} + 1}$ and $\mathbf{z}_2 \in \mathbb{R}^{\sum_{i=1}^k r_i - r_{g_{k+1}} + 1}$ coprime vectors such that $\mathbf{z}_1 * \mathbf{g}_{k+1} = \mathbf{g}_k * \mathbf{x}_{k+1}$ and $\mathbf{z}_2 * \mathbf{g}_{k+1} = \mathbf{x}_1 * \dots * \mathbf{x}_k$. Thus, resuming (35),

$$\bigcap_{i=1}^{k+1} \text{Ker}(\mathbf{T}(\mathbf{x}_{(i)}))' = \text{Ker} \left(\begin{bmatrix} \mathbf{T}(\mathbf{z}_1)' \\ \mathbf{T}(\mathbf{z}_2)' \end{bmatrix} \mathbf{T}(\mathbf{g}_{k+1})' \right) = \text{Ker}(\mathbf{T}(\mathbf{g}_{k+1}))'$$

where last equality is analogous as for the case $k = 2$, given that the number of columns of $\begin{bmatrix} \mathbf{T}(\mathbf{z}_1) & \mathbf{T}(\mathbf{z}_2) \end{bmatrix}' \in \mathbb{R}^{\sum_{i=1}^{k+1} r_i - r_{g_k} + 2} \times \mathbb{R}^{\sum_{i=1}^{k+1} r_i - r_{g_{k+1}} + 1}$ is lesser than the number of rows, that is

$$\sum_{i=1}^{k+1} r_i - r_{g_{k+1}} + 1 \leq \sum_{i=1}^{k+1} r_i - r_{g_k} + 2 \quad \Leftrightarrow \quad r_{g_{k+1}} + 1 \geq r_{g_k},$$

which holds true by (36). To prove (19), we note that the first equation therein is established by (18), thus only the second equivalence shall be proved. Once more, we define \mathbf{z}_1 and \mathbf{z}_2 such that such that $\begin{bmatrix} \mathbf{T}(\mathbf{z}_1) & \mathbf{T}(\mathbf{z}_2) \end{bmatrix}' \mathbf{T}(\tilde{\mathbf{g}})' = \begin{bmatrix} \mathbf{T}(\mathbf{g}_k) & \mathbf{T}(\bar{\mathbf{g}}_h) \end{bmatrix}'$ with $\mathbf{T}(\mathbf{z}_1)' \in \mathbb{R}^{\sum_{i=1}^k r_i - r_{g_k} + 1} \times \mathbb{R}^{\sum_{i=1}^k r_i - r_{\tilde{\mathbf{g}}} + 1}$, $\mathbf{T}(\mathbf{z}_2)' \in \mathbb{R}^{\sum_{i=1}^k r_i - r_{\bar{\mathbf{g}}_h} + 1} \times \mathbb{R}^{\sum_{i=1}^k r_i - r_{\tilde{\mathbf{g}}} + 1}$. Again, the conclusion is verified by checking that, in the matrix $\begin{bmatrix} \mathbf{T}(\mathbf{z}_1) & \mathbf{T}(\mathbf{z}_2) \end{bmatrix}'$, there are less rows than columns, namely

$$\sum_{i=1}^k r_i + r_{\tilde{\mathbf{g}}} + 1 \geq r_{g_k} + r_{\bar{\mathbf{g}}_h},$$

which holds true since $\mathbf{g}_k * \bar{\mathbf{g}}_h = \text{gcd}(\mathbf{g}_k, \bar{\mathbf{g}}_h) * \text{lcm}(\mathbf{g}_k, \bar{\mathbf{g}}_h) = \tilde{\mathbf{g}} * \text{lcm}(\mathbf{g}_k, \bar{\mathbf{g}}_h)$ and $\text{lcm}(\mathbf{g}_k, \bar{\mathbf{g}}_h)$ is a divisor of $\mathbf{z} = \mathbf{x}_1 * \dots * \mathbf{x}_k = \mathbf{y}_1 * \dots * \mathbf{y}_h$ (as polynomials), so $\text{deg } \text{lcm}(\mathbf{g}_k, \bar{\mathbf{g}}_h) \leq \sum_{i=1}^k r_i$. \square

Proof of Theorem 1. The relations (20) and (21) derive as applications of Lemma 2 to (14) and (15), respectively. Lastly, (22) and (23) follow from (20) and (21), respectively, through rank-nullity properties of linear transformations from a finite-dimensional domain. \square

Proof of Corollary 1. This follows from Theorem 1, as $\text{Ker}(\mathbf{T}(\mathbf{g}_k)') = \text{Ker}(\mathbf{T}(\tilde{\mathbf{g}})') = \{(0, \dots, 0)\} \subseteq \mathbb{R}^{s+1}$ by coprimeness. \square

Proof of Corollary 2. It suffices to show (24), since (25) follows from the same reasoning. Relation (13), Lemma 2 and Theorem 1 imply

$$\begin{aligned} \text{Ker}(\Psi) &= \text{Ker}(\mathbf{T}(\mathbf{g}_k)') \oplus \left(\bigcap_{i=1}^k \{ \mathbf{v} \in \mathbb{R}^{s+1} : \mathbf{T}(\mathbf{g}_k)' \mathbf{v} \in \text{Ker}(\Sigma(\mathbf{x}_i)) \} \right) \\ &\subseteq \langle \mathbf{1}_s \rangle \oplus \text{Ker}(\mathbf{T}(\mathbf{g}_k)') \oplus_{i=1}^k \{ \mathbf{v} \in \mathbb{R}^{s+1} : \mathbf{T}(\mathbf{g}_k)' \mathbf{v} \in E_i \}, \end{aligned}$$

since $\text{Ker}(\Sigma(\mathbf{x}_i)) = \langle \mathbf{1}_{r_i} \rangle \oplus E_i$, given $E_i = \langle \{ \mathbf{e}_l^i : l \in L_i \} \rangle$ with $e_{lu}^i = \delta_{l,u}$ for $u = 0, \dots, r_i$ and $i = 1, \dots, k$. As the dimension of E_i equals the cardinality of L_i , (24) follows from

$$\text{nul}(\Psi) \leq 1 + r_{g_k} + \sum_{i=1}^k |L_i|. \quad \square$$

Proof of Corollary 3. To prove (27) we write the relation

$$\sqrt{m} \begin{pmatrix} \hat{\mathbf{x}}_{1m} - \mathbf{x}_1 \\ \vdots \\ \hat{\mathbf{x}}_{km} - \mathbf{x}_k \\ \hat{\mathbf{z}}_m - \mathbf{z} \end{pmatrix} \underset{m \rightarrow \infty}{\sim} \mathcal{N} \left(\begin{bmatrix} \Sigma(\mathbf{x}_1) & \mathbf{0} & \dots & \mathbf{0} & \mathbf{A}_1 \\ \mathbf{0} & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \Sigma(\mathbf{x}_k) & \mathbf{A}_k \\ \mathbf{A}'_1 & \dots & \dots & \mathbf{A}'_k & \Sigma(\mathbf{z}) \end{bmatrix} \right), \quad (37)$$

where the $\mathbf{0}$ entries follow from the uncorrelatedness implied by sub-independence, and $\mathbf{A}_i = \mathbf{D}(\mathbf{x}_i)(\mathbf{T}(\mathbf{x}_i)' - (\mathbf{z} \dots \mathbf{z})')$, given $\mathbf{D}(\mathbf{x}_i)$ the matrix with \mathbf{x}_i at the diagonal and 0 elsewhere. Then, (27) is derived by applying the delta method to (37), with

$$\Upsilon = \Psi - \sum_{i=1}^k \mathbf{A}'_i \mathbf{T}(\mathbf{x}_i)' - \sum_{i=1}^k \mathbf{T}(\mathbf{x}_i) \mathbf{A}_i + \Sigma(\mathbf{z}) = \Sigma(\mathbf{z}) - \Psi,$$

since

$$\sum_{i=1}^k \mathbf{T}(\mathbf{x}_{(i)}) \mathbf{A}_i = \sum_{i=1}^k \mathbf{T}(\mathbf{x}_{(i)}) (\mathbf{D}(\mathbf{x}_i) - \mathbf{x}_i \mathbf{x}_i') \mathbf{T}(\mathbf{x}_{(i)})' = \sum_{i=1}^k \mathbf{T}(\mathbf{x}_{(i)}) \boldsymbol{\Sigma}(\mathbf{x}_i) \mathbf{T}(\mathbf{x}_{(i)})' = \boldsymbol{\Psi}.$$

To obtain $\text{rk}(\boldsymbol{\Upsilon}) = s$, $\boldsymbol{\Upsilon}$ is rewritten as

$$\begin{aligned} \boldsymbol{\Upsilon} &= \boldsymbol{\Sigma}(\mathbf{z}) - \boldsymbol{\Psi} = \\ &= \begin{bmatrix} \mathbf{I} & \mathbf{T}(\mathbf{x}_{(1)}) & \cdots & \mathbf{T}(\mathbf{x}_{(k)}) \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}(\mathbf{z}) & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}(\mathbf{x}_1) & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \boldsymbol{\Sigma}(\mathbf{x}_k) \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ -\mathbf{T}(\mathbf{x}_{(1)})' \\ \vdots \\ -\mathbf{T}(\mathbf{x}_{(k)})' \end{bmatrix} \\ &= \mathbf{M}_1 \mathbf{M}_2 \mathbf{M}_3, \end{aligned}$$

where \mathbf{I} is the identity matrix of dimension $s + 1 \times s + 1$. In fact, following the rationale as for Theorem 1 and (13), $\text{Ker}(\mathbf{M}_2 \mathbf{M}_3) = \langle \mathbf{1}_s \rangle$, hence $\text{Ker}(\boldsymbol{\Upsilon}) = \langle \mathbf{1}_s \rangle \oplus \{ \mathbf{v} \in \mathbb{R}^{s+1} : \mathbf{M}_2 \mathbf{M}_3 \mathbf{v} \in \text{Ker}(\mathbf{M}_1) \}$. Since $\text{Ker}(\mathbf{M}_1)$ is a space orthogonal to $\text{Im}(\mathbf{M}_1')$, then $\text{Ker}(\mathbf{M}_1) \subseteq \{ (\mathbf{0}, \mathbf{w}) \in \mathbb{R}^{2s+k+1} : \mathbf{w} \in \mathbb{R}^{s+k} \}$ and therefore $\{ \mathbf{v} \in \mathbb{R}^{s+1} : \mathbf{M}_2 \mathbf{M}_3 \mathbf{v} \in \text{Ker}(\mathbf{M}_1) \} = \langle \mathbf{1}_s \rangle$, which leads to $\text{rk}(\boldsymbol{\Upsilon}) = s$. To conclude, (28) derives from Proposition 3. \square