HKU CS Tech Report TR-2006-12

The Mathematics of the Overlapping Chi-square Test[1]

Wai Wan Tsang and Chi-yin Pang

Addresses: Wai Wan Tsang, Computer Science Department, The University of Hong Kong, Pokfulam Road, Hong Kong; Chi-yin Pang, Mathematics Department, Evergreen Valley College, 3095 Yerba Buena Road, San Jose, CA 95135, USA.

Emails: tsang@cs.hku.hk, pangchiyin@yahoo.com

Keywords: Goodness-of-fit test, Chi-square test, Overlapping *m*-tuple test, Serial test.

Abstract

One shortcoming of the Pearson chi-square test is that it only examines the distribution of experiment outcomes, but not their interdependence. For example, the test statistic is invariant against re-ordering the outcomes. To remedy the shortcoming, G. Marsaglia suggested combining consecutive outcomes to form tuples and then examining the distribution of the tuples. The statistic of the test is a quadratic form in the weak inverse of the covariance matrix of the counts of all tuples. As consecutive tuples overlap, we call this test the overlapping chi-square test. Compared with the conventional one, this test checks the interdependence as well as the distribution of experiment outcomes and is thus more comprehensive.

Marsaglia stated that the statistic of the overlapping chi-square test follows a chi-square distribution with the degrees of freedom equal to the rank of the covariance matrix. He sketched the background theory but did not publish any proof. This paper gives a detailed derivation for the distribution of the test statistic.

[1] ***Acknowledgement***: Research supported by the Hong Kong Research Grants Council, Grant HKU 7143/04E.

1. Introduction

In statistics, a goodness-of-fit test is used to check whether a set of samples follows a hypothetical distribution. Such tests are used in model validation and random number generator testing. The most common test for checking discrete samples is the Pearson chi-square test [PK00]. The null hypothesis is that the test samples follow a purported discrete distribution. Suppose that the possible outcomes of an experiment are 0, 1, 2, . . ., $d-1$, with probabilities $P(0)$, $P(1)$, . . ., $P(d-1)$, respectively. The experiment is carried out n times independently and the outcomes are Y_1, Y_2, …, Y_n. Let $N(i)$, $0 \leq \mathrm{i} < d$, be the number of times that i occurs in the outcomes. Note that $\sum_{i=0}^{d-1} P(i) = 1$ and $\sum_{i=0}^{d-1} N(i) = n$. The chi-square statistic is defined as $V = \sum_{i=0}^{d-1} \frac{(N(i) - nP(i))^2}{nP(i)}$. Asymptotically, V follows the chi-square distribution of $d-1$ degrees of freedom.

As pointed out by G. Marsaglia in [MR05], the chi-square test only examines the distribution of the samples but not the interdependency, e.g., the correlation between adjacent outcomes. In order to check the dependency, Marsaglia suggested viewing the outcomes as a circular string and taking $S_1=(Y_1, Y_2, \ldots, Y_t)$, $S_2= (Y_2, Y_3, \ldots, Y_{t+1})$, …, $S_n= (Y_\mathrm{n}, Y_1, \ldots, Y_{t-1})$ as test samples. To avoid two samples overlapping at both ends, we impose a constraint that $n \geq 2t$. The test statistic is generalized to

$$V = \sum_{|\alpha|=t} \frac{(N(\alpha) - nP(\alpha))^2}{nP(\alpha)} - \sum_{|\alpha|=t-1} \frac{(N(\alpha) - nP(\alpha))^2}{nP(\alpha)} . \qquad (1.1)$$

In the formula, $\alpha = a_1 a_2 \ldots a_t$ with $0 \leq a_i < d$. $N(\alpha)$ is the number of times that $S_i=\alpha$, for $i = 1, \ldots, n$. $P(\alpha) = P(a_1)P(a_2)\ldots P(a_t)$ is the probability that a particular test sample, say, S_1, equals α. Asymptotically, V follows the chi-square distribution of $d^t - d^{t-1}$ degrees of freedom. Note that when t=1, the test degenerates to the standard chi-square test. This test has been called the overlapping m-tuple test, overlapping occupancy test, overlapping serial test, generalized chi-square test in the literature. We incline to call it the overlapping chi-square test as it is basically a chi-square test and its samples consist of overlapping outcomes.

George Marsaglia was the first who suggested the test and worked out the distribution of V. However, he only gave a brief background on the theory but did not publish any proof [MR85, MR05]. Wegenkittl described a proof on the distribution of V in his master's thesis. His notation and derivations are complicated [WS96]. Knuth made up two exercises for proving the distribution and sketched the key steps in the suggested answers [No. 25 in 3.3.1 and No. 24 in 3.3.2, KN98]. In this paper, we give a refined and simpler proof for the distribution of V. We mainly use Knuth's notations and fill in all gaps in his sketches. As a matter of fact, some of these gaps are actually grand canyons.

To derive the distribution of V, we need to use a more general theorem on the distribution of a quadratic form of joint normal variables. This theorem was discovered by Good [GD53] and independently by Marsaglia in early the 1950s. The proof of this theorem is described in Section 2. In Section 3, we derive the means and the covariance matrix, C, of the $N(\alpha)$s. Section 4 verifies the formula of a weak inverse of C discovered by G. Marsaglia. This is the most complicated portion of the paper. Section 5 derives the rank of C by an approach that makes use of the weak inverse. This is the only portion of the proof where we deviate from Knuth's suggestions. In Section 6, we show that V is actually the quadratic form defined in Section 2 and therefore follows a chi-square distribution with the degrees of freedom equal to the rank of C. Finally, in the last section, a few practical issues of the test are addressed.

2. Distribution of a quadratic form of joint normal variables

This section derives the distribution of a quadratic form of random variables that have the joint normal distributions.

Suppose $X = \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_n \end{bmatrix}$ is a vector of *n* independent standard normal variables, i.e., with zero mean and unit variance, and $\begin{bmatrix} N_1 \\ N_2 \\ \cdots \\ N_m \end{bmatrix} = AX + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_m \end{bmatrix}$. Then $N = \begin{bmatrix} N_1 \\ N_2 \\ \cdots \\ N_m \end{bmatrix}$ is a vector of joint normal variables with mean equal to $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_m \end{bmatrix}$. The covariance matrix of *N* is $C = (c_{ij})$. Its dimensions are *m*×*m* and $c_{ij} = E[(N_i - \mu_i)(N_j - \mu_j)]$. Suppose $A = (a_{ij})$ is an $m \times n$ matrix of rank *n*.

Theorem 2.1. $C = AA^T$.

Proof:

$c_{ij} = E[(N_i - \mu_i)(N_j - \mu_j)]$

$= E\left[\sum_{1 \le k \le n} a_{ik} X_k \sum_{1 \le l \le n} a_{jl} X_l \right]$

$= \sum_{1 \le k \le n} \sum_{1 \le l \le n} a_{ik} a_{jl} E[X_k X_l]$

$= \sum_{1 \le k \le n} a_{ik} a_{jk}$

The last equality holds because $X_1, ..., X_n$ are independent. $E[X_k X_l]$ equals zero when $k \ne l$. It equals $E[X_k^2]$ when $k = l$. As $E[X_k^2]$ is the variance of X_k, it equals one. Thus, $C = AA^T$.

QED

A weak inverse of *C* is a matrix, $\overline{C} = (\overline{c}_{ij})$, such that $C = C\overline{C}C$. The dimensions of $\overline{C}$ are $m \times n$. The mean-adjusted quadratic form of *N* in $\overline{C}$ is $(N - \mu)^T \overline{C} (N - \mu) = \sum_{1 \le i \le n} \sum_{1 \le j \le n} (N_i - \mu_i)(N_j - \mu_j)\overline{c}_{ij}$. Note that $N - \mu = AX$.

Theorem 2.2. The mean-adjusted quadratic form $(N-\mu)^T\overline{C}(N-\mu)$ follows the chi-square distribution of *n* degrees of freedom.

Proof:
We first derive an identity from the singular value decomposition of *A* and the definition of weak inverse. Then we show that $(N-\mu)^T\overline{C}(N-\mu)=X^TX$. The right hand side is a sum of *n* squares of independent standard normal variables. Thus, $(N-\mu)^T\overline{C}(N-\mu)$ follows the chi-square distribution of *n* degrees of freedom.

Using the singular value decomposition, $A=UDV^T$ where U, V have dimensions $m\times m$ and $n\times n$ respectively. Furthermore, $UU^T=U^TU=I_m$, $VV^T=V^TV=I_n$. D has dimensions $m\times n$ and has all zero off its diagonal. Since A has rank n, D also has rank *n*, and the diagonal elements, $d_{11},d_{22}...,d_{nn}$, of D are non-zero. From Theorem 2.1,
$C=AA^T=(UDV^T)(UDV^T)^T=UDV^TVD^TU^T=UDD^TU^T$.

Let E be an $n\times m$ "diagonal matrix" with diagonal elements defined by $e_{11}=\frac{1}{d_{11}},e_{22}=\frac{1}{d_{22}}...,e_{nn}=\frac{1}{d_{nn}}$, and all other elements equal to 0. Note that $ED=D^TE^T=I_n$. The following deducts the identity $D^TU^T\overline{C}UD=I_n$ from $C\overline{C}C=C$ by substituting *C* with UDD^TU^T and eliminating excessive *D*s using *E*s.

$$C\overline{C}C=C$$
$$(UDD^TU^T)\overline{C}(UDD^TU^T)=(UDD^TU^T)$$
$$U^T(UDD^TU^T)\overline{C}(UDD^TU^T)U=U^T(UDD^TU^T)U$$
$$DD^TU^T\overline{C}UDD^T=DD^T$$
$$EDD^TU^T\overline{C}UDD^TE^T=EDD^TE^T$$
$$D^TU^T\overline{C}UD=I_n$$

The following shows that the quadratic form equals X^TX.
$$(N-\mu)^T\overline{C}(N-\mu)$$
$$=(AX)^T\overline{C}(AX)$$
$$=X^TA^T\overline{C}AX$$
$$=X^T(UDV^T)^T\overline{C}(UDV^T)X \qquad \text{//using } A=UDV^T$$
$$=X^TV(D^TU^T\overline{C}UD)V^TX$$
$$=X^TVI_nV^TX \qquad \text{//using the above identity}$$
$$=X^TVV^TX$$
$$=X^TX$$

QED

3. Mean and covariance matrix of *N*

With reference to the definitions given in Section 1, Let $N = \begin{bmatrix} N(\alpha_1) \\ N(\alpha_2) \\ \cdots \\ N(\alpha_{d^t}) \end{bmatrix}$ be the vector consisting all $N(\alpha)$s where the length of α is *t*. In this section, we derive the formula for the mean and the covariance matrix of *N*.

First we give a few definitions needed for formulation. These definitions are suggested by D. Knuth. Let $\alpha = a_1 a_2 ... a_t$ and $\beta = b_1 b_2 ... b_t$ be two strings of the experiment outcomes.

Definition 3.1. $K(\alpha,\beta) \equiv \begin{cases} \dfrac{1}{P(\alpha)} & \textit{if } \alpha = \beta, \\ 0 & \textit{otherwise.} \end{cases}$

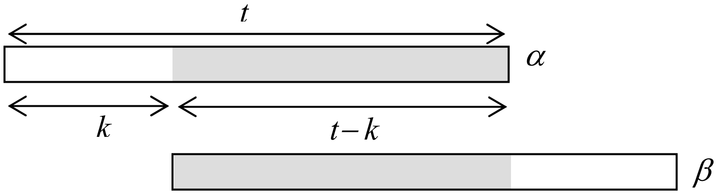
For convenience, let *K*(*null, null*) = 1. That happens when both α and β are empty string.

Definition 3.2. Let $k:\alpha = a_1 a_2 ... a_{k-1} a_k$ denote the leftmost *k* digits of α and $\alpha : k = a_{t-k+1} a_{t-k+2} ... a_{t-1} a_t$ denote the rightmost *k* digits of *α*.

Definition 3.3. Given strings, α and β, of length *t*, and a "shifting index" $-t < k < t$,

$$T_k(\alpha,\beta) \equiv \begin{cases} K(t+k:\alpha, \beta:t+k) - 1, & \text{if } k < 0, \\ K(\alpha:t-k, t-k:\beta) - 1, & \text{if } k \geq 0. \end{cases}$$

The following figure pictures what $T_k(\alpha,\beta)$ operates on when *k* > 0. First align α and β at the left end; shift β to the right by *k* digits; then compare the overlapping portions of α and β.



When *k* = 0, there is no shifting. When *k* < 0, β is shifted to the left by –*k* digits.

The following are several properties of $K(\alpha,\beta)$ and $T_k(\alpha,\beta)$ that are useful in proving theorems.

Property 3.4. $K(\alpha,\beta) = K(\beta,\alpha)$ and $T_k(\alpha,\beta) = T_{-k}(\beta,\alpha)$.

Property 3.5. If $\alpha = \alpha_1\alpha_2$, $\beta = \beta_1\beta_2$ and $|\alpha_1| = |\beta_1|$, $|\alpha_2| = |\beta_2|$, then $K(\alpha,\beta) = K(\alpha_1,\beta_1)K(\alpha_2,\beta_2)$.

Property 3.6. $\sum_{|\gamma|=k} P(\gamma) = 1$.

Property 3.7. $\sum_{0 \leq a < d} P(a)K(a,b) = P(b)K(b,b) = 1$.

Theorem 3.8. The mean of *N* is $\mu = \begin{bmatrix} nP(\alpha_1) \\ nP(\alpha_2) \\ \ldots \\ nP(\alpha_{d^t}) \end{bmatrix}$.

Proof:
For $1 \le i \le n$, let Z_i be a random variable such that $Z_i = 1$ if S_i equals α, otherwise 0. The mean of Z_i is equal to $P(\alpha)$ which does not depend on *i*. $N(\alpha) = \sum_{i=1}^{n} Z_i$ and its mean is the sum of the means of Z_i's. Thus, the mean of $N(\alpha)$, denoted as $\mu_{N(\alpha)}$, is $nP(\alpha)$.

QED

Theorem 3.9. Let $C = (c_{\alpha\beta})$ be the covariance matrix of *N*. $c_{\alpha\beta} = nP(\alpha\beta)\sum_{|k|<t} T_k(\alpha,\beta)$.

Proof:
Let Z_i be defined as in Theorem 3.8 and W_i be defined in the same way but with respect toβ. That is, $W_i = 1$ if S_i equals β, otherwise 0. $N(\alpha) = \sum_{i=1}^{n} Z_i$ and $N(\beta) = \sum_{i=1}^{n} W_i$.

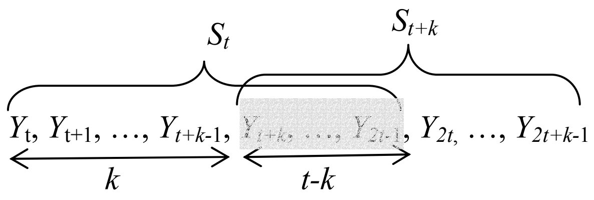
$$
\begin{aligned}
& c_{\alpha\beta} \\
& = E[(N(\alpha) - \mu_{N(\alpha)})(N(\beta) - \mu_{N(\beta)})] \\
& = E[N(\alpha)N(\beta)] - \mu_{N(\alpha)}\mu_{N(\beta)} \\
& = E\left[\left(\sum_{1\le i\le n} Z_i\right)\left(\sum_{1\le j\le n} W_j\right)\right] - nP(\alpha)nP(\beta) \\
& = \sum_{1\le i\le n}\left(\sum_{1\le j\le n} E\left[Z_i W_j\right]\right) - n^2 P(\alpha\beta) \\
& = n\sum_{1\le j\le n} E\left[Z_t W_j\right] - n^2 P(\alpha\beta) \\
& = nP(\alpha\beta)\sum_{1\le j\le n} \frac{E\left[Z_t W_j\right]}{P(\alpha\beta)} - n \\
& = nP(\alpha\beta)\sum_{1\le j\le n}\left(\frac{E\left[Z_t W_j\right]}{P(\alpha\beta)} - 1\right) \\
& == nP(\alpha\beta)\sum_{|k|<t}\left(\frac{E\left[Z_t W_{t+k}\right]}{P(\alpha\beta)} - 1\right)
\end{aligned}
$$

The fifth equality holds because the experiment outcomes are treated as a ring and all positions are "equal". Thus, the value of the inner summation is identical for all *i*. In the expression, we set it equal to the sum where $i = t$.

The last equality holds because when S_t and S_j do not overlap, Z_t and W_j are independent random variables. Therefore, $E[Z_t W_j] = E[Z_t]E[W_j] = P(\alpha\beta)$. Consequently, $\frac{E\left[Z_t W_j\right]}{P(\alpha\beta)} - 1 = 0$ for these cases.

The following shows that $\frac{E[Z_t W_{t+k}]}{P(\alpha\beta)} - 1$ is indeed $T_k(\alpha,\beta)$ and that completes the proof.

Consider the situation when $k > 0$.


S_t
S_{t+k}
Y_t, Y_{t+1}, …, Y_{t+k-1}, Y_{t+k}, …, Y_{2t-1}, Y_{2t}, …, Y_{2t+k-1}
k
t-k

The only case where $Z_t W_j = 1$ is when S_t=α , S_{t+k}=β , provided that α:t–k equals t–k:β . For any α and β, the probability of $Z_t W_j = 1$ is $P(\alpha\beta)K(\alpha : t-k, t-k : \beta)$ and this is also the expected value of $Z_t W_j$. Now it is easy to see that

$$
\begin{aligned}
&\frac{E[Z_t W_{t+k}]}{P(\alpha\beta)} - 1 \\
&= \frac{P(\alpha\beta)K(\alpha : t-k, t-k : \beta)}{P(\alpha\beta)} - 1 . \\
&= K(\alpha : t-k, t-k : \beta) - 1 \\
&= T_k(\alpha, \beta)
\end{aligned}
$$

The case of $k < 0$ is identical to the above except S_{t+k} is in the left hand side of S_t. The case of $k = 0$ is trivial.

QED

4. Weak inverse of the covariance matrix

Let $\overline{C}$ be the $d^t \times d^t$ matrix with entries $\overline{c}_{\alpha\beta} \equiv \frac{1}{n}\left(K(\alpha,\beta) - K(t-1:\alpha, t-1:\beta)\right)$. We verify that $\overline{C}$ is a "weak inverse" of *C*, i.e., $C = C\overline{C}C$, in Theorem 4.1- 4.4.

Theorem 4.1. The $\alpha\beta$ entry of $C\overline{C}C$ is:

$$\left(C\overline{C}C\right)_{\alpha\beta} = nP(\alpha\beta)\sum_{-t<k<t}\sum_{-t<l<t}\sum_{|\gamma|=t-1}\sum_{\substack{0\le a<d\\0\le b<d}} P(\gamma ab)T_k(\alpha,\gamma a)(K(a,b)-1)T_l(\gamma b,\beta)$$

Proof:

$$\left(C\overline{C}C\right)_{\alpha\beta}$$

$$= \sum_{|\delta|=t}\sum_{|\varepsilon|=t} c_{\alpha\delta}\overline{c}_{\delta\varepsilon}c_{\varepsilon\beta}$$

$$= \sum_{|\delta|=t}\sum_{|\varepsilon|=t}\left(nP(\alpha\delta)\sum_{|k|<t}T_k(\alpha,\delta)\right)\frac{1}{n}\left(K(\delta,\varepsilon) - K(t-1:\delta, t-1:\varepsilon)\right)\left(nP(\varepsilon\beta)\sum_{|l|<t}T_l(\varepsilon,\beta)\right)$$

$$= nP(\alpha\beta)\sum_{|\delta|=t}\sum_{|\varepsilon|=t}P(\delta\varepsilon)\sum_{|k|<t}T_k(\alpha,\delta)\left(K(\delta,\varepsilon) - K(t-1:\delta, t-1:\varepsilon)\right)\sum_{|l|<t}T_l(\varepsilon,\beta)$$

$$= nP(\alpha\beta)\sum_{|\gamma|=t-1}\sum_{0\le a<d}\sum_{|\zeta|=t-1}\sum_{0\le b<d}P(\gamma a\zeta b)\sum_{|k|<t}T_k(\alpha,\gamma a)\left(K(\gamma a,\zeta b) - K(\gamma,\zeta)\right)\sum_{|l|<t}T_l(\zeta b,\beta)$$

Replace δ with γa, ε with ζb.

$$= nP(\alpha\beta)\sum_{|\gamma|=t-1}\sum_{\substack{0\le a<d\\0\le b<d}}P(\gamma a\gamma b)\sum_{|k|<t}T_k(\alpha,\gamma a)\left(K(\gamma a,\gamma b) - K(\gamma,\gamma)\right)\sum_{|l|<t}T_l(\gamma b,\beta)$$

$$= nP(\alpha\beta)\sum_{|\gamma|=t-1}\sum_{\substack{0\le a<d\\0\le b<d}}P(\gamma a\gamma b)\sum_{|k|<t}T_k(\alpha,\gamma a)\left(\frac{K(a,b)}{P(\gamma)} - \frac{1}{P(\gamma)}\right)\sum_{|l|<t}T_l(\gamma b,\beta)$$

$$= nP(\alpha\beta)\sum_{|\gamma|=t-1}\sum_{\substack{0\le a<d\\0\le b<d}}P(\gamma ab)\sum_{|k|<t}\sum_{|l|<t}T_k(\alpha,\gamma a)(K(a,b)-1)T_l(\gamma b,\beta)$$

$$= nP(\alpha\beta)\sum_{-t<k<t}\sum_{-t<l<t}\left(\sum_{|\gamma|=t-1}\sum_{\substack{0\le a<d\\0\le b<d}}P(\gamma ab)T_k(\alpha,\gamma a)(K(a,b)-1)T_l(\gamma b,\beta)\right)$$

In Step 5, the summation over ζ is eliminated because $K(\gamma a,\zeta b) - K(\gamma,\zeta)$ is zero except possibly when $\zeta = \gamma$.

QED

The next theorem shows that the bracketed expression in the proof above is zero for $-t < l < 0$. By symmetry, the expression is also zero when $0 < k < t$. Thus,

$$\left(C\overline{C}C\right)_{\alpha\beta} = nP(\alpha\beta)\sum_{-t<k\le 0}\sum_{0\le l<t}\sum_{|\gamma|=t-1}\sum_{\substack{0\le a<d\\0\le b<d}}P(\gamma ab)T_k(\alpha,\gamma a)(K(a,b)-1)T_l(\gamma b,\beta).$$

Theorem 4.2. For $-t<l<0$, $\sum_{|\gamma|=t-1}\sum_{\substack{0\le a<d\\ 0\le b<d}} P(\gamma ab)T_k(\alpha,\gamma a)(K(a,b)-1)T_l(\gamma b,\beta)=0$.

Proof:
When $l<0$, $T_l(\gamma b,\beta)$ concerns with the *t*+*l*: γb and β :*t*+*l*. The value of *b* does not affect the value of $T_l(\gamma b,\beta)$ at all. Thus, we can replace $T_l(\gamma b,\beta)$ with $T_l(\gamma a,\beta)$ and factor it out from the summation over *b*.

$$\sum_{|\gamma|=t-1}\sum_{\substack{0\le a<d\\ 0\le b<d}} P(\gamma ab)T_k(\alpha,\gamma a)(K(a,b)-1)T_l(\gamma b,\beta)$$

$$=\sum_{|\gamma|=t-1}\sum_{0\le a<d} P(\gamma a)T_k(\alpha,\gamma a)T_l(\gamma a,\beta)\sum_{0\le b<d}P(b)(K(a,b)-1)$$

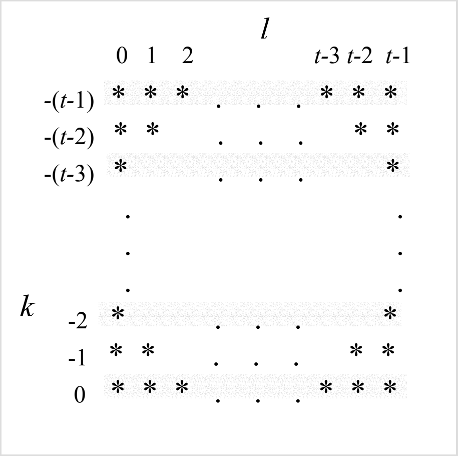
$$=\sum_{|\gamma|=t-1}\sum_{0\le a<d} P(\gamma a)T_k(\alpha,\gamma a)T_l(\gamma b,\beta)\left(\sum_{0\le b<d}P(b)K(a,b)-\sum_{0\le b<d}P(b)\right)$$

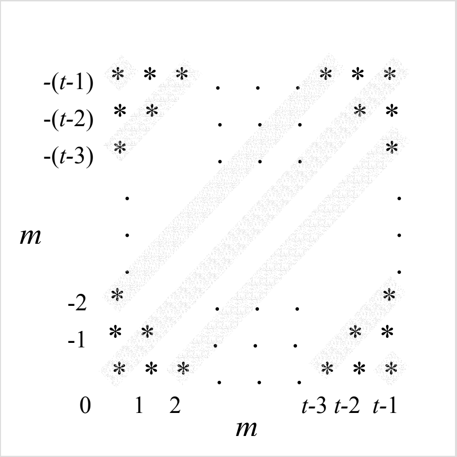
$$=\sum_{|\gamma|=t-1}\sum_{0\le a<d} P(\gamma a)T_k(\alpha,\gamma a)T_l(\gamma b,\beta)(1-1)$$

$$=0$$

QED

Let $F_{kl}=\sum_{|\gamma|=t-1}\sum_{\substack{0\le a<d\\ 0\le b<d}} P(\gamma ab)T_k(\alpha,\gamma a)(K(a,b)-1)T_l(\gamma b,\beta)$, we have proven $\left(C\overline{C}C\right)_{\alpha\beta}=nP(\alpha\beta)\sum_{-t<k\le 0}\sum_{0\le l<t}F_{kl}$ which sums all F_{kl} row by row as shown in the left figure below.





Let us re-index the element so that the summation is carried out diagonal by diagonal as shown in the right. The new formula is $\left(C\overline{C}C\right)_{\alpha\beta}=nP(\alpha\beta)\left(\sum_{-t<m\le 0}\sum_{-t<k\le m}F_{k,m-k}+\sum_{0<m<t}\sum_{m-t<k\le 0}F_{k,m-k}\right)$. From top-left to bottom-right, each diagonal corresponds to one value of *m*, where *m* ranges from –(*t*–1) to *t*–1. Amazingly, the sum of the elements in a diagonal of *m* is equal to $T_m(\alpha,\beta)$.

Theorem 4.3. For $-t < m \le 0$, $\sum_{-t<k\le m} F_{k,m-k} = T_m(\alpha,\beta)$.

Proof:

$$\sum_{-t<k\le m} F_{k,m-k}$$

$$= \sum_{-t<k\le m}\sum_{|\gamma|=t-1}\sum_{\substack{0\le a<d\\ 0\le b<d}} P(\gamma ab)T_k(\alpha,\gamma a)(K(a,b)-1)T_{m-k}(\gamma b,\beta)$$

$$= \sum_{-t<k\le m}\sum_{|\gamma|=t-1} P(\gamma)\sum_{0\le a<d} P(a)T_k(\alpha,\gamma a)\left(\sum_{0\le b<d} P(b)K(a,b)T_{m-k}(\gamma b,\beta) - \sum_{0\le b<d} P(b)T_{m-k}(\gamma b,\beta)\right)$$

$$= \sum_{-t<k\le m}\sum_{|\gamma|=t-1} P(\gamma)\sum_{0\le a<d} P(a)T_k(\alpha,\gamma a)\left(T_{m-k}(\gamma a,\beta) - \sum_{0\le b<d} P(b)T_{m-k}(\gamma b,\beta)\right) \qquad \text{// } m-k \text{ is positive}$$

$$= \sum_{-t<k\le m}\sum_{|\gamma|=t-1} P(\gamma)\sum_{0\le a<d} P(a)T_k(\alpha,\gamma a)\left(K(\gamma a : t-m+k, t-m+k : \beta) - \sum_{0\le b<d} P(b)K(\gamma b : t-m+k, t-m+k : \beta)\right)$$

$$= \sum_{-t<k\le m}\sum_{|\gamma|=t-1} P(\gamma)\sum_{0\le a<d} P(a)T_k(\alpha,\gamma a)\left(\begin{array}{l} K(\gamma a : t-m+k, t-m+k : \beta) \\ -\sum_{0\le b<d} P(b)K(\gamma : t-m+k-1, t-m+k-1 : \beta)K(b, b_{t-m+k}) \end{array}\right)$$

$$= \sum_{-t<k\le m}\sum_{|\gamma|=t-1} P(\gamma)\sum_{0\le a<d} P(a)T_k(\alpha,\gamma a)\big(K(\gamma a : t-m+k, t-m+k : \beta) - K(\gamma : t-m+k-1, t-m+k-1 : \beta)\big)$$

$$= \sum_{-t<k\le m}\sum_{|\gamma|=t-1} P(\gamma)\sum_{0\le a<d} P(a)(K(t+k : \alpha, \gamma a : t+k)-1)\left(\begin{array}{l} K(\gamma a : t-m+k, t-m+k : \beta) \\ -K(\gamma : t-m+k-1, t-m+k-1 : \beta) \end{array}\right)$$

$$= \sum_{-t<k\le m}\sum_{|\gamma|=t-1} P(\gamma)\left(\begin{array}{l} \sum_{0\le a<d} P(a)K(t+k : \alpha, \gamma a : t+k)K(\gamma a : t-m+k, t-m+k : \beta) \\ -\sum_{0\le a<d} P(a)K(t+k : \alpha, \gamma a : t+k)K(\gamma : t-m+k-1, t-m+k-1 : \beta) \\ -\sum_{0\le a<d} P(a)K(\gamma a : t-m+k, t-m+k : \beta) \\ +\sum_{0\le a<d} P(a)K(\gamma : t-m+k-1, t-m+k-1 : \beta) \end{array}\right)$$

$$= \sum_{-t<k\le m}\sum_{|\gamma|=t-1} P(\gamma)\big(Q_1 - Q_2 - Q_3 + Q_4\big) \qquad \text{//Each } Q \text{ represents one summation in the above bracket.}$$

The following computes the summations Q_1, Q_2, Q_3 and Q_4 one at a time:

$$Q_1 = \sum_{0\le a<d} P(a)K(t+k:\alpha,\gamma a:t+k)K(\gamma a:t-m+k,t-m+k:\beta)$$

$$= \sum_{0\le a<d} P(a)K(t+k-1:\alpha,\gamma:t+k-1)K(a_{t+k},a)K(\gamma:t-m+k-1,t-m+k-1:\beta)K(a,b_{t-m+k})$$

$$= K(t+k-1:\alpha,\gamma:t+k-1)K(\gamma:t-m+k-1,t-m+k-1:\beta)\sum_{0\le a<d} P(a)K(a_{t+k},a)K(a,b_{t-m+k})$$

$$= K(t+k-1:\alpha,\gamma:t+k-1)K(\gamma:t-m+k-1,t-m+k-1:\beta)K(a_{t+k},b_{t-m+k})$$

$$Q_2 = \sum_{0\le a<d} P(a)K(t+k:\alpha,\gamma a:t+k)K(\gamma:t-m+k-1,t-m+k-1:\beta)$$

$$= K(\gamma:t-m+k-1,t-m+k-1:\beta)\sum_{0\le a<d} P(a)K(t+k:\alpha,\gamma a:t+k)$$

$$= K(\gamma:t-m+k-1,t-m+k-1:\beta)\sum_{0\le a<d} P(a)K(t+k-1:\alpha,\gamma:t+k-1)K(a_{t+k},a)$$

$$= K(\gamma:t-m+k-1,t-m+k-1:\beta)K(t+k-1:\alpha,\gamma:t+k-1)\sum_{0\le a<d} P(a)K(a_{t+k},a)$$

$$= K(\gamma:t-m+k-1,t-m+k-1:\beta)K(t+k-1:\alpha,\gamma:t+k-1)$$

$$Q_3 = \sum_{0\le a<d} P(a)K(\gamma a:t-m+k,t-m+k:\beta)$$

$$= \sum_{0\le a<d} P(a)K(\gamma:t-m+k-1,t-m+k-1:\beta)K(a,b_{t-m+k})$$

$$= K(\gamma:t-m+k-1,t-m+k-1:\beta)\sum_{0\le a<d} P(a)K(a,b_{t-m+k})$$

$$= K(\gamma:t-m+k-1,t-m+k-1:\beta)$$

$$Q_4 = \sum_{0\le a<d} P(a)K(\gamma:t-m+k-1,t-m+k-1:\beta)$$

$$= K(\gamma:t-m+k-1,t-m+k-1:\beta)\sum_{0\le a<d} P(a)$$

$$= K(\gamma:t-m+k-1,t-m+k-1:\beta)$$

Substituting the last four summations into the main computation, Q_3 is canceled with Q_4 and we get

$$\sum_{-t<k\le m} F_{k,m-k}$$

$$= \sum_{-t<k\le m}\sum_{|\gamma|=t-1} P(\gamma)\begin{pmatrix} K(t+k-1:\alpha,\gamma:t+k-1)K(\gamma:t-m+k-1,t-m+k-1:\beta)K(a_{t+k},b_{t-m+k}) \\ -K(\gamma:t-m+k-1,t-m+k-1:\beta)K(t+k-1:\alpha,\gamma:t+k-1) \end{pmatrix}$$

$$= \sum_{-t<k\le m}(K(a_{t+k},b_{t-m+k})-1)\sum_{|\gamma|=t-1}P(\gamma)K(t+k-1:\alpha,\gamma:t+k-1)K(\gamma:t-m+k-1,t-m+k-1:\beta)$$

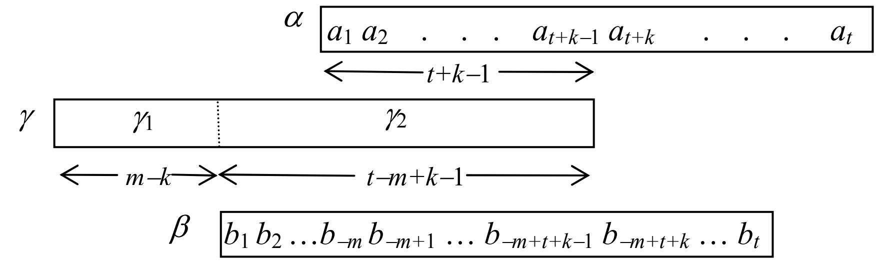
$$= \sum_{-t<k\le m}(K(a_{t+k},b_{t-m+k})-1)K(a_1a_2...a_{t+k-1},b_{-m+1}b_{-m+2}...b_{-m+t+k-1})$$

$$= \sum_{-t<k\le m}(K(a_1a_2...a_{t+k-1},b_{-m+1}b_{-m+2}...b_{-m+t+k-1})K(a_{t+k},b_{t-m+k})-K(a_1a_2...a_{t+k-1},b_{-m+1}b_{-m+2}...b_{-m+t+k-1}))$$

$$= \sum_{-t<k\le m}(K(a_1a_2...a_{t+k},b_{-m+1}b_{-m+2}...b_{-m+t+k})-K(a_1a_2...a_{t+k-1},b_{-m+1}b_{-m+2}...b_{-m+t+k-1}))$$

$$\begin{aligned} = \quad & K(a_1,b_{-m+1})-K(null,null) \\ & +K(a_1a_2,b_{-m+1}b_{-m+2})-K(a_1,b_{-m+1}) \\ & \ldots \\ & +K(a_1a_2...a_{t+m},b_{-m+1}b_{-m+2}...b_t)-K(a_1a_2...a_{t+m-1},b_{-m+1}b_{-m+2}...b_{t-1}) \\ = {} & K(a_1a_2...a_{t+m},b_{-m+1}b_{-m+2}...b_t)-1 \\ = {} & K(t+m:\alpha,\beta:t+m)-1 \\ = {} & T_m(\alpha,\beta) \end{aligned}$$

The following illustrates the relations between α, β and γ in the second step of the above derivation.


α
$a_1\ a_2\ \ .\ \ .\ \ .\ \ a_{t+k-1}\ a_{t+k}\ \ .\ \ .\ \ .\ \ a_t$
$t+k-1$
γ
γ_1
γ_2
$m-k$
$t-m+k-1$
β
$b_1\ b_2\ \ldots b_{-m}\ b_{-m+1}\ \ldots\ b_{-m+t+k-1}\ b_{-m+t+k}\ \ldots\ b_t$

$$\sum_{|\gamma|=t-1}P(\gamma)K(t+k-1:\alpha,\gamma:t+k-1)K(\gamma:t-m+k-1,t-m+k-1:\beta)$$

$$= \sum_{|\gamma_1|=m-k}\ \sum_{|\gamma_2|=t-m+k-1}P(\gamma_1)P(\gamma_2)K(t+k-1:\alpha,\gamma_2:t+k-1)K(\gamma_2,t-m+k-1:\beta)$$

$$= \sum_{|\gamma_1|=m-k}P(\gamma_1)P(t-m+k-1:\beta)K(t+k-1:\alpha,(t-m+k-1:\beta):t+k-1)\frac{1}{P(t-m+k-1:\beta)}$$

$$= K(t+k-1:\alpha,(t-m+k-1:\beta):t+k-1)\left(\sum_{|\gamma_1|=m-k}P(\gamma_1)\right)$$

$$= K(a_1a_2...a_{t+k-1},b_{-m+1}b_{-m+2}...b_{-m+t+k-1})$$

QED

Theorem 4.4. For $0 < m < t$, $\sum_{m-t<k\le 0} F_{k,m-k} = T_m(\alpha,\beta)$.

Theorem 4.3 considers the cases where *m* is negative (β is shifted to the left of α). This theorem takes care of the cases where *m* is positive (β is shifted to the right of α). The proof is parallel to that of Theorem 4.3 and is therefore skipped.

Recall that we have proven $\left(C\overline{C}C\right)_{\alpha\beta} = nP(\alpha\beta)\left(\sum_{-t<m\le 0}\sum_{-t<k\le m} F_{k,m-k} + \sum_{0<m<t}\sum_{m-t<k\le 0} F_{k,m-k}\right)$. By Theorems 4.3 and 4.4, the inner summations are equal to $T_m(\alpha,\beta)$. Thus,

$$\begin{aligned}
&\left(C\overline{C}C\right)_{\alpha\beta} \\
&= nP(\alpha\beta)\left(\sum_{-t<m\le 0} T_m(\alpha,\beta) + \sum_{0<m<t} T_m(\alpha,\beta)\right) \\
&= nP(\alpha\beta)\sum_{|m|<t} T_m(\alpha,\beta) \\
&= c_{\alpha\beta}
\end{aligned}.$$

We conclude that $C\overline{C}C = C$. By definition, $\overline{C}$ is a weak inverse of *C*.

5. The rank of the covariance matrix

We encounter problems in following Knuth's hints on finding the rank of the covariance matrix *C*. In the suggested solution of Ex. 24, Section 3.3.2 [KN98], "These variables ($N(\alpha)$s) are subject to the constraint $\sum_{a=0}^{d-1} N(\alpha a) = \sum_{a=0}^{d-1} N(a\alpha)$ for each of the d^{t-1} strings α, but all other linear constraints are derivable from these." First, among the d^{t-1} constraints of the form $\sum_{a=0}^{d-1} N(\alpha a) = \sum_{a=0}^{d-1} N(a\alpha)$, only $d^{t-1}-1$ are independent. The remaining one is a linear combination of the others. Second, the constraint $\sum_{|\alpha|=t} N(\alpha) = n$ cannot be derived from the above. Lastly, we do not see why all other linear constraints are derivable from this set of constraints.

A different approach for finding the rank of *C* was suggested in [WS96]. An idempotent matrix is a matrix such that its square equals itself, i.e., $BB = B$. The rank of an idempotent matrix is equal to the sum of its diagonal element. It is easy to see that $C\overline{C}$ is an idempotent matrix as $(C\overline{C})(C\overline{C}) = C\overline{C}$. The following theorem states the rank of $C\overline{C}$. Theorem 5.2 argues that the rank of C equals to that of $C\overline{C}$.

Theorem 5.1. The rank of $C\overline{C}$ is $d^t - d^{t-1}$.

Proof:
$C\overline{C}$ is an idempotent matrix. Its rank equals the sum of its diagonal elements, say, Q.

$$Q = \sum_{|\alpha|=t}\sum_{|\beta|=t} c_{\alpha\beta}\bar{c}_{\beta\alpha}$$

$$= \sum_{|\alpha|=t}\sum_{|\beta|=t}\left(P(\alpha\beta)\sum_{|k|<t} T_k(\alpha,\beta)\right)\left(K(\beta,\alpha) - K(t-1:\beta, t-1:\alpha)\right)$$

$$= \sum_{|\alpha|=t}\sum_{|\beta|=t} P(\alpha\beta)K(\beta,\alpha)\sum_{|k|<t} T_k(\alpha,\beta) - \sum_{|\alpha|=t}\sum_{|\beta|=t} P(\alpha\beta)K(t-1:\beta, t-1:\alpha)\sum_{|k|<t} T_k(\alpha,\beta)$$

Let Q_1 and Q_2 be the first and second terms such that $Q = Q_1 - Q_2$. We compute Q_1 and Q_2 separately.

$$Q_1 = \sum_{|\alpha|=t} \sum_{|\beta|=t} P(\alpha\beta) K(\beta,\alpha) \sum_{|k|<t} T_k(\alpha,\beta)$$

$$= \sum_{|\alpha|=t} P(\alpha\alpha) K(\alpha,\alpha) \sum_{|k|<t} T_k(\alpha,\alpha)$$

$$= \sum_{|\alpha|=t} P(\alpha) \sum_{|k|<t} T_k(\alpha,\alpha)$$

$$= \sum_{|\alpha|=t} P(\alpha) \left(\sum_{-t<k<0} T_k(\alpha,\alpha) + T_0(\alpha,\alpha) + \sum_{0<k<t} T_k(\alpha,\alpha) \right)$$

$$= \sum_{|\alpha|=t} P(\alpha) \left(T_0(\alpha,\alpha) + 2 \sum_{0<k<t} T_k(\alpha,\alpha) \right) \qquad // \because T_k(\alpha,\alpha) = T_{-k}(\alpha,\alpha)$$

$$= \sum_{|\alpha|=t} P(\alpha) \big(K(\alpha,\alpha) - 1 \big) + 2 \sum_{|\alpha|=t} P(\alpha) \sum_{0<k<t} T_k(\alpha,\alpha)$$

$$= \sum_{|\alpha|=t} P(\alpha) \left(\frac{1}{P(\alpha)} - 1 \right) + 2 \sum_{|\alpha|=t} P(\alpha) \sum_{0<k<t} T_k(\alpha,\alpha)$$

$$= \sum_{|\alpha|=t} \big(1 - P(\alpha) \big) + 2 \sum_{|\alpha|=t} P(\alpha) \sum_{0<k<t} T_k(\alpha,\alpha)$$

$$= \sum_{|\alpha|=t} 1 - \sum_{|\alpha|=t} P(\alpha) + 2 \sum_{|\alpha|=t} P(\alpha) \sum_{0<k<t} T_k(\alpha,\alpha)$$

$$= \big(d^t - 1 \big) + 2 \sum_{|\alpha|=t} P(\alpha) \sum_{0<k<t} T_k(\alpha,\alpha)$$

$$Q_2 = \sum_{|\alpha|=t} \sum_{|\beta|=t} P(\alpha\beta) K(t-1:\beta, t-1:\alpha) \sum_{|k|<t} T_k(\alpha,\beta)$$

$$= \sum_{|\alpha'|=t-1} \sum_{0\le a<d} \sum_{0\le b<d} \sum_{|\beta'|=t-1} P(\alpha' a \beta' b) K(\beta',\alpha') \sum_{|k|<t} T_k(\alpha' a, \beta' b) \qquad // \alpha = \alpha' a, \beta = \beta' b, \text{where } |\alpha'| = |\beta'| = t$$

$$= \sum_{|\alpha'|=t-1} \sum_{0\le a<d} \sum_{0\le b<d} P(\alpha' a \alpha' b) K(\alpha',\alpha') \sum_{|k|<t} T_k(\alpha' a, \alpha' b)$$

$$= \sum_{|\alpha'|=t-1} \sum_{0\le a<d} \sum_{0\le b<d} P(\alpha' a \alpha' b) \frac{1}{P(\alpha')} \left(T_0(\alpha' a, \alpha' b) + \sum_{-t<k<0} T_k(\alpha' a, \alpha' b) + \sum_{0<k<t} T_k(\alpha' a, \alpha' b) \right)$$

$$= \sum_{|\alpha'|=t-1} \sum_{0\le a<d} \sum_{0\le b<d} P(\alpha' ab) \big(K(\alpha' a, \alpha' b) - 1 \big)$$

$$+ \sum_{|\alpha'|=t-1} \sum_{0\le a<d} \sum_{0\le b<d} P(\alpha' ab) \sum_{-t<k<0} T_k(\alpha' a, \alpha' b)$$

$$+ \sum_{|\alpha'|=t-1} \sum_{0\le a<d} \sum_{0\le b<d} P(\alpha' ab) \sum_{0<k<t} T_k(\alpha' a, \alpha' b)$$

Let Q_{21}, Q_{22}, Q_{22} be the three terms of the last expression, then $Q_2 = Q_{21} + Q_{22} + Q_{23}$.

$$Q_{21} = \sum_{|\alpha'|=t-1} \sum_{0\le a<d} \sum_{0\le b<d} P(\alpha' ab)\big(K(\alpha' a, \alpha' b) - 1\big)$$

$$= \sum_{|\alpha'|=t-1} P(\alpha') \sum_{0\le a<d} P(a) \left(\sum_{0\le b<d} P(b) K(\alpha' a, \alpha' b) - \sum_{0\le b<d} P(b) \right)$$

$$= \sum_{|\alpha'|=t-1} \sum_{0\le a<d} \left(P(\alpha' aa) \frac{1}{P(\alpha' a)} - P(\alpha' a) \sum_{0\le b<d} P(b) \right)$$

$$= \sum_{|\alpha'|=t-1} \sum_{0\le a<d} \big(P(a) - P(\alpha' a)\big)$$

$$= \sum_{|\alpha'|=t-1} \big(1 - P(\alpha')\big) \sum_{0\le a<d} P(a)$$

$$= \sum_{|\alpha|=t-1} 1 - \sum_{|\alpha'|=t-1} P(\alpha')$$

$$= d^{t-1} - 1$$

$$Q_{22} = \sum_{|\alpha'|=t-1} \sum_{0\le a<d} \sum_{0\le b<d} P(\alpha' ab) \sum_{-t<k<0} T_k(\alpha' a, \alpha' b)$$

$$= \sum_{|\alpha'|=t-1} \sum_{0\le a<d} \sum_{0\le b<d} P(\alpha' ab) \sum_{0<k<t} T_k(\alpha' b, \alpha' a)$$

$$= \sum_{|\alpha'|=t-1} \sum_{0\le a<d} \sum_{0\le b<d} P(\alpha' ab) \sum_{0<k<t} T_k(\alpha' a, \alpha' b) \qquad // \because \text{The roles of } a \text{ and } b \text{ are symmetric in the formula.}$$

$$= Q_{23}$$

$$Q_{23} = \sum_{|\alpha'|=t-1} \sum_{0\le a<d} \sum_{0\le b<d} P(\alpha' ab) \sum_{0<k<t} T_k(\alpha' a, \alpha' b)$$

$$= \sum_{|\alpha'|=t-1} \sum_{0\le a<d} \sum_{0\le b<d} P(\alpha' ab) \sum_{0<k<t} T_k(\alpha' a, \alpha' a)$$

$$= \sum_{|\alpha|=t} \sum_{0\le b<d} P(\alpha b) \sum_{0<k<t} T_k(\alpha, \alpha)$$

$$= \sum_{|\alpha|=t} P(\alpha) \sum_{-t<k<0} T_k(\alpha, \alpha) \sum_{0\le b<d} P(b)$$

$$= \sum_{|\alpha|=t} P(\alpha) \sum_{-t<k<0} T_k(\alpha, \alpha)$$

The first equality holds because when $k > 0$, $T_k(\alpha' a, \alpha' b)$ does not really depend on the last digit of the second argument. Therefore, we can replace $T_k(\alpha' a, \alpha' b)$ with $T_k(\alpha' a, \alpha' a)$.

Finally, $Q = Q_1 - Q_{21} - Q_{22} - Q_{23} = d^t - d^{t-1}$.

QED

Theorem 5.2. The rank of C is $d^t - d^{t-1}$.

Proof:

Suppose *RS*=*T*. The rank of *T* is less than or equal to the rank of *R* or *S*. Therefore, the rank of $C\overline{C}$ is bounded by the rank of *C*, i.e., rank($C\overline{C}$) ≤ rank(*C*) . On the other hand, since $(C\overline{C})C = C$, rank($C\overline{C}$) ≥ rank(*C*). Thus, rank(*C*) = rank($C\overline{C}$) = $d^t - d^{t-1}$.

QED

6. The distribution of the test statistic

In Section 1, we give the formula of the test statistic $V = \sum_{|\alpha|=t} \frac{\left(N(\alpha) - nP(\alpha)\right)^2}{nP(\alpha)} - \sum_{|\alpha|=t-1} \frac{\left(N(\alpha) - nP(\alpha)\right)^2}{nP(\alpha)}$.

This formula is actually simplified from the mean-adjusted quadratic form of $N(\alpha)$s in the weak inverse $\overline{C}$. Assume that $N(\alpha)$s approximate the joint normal distribution when n is large. According to Theorem 2.2, V follows the chi-square distribution with degrees of freedom equal to the rank of the covariance matrix C. In Section 5, we have worked out the rank of C which is $d^t - d^{t-1}$. Thus, we conclude that V follows the chi-square distribution with $d^t - d^{t-1}$ degrees of freedom.

Recall from Section 3 that $N = \begin{bmatrix} N(\alpha_1) \\ N(\alpha_2) \\ \cdots \\ N(\alpha_{d^t}) \end{bmatrix}$ and its mean is $\mu = \begin{bmatrix} nP(\alpha_1) \\ nP(\alpha_2) \\ \cdots \\ nP(\alpha_{d^t}) \end{bmatrix}$, the following theorem verifies that $V = (N - \mu)^T \overline{C} (N - \mu)$.

Theorem 6.1. $(N - \mu)^T \overline{C} (N - \mu) = \sum_{|\alpha|=t} \frac{\left(N(\alpha) - nP(\alpha)\right)^2}{nP(\alpha)} - \sum_{|\alpha|=t-1} \frac{\left(N(\alpha) - nP(\alpha)\right)^2}{nP(\alpha)}$.

Proof:

$$
\begin{aligned}
&(N-\mu)^T \overline{C} (N-\mu) \\
&= \sum_{|\alpha|=t} \sum_{|\beta|=t} \left(N(\alpha) - nP(\alpha)\right) \overline{c}_{\alpha\beta} \left(N(\beta) - nP(\beta)\right) \\
&= \frac{1}{n} \sum_{|\alpha|=t} \sum_{|\beta|=t} \left(N(\alpha) - nP(\alpha)\right)\left(K(\alpha,\beta) - K(t-1:\alpha, t-1:\beta)\right)\left(N(\beta) - nP(\beta)\right) \\
&= \frac{1}{n} \sum_{|\alpha|=t} \sum_{|\beta|=t} \left(N(\alpha) - nP(\alpha)\right) K(\alpha,\beta) \left(N(\beta) - nP(\beta)\right) - \frac{1}{n} \sum_{|\alpha|=t} \sum_{|\beta|=t} \left(N(\alpha) - nP(\alpha)\right) K(t-1:\alpha, t-1:\beta) \left(N(\beta) - nP(\beta)\right) \\
&= \frac{1}{n} \sum_{|\alpha|=t} \left(N(\alpha) - nP(\alpha)\right)^2 K(\alpha,\alpha) - \frac{1}{n} \sum_{|\alpha|=t-1} \sum_{|\beta|=t-1} \sum_{0 \le a < d} \sum_{0 \le b < d} \left(N(\alpha a) - nP(\alpha a)\right) K(\alpha,\beta) \left(N(\beta b) - nP(\beta b)\right) \\
&= \sum_{|\alpha|=t} \frac{\left(N(\alpha) - nP(\alpha)\right)^2}{nP(\alpha)} - \frac{1}{n} \sum_{|\alpha|=t-1} \sum_{0 \le a < d} \sum_{0 \le b < d} \left(N(\alpha a) - nP(\alpha a)\right) K(\alpha,\alpha) \left(N(\alpha b) - nP(\alpha b)\right) \\
&= \sum_{|\alpha|=t} \frac{\left(N(\alpha) - nP(\alpha)\right)^2}{nP(\alpha)} - \sum_{|\alpha|=t-1} \frac{1}{nP(\alpha)} \left(\sum_{0 \le a < d} \left(N(\alpha a) - nP(\alpha a)\right) \right) \left(\sum_{0 \le b < d} \left(N(\alpha b) - nP(\alpha b)\right) \right) \\
&= \sum_{|\alpha|=t} \frac{\left(N(\alpha) - nP(\alpha)\right)^2}{nP(\alpha)} - \sum_{|\alpha|=t-1} \frac{1}{nP(\alpha)} \left(\sum_{0 \le a < d} \left(N(\alpha a) - nP(\alpha a)\right) \right)^2 \\
&= \sum_{|\alpha|=t} \frac{\left(N(\alpha) - nP(\alpha)\right)^2}{nP(\alpha)} - \sum_{|\alpha|=t-1} \frac{1}{nP(\alpha)} \left(N(\alpha) - nP(\alpha) \sum_{0 \le a < d} P(a) \right)^2 \\
&= \sum_{|\alpha|=t} \frac{\left(N(\alpha) - nP(\alpha)\right)^2}{nP(\alpha)} - \sum_{|\alpha|=t-1} \frac{\left(N(\alpha) - nP(\alpha)\right)^2}{nP(\alpha)}
\end{aligned}
$$

The second to the last step uses the equality $\sum_{0 \le a < d} N(\alpha a) = N(\alpha)$.

QED

7. Future work

One big assumption in the derivation of the distribution of V is that $N(\alpha)$s have the joint normal distribution when n is large. Note that $N(\alpha)$s are discrete variables and the normal is continuous. The distribution of $N(\alpha)$s is at best close to a joint normal but will not be identical. The closeness of the approximation is positively correlated to n. One problem is how large n is needed such that the test has sufficient accuracy. For the standard chi-square test, n is required to be large enough such that all expected frequencies are at least 5. We need a similar rule-of-thumb for this overlapping version. We plan to compute the exact distribution of V for a range of tests with small n, d and t. Then compare these exact distributions with the corresponding chi-square distributions. We anticipate to find a rule on the size of n or on the expected frequencies of $N(\alpha)$s such that the resulting test has sufficient accuracy in practice.

One optimization problem arising from the overlapping chi-square test is: given n outcomes of an experiment, what choice of t would yield the highest power, i.e., the highest probability of rejecting the null hypothesis when it is actually false. The answer to this problem is crucial in applications.

In addition to serving as a goodness-of-fit test, the overlapping chi-square test can be applied directly for examining the randomness of the outcomes from a random number generator (RNG) [MR85]. We intend to compare the power of the test with other stringent tests for RNGs, e.g., the monkey test and the birthday spacing test [MT02].

References

[GD53] Good, I.J., The serial test for sampling numbers and other tests for randomness. *Proc. Cambridge Philosophical Society,* **49**, 1953.

[KN98] Knuth, D. E., *The Art of Computer Programming*, Volume 2, 3rd ed., Addison-Wesley, 1998.

[MR85] Marsaglia, G, A current view of random number generators, Keynote Address, *Proc. Statistics and Computer Science: 16th Symposium on the Interface*, Atlanta, 1985.

[MT02] Marsaglia, G and Tsang, W. W.. Some difficult-to-pass tests of randomness. *Journal of Statistical Software*, **7**, Issue 3, 2002.

[MR05] Marsaglia, G, Monkeying with the Goodness-of-Fit Test. *Journal of Statistical Software*, **14**, Issue 13, 2005.

[PK00] Pearson, K., On the Criterion that a Given System of Deviations from the Probable in the Case of Correlated System of Variables is such that it can be Reasonably Supposed to have Arisen from Random Sampling, *Philosophical Magazine*, **50**, Issue 5, 157 – 175, 1900.

[WS96] Wegenkittl, S., *Empirical testing of Pseudorandom number generators*, Master thesis, Universität Salzburg, 1996.