Manuscript Number: JSPI-D-07-00153R1

Title: A-dependence statistics for mutual and serial independence of categorical variables

Article Type: Full-Length Article

Keywords: Categorical variables; chi-square tests; mutual independence; serial independence

Corresponding Author: Martin Bilodeau,

Corresponding Author's Institution: University of Montreal

First Author: Martin Bilodeau

Order of Authors: Martin Bilodeau; Pierre Lafaye de Micheaux, Ph.D.

Abstract: The M\"{o}bius transformation of probability cells in a multi-way contingency table is used to partition the Pearson chi-square test of mutual independence into A-de\-pen\-dence statistics. A similar partition is proposed for a universal and consistent test of serial independence in a stationary sequence of a categorical variable. The partition proposed can be adapted whether using estimated or theoretical marginal probabilities. With the aim of detecting a dependence of high order in a long sequence, A-de\-pen\-dence terms of the partition measuring increasing lagged dependences can be combined in a Box-Pierce type test of serial independence. A real data analysis of a nucleotides sequence using the Box-Pierce type test is provided.

Response to reviewers' comments

1. My paper treats only of categorical variables as the title indicates. The two references given by the reviewer are relevant to quantitative variables. For reasons of lack of invariance, categorical data, especially those of a nominal nature, should not be analyzed using quantitative methods. The test of Delgado (1996) is for continuous variables with critical points estimated by random permutation. A better approach is that of Genest and Rémillard (2004), also for continuous variables, with critical points which can be tabulated. The analysis of the DNA sequence with the method in Genest and Rémillard (2004) using the assignment (A=1, G=2, C=3, T=4) or (T=1, A=2, G=3, C=4) would give different values of the test statistic. Moreover, the tabulated critical values could not be used since the variables are discrete. Critical values could be estimated by the bootstrap. This is explained in the last paragraph of the introduction which now excludes notations and mathematical formulations.

2. A new Section 2 was added to introduce notations and the chi-square tests.

3. A non binary categorical time series is now considered in Section 7.2. Table 2 was modified accordingly. Simulations could not cover all cases of serial or non-serial, sample size, number of cells, and choice of dimension d. It is preferable to simulate the situation at hand for a given problem, as I have done in Section 7.4 prior to the data analysis in Section 8. A comment along these lines was added at the beginning of Section 7.

4. Comparison of the power of the test proposed in my paper to that of other tests existing in the litterature was not done for the reasons given in 1.

A-dependence statistics for mutual and serial independence of categorical variables

M. Bilodeau [a,*] and P. Lafaye de Micheaux [b]

[a] *Département de mathématiques et de statistique, université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Canada H3C 3J7*

[b] *Université Pierre Mendès France, Laboratoire de Statistique et Analyse de Données (LabSAD), BP 47 / F-38040 Grenoble Cedex 9, France*

Abstract

The Möbius transformation of probability cells in a multi-way contingency table is used to partition the Pearson chi-square test of mutual independence into A-dependence statistics. A similar partition is proposed for a universal and consistent test of serial independence in a stationary sequence of a categorical variable. The partition proposed can be adapted whether using estimated or theoretical marginal probabilities. With the aim of detecting a dependence of high order in a long sequence, A-dependence terms of the partition measuring increasing lagged dependences can be combined in a Box-Pierce type test of serial independence. A real data analysis of a nucleotides sequence using the Box-Pierce type test is provided.

Key words: Categorical variables, chi-square tests, mutual independence, serial independence
1991 MSC: 62H15, 62G09, 60F05

---

1 Introduction

When the Pearson chi-square test of mutual independence is significant, it gives little indication to the way that the null hypothesis disagrees with the data. This paper uses the Möbius transformation to partition Pearson chi-square into components measuring what was termed by Deheuvels (1979) as

* Corresponding author
Email addresses: bilodeau@dms.umontreal.ca (M. Bilodeau), Pierre.Lafaye-de-Micheaux@upmf-grenoble.fr (P. Lafaye de Micheaux).

the A-dependence. The A-dependence statistics of the partition are mutually independent and asymptotically distributed as chi-square. They are associated with additive interactions. The dependence structure of a multi-way contingency table can also be investigated by multiplicative interactions in a log-linear model, see Santner and Duffy (1989).

The main contribution of this paper relates to the use of A-dependence statistics to test for serial independence of a stationary sequence of a categorical variable. The adaptation of log-linear models to stationary sequences seems unwieldy, whereas the A-dependence approach extends very naturally. A universal and consistent test of serial independence is constructed using A-dependence statistics. This test is also a chi-square test and it can be partitioned into A-dependence statistics which are asymptotically independent and distributed as chi-square. Similar properties are obtained for the chi-square test to be used with theoretical marginal probabilities. A chi-square test of the Box-Pierce type is also proposed to detect large lagged dependences with an application to a nucleotides sequence. The A-dependence statistics have a closed form. They can be evaluated by the most common statistical softwares which provide the Pearson chi-square test of independence for multi-way contingency tables. The Möbius transformation sheds new light on the partition of the mutual independence chi-square test in Lancaster (1951) and the serial independence chi-square test in Good (1953).

Tests of serial independence for a stationary quantitative sequence are numerous. Delgado (1996) extends the work of Blum *et al.* (1961) in the higher dimensional case to obtain a test of serial independence in the continuous case. Genest and Rémillard (2004) propose another test based on the Möbius transformation with a simpler covariance function than that of Delgado (1996). Both of these tests are distribution free in the continuous case. They could also be used in the discrete case with the use of the bootstrap distribution even though they are no longer distribution free, see Beran *et al.* (2007). Hong (1998) obtains a test for pairwise serial independence applicable in discrete or continuous models with a standard normal approximation for large larged dependences. All of the above tests are based on the empirical distribution function and should not be applied to categorical time series, especially those of a nomimal (or non-ordinal) nature. Nominal categories are mere labels and their quantification can yield different orderings of the labels. The tests based on ranks, which are invariant to monotone transformations of the data, are not invariant to permutation of the labels. In a genetic application, the assignment of nucleotides (A=1, G=2, C=3, T=4) or (T=1, A=2, G=3, C=4) would give different values of the test statistic.

2 Chi-square tests

A d-dimensional categorical random vector is denoted $X = (X^{(1)}, \ldots, X^{(d)})$. As for the vector $t = (t_1, \ldots, t_d)$, it represents a d-dimensional cell in a multi-way contingency table. The joint and marginal cell probabilities are

$$\begin{aligned} p(t) &= \mathrm{pr}(X^{(1)} = t_1, \ldots, X^{(d)} = t_d), \\ p^{(k)}(t_k) &= \mathrm{pr}(X^{(k)} = t_k), \ k = 1, \ldots, d. \end{aligned}$$

The cell counts in the d-dimensional table are based on n independent realizations

$$X_i = (X_i^{(1)}, \ldots, X_i^{(d)}), \ i = 1, \ldots, n,$$

from the distribution of X. The sampling distribution of the counts is a single multinomial experiment, where only the total n is fixed. The joint and marginal empirical cell probabilities are

$$\begin{aligned} p_n(t) &= \frac{1}{n} \sum_{i=1}^{n} \prod_{k=1}^{d} \mathbb{I}\{X_i^{(k)} = t_k\}, \\ p_n^{(k)}(t_k) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i^{(k)} = t_k\}. \end{aligned}$$

Pearson chi-square

$$X^2_{1,\ldots,d} = \sum_t \frac{n[p_n(t) - \prod_{k=1}^d p_n^{(k)}(t_k)]^2}{\prod_{k=1}^d p_n^{(k)}(t_k)} \tag{1}$$

is often used to test the hypothesis of mutual independence among d categorical variables, where $d \geq 2$. Assuming there are I_k categories associated with the variable $X^{(k)}$, then the number of cells of the d-dimensional table is $\prod_{k=1}^d I_k$. The asymptotic null distribution of this test is chi-square with

$$f = \prod_{k=1}^{d} I_k - 1 - \sum_{k=1}^{d} (I_k - 1)$$

degrees of freedom.

Sometimes, the chi-square test is used with theoretical marginal probabilities in which case it is defined as

$$\breve{X}^2_{1,\ldots,d} = \sum_t \frac{n[p_n(t) - \prod_{k=1}^d p^{(k)}(t_k)]^2}{\prod_{k=1}^d p^{(k)}(t_k)} \tag{2}$$

and it has $f = \prod_{k=1}^d I_k - 1$ degrees of freedom.

3 A-dependence

Let $\mathcal{I}_d$ be the family of all subsets A of $\{1, \ldots, d\}$ of size $|A| > 1$. There are $2^d - d - 1$ such subsets in $\mathcal{I}_d$. For a subset $B \in \mathcal{I}_d$, let $t^{(B)} = (t_k)_{k \in B}$ be a sub-cell of dimensionality $|B|$, and let its probability be

$$p^{(B)}(t^{(B)}) = \mathrm{pr}(\cap_{k \in B}\{X^{(k)} = t_k\}).$$

For all subsets B, the empirical sub-cell probability is

$$p_n^{(B)}(t^{(B)}) = \frac{1}{n}\sum_{i=1}^n \prod_{k \in B} \mathbb{I}\{X_i^{(k)} = t_k\}.$$

The Möbius transformation of cell probabilities is used to characterize mutual independence. The variables $X^{(1)}, \ldots, X^{(d)}$ are mutually independent if and only if $\mu_A(t^{(A)}) = 0$, for all $A \in \mathcal{I}_d$ and all sub-cells $t^{(A)}$, where

$$\mu_A(t^{(A)}) = \sum_{B \subset A} (-1)^{|A \setminus B|} p^{(B)}(t^{(B)}) \prod_{k \in A \setminus B} p^{(k)}(t_k), \tag{3}$$

with the usual convention $\prod_{k \in \emptyset} = 1$. The quantities μ_A are additive interactions. The inverse transformation yields the cell probabilities

$$p(t) = \prod_{k=1}^d p^{(k)}(t_k) + \sum_{A \in \mathcal{I}_d} \mu_A(t^{(A)}) \prod_{k \notin A} p^{(k)}(t_k).$$

The following expressions of additive interactions are given for subsets A of size two and three

$$\begin{aligned}
\mu_{1,2}(t_1, t_2) &= p^{(1,2)}(t_1, t_2) - p^{(1)}(t_1)p^{(2)}(t_2), \\
\mu_{1,2,3}(t_1, t_2, t_3) &= p^{(1,2,3)}(t_1, t_2, t_3) - p^{(1,2)}(t_1, t_2)p^{(3)}(t_3) - p^{(1,3)}(t_1, t_3)p^{(2)}(t_2) \\
&\quad - p^{(2,3)}(t_2, t_3)p^{(1)}(t_1) + 2p^{(1)}(t_1)p^{(2)}(t_2)p^{(3)}(t_3).
\end{aligned}$$

In a contingency table of dimension three, the hypothesis of no three-variable additive interaction was first considered by Lancaster (1951). Lancaster (1969) obtained different expressions for this hypothesis. Among them, the expression (7) in Darroch (1974), in a comparison of additive and multiplicative interactions, corresponds to the hypothesis $\mu_A(t^{(A)}) = 0$ for $A = \{1, 2, 3\}$. Instead of additive interactions, log-linear models use multiplicative interactions for the analysis of contingency tables.

Similar characterizations to (3), based on the distribution function or characteristic function, for numerical variables are those in Ghoudi *et al.* (2001) or Bilodeau and Lafaye de Micheaux (2005). In a non-parametric test of mutual independence among numerical vectors, Beran *et al.* (2007) introduced a variant of this characterization based on half-spaces. All these characterizations were inspired by the work of Deheuvels (1981) who himself drew on the seminal work of Blum *et al.* (1961). Unbeknownst to the authors of these recent papers, the Möbius transformation goes back even further as it is implicit in the work of Lancaster (1951) on contingency tables.

The A-dependence statistics proposed here are functions of the variables

$$R_{n,A}(t^{(A)}) = \sqrt{n} \sum_{B \subset A} (-1)^{|A \setminus B|} p_n^{(B)}(t^{(B)}) \prod_{k \in A \setminus B} p_n^{(k)}(t_k).$$

Explicit expressions for the variables $R_{n,A}(t^{(A)})/\sqrt{n}$ are obtained by replacing the theoretical probabilities by the empirical ones in the expressions above for $\mu_A(t^{(A)})$. For a given subset A, the A-dependence statistic is

$$T_A = \sum_{t^{(A)}} \frac{[R_{n,A}(t^{(A)})]^2}{\prod_{k \in A} p_n^{(k)}(t_k)}.$$

The multinomial formula

$$\sum_{B \subset A} \prod_{k \in B} u^{(k)} \cdot \prod_{k \in A \setminus B} v^{(k)} = \prod_{k \in A} \left(u^{(k)} + v^{(k)} \right)$$

gives the following representation

$$R_{n,A}(t^{(A)}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \prod_{k \in A} \left[\mathbb{I}\{X_i^{(k)} = t_k\} - p_n^{(k)}(t_k) \right].$$

Another useful set of variables is

$$\breve{R}_{n,A}(t^{(A)}) = \sqrt{n} \sum_{B \subset A} (-1)^{|A \setminus B|} p_n^{(B)}(t^{(B)}) \prod_{k \in A \setminus B} p^{(k)}(t_k)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \prod_{k \in A} \left[\mathbb{I}\{X_i^{(k)} = t_k\} - p^{(k)}(t_k) \right].$$

The associated A-dependence statistic becomes

$$\breve{T}_A = \sum_{t^{(A)}} \frac{[\breve{R}_{n,A}(t^{(A)})]^2}{\prod_{k \in A} p^{(k)}(t_k)}.$$

The terms $\breve{T}_A$ can also be defined for singletons $A = \{k\}$ as

$$\breve{T}_k = \sum_{t_k} \frac{n[p_n^{(k)}(t_k) - p^{(k)}(t_k)]^2}{p^{(k)}(t_k)}, \ k = 1, \dots, d,$$

and they are the Pearson goodness-of-fit tests on each variable.

It may be noted that both $R_{n,A}(t^{(A)})/\sqrt{n}$ and $\breve{R}_{n,A}(t^{(A)})/\sqrt{n}$ converge with probability one to $\mu_A(t^{(A)})$ and $E[\breve{R}_{n,A}(t^{(A)})/\sqrt{n}] = \mu_A(t^{(A)})$. Also, the variables $\breve{R}_{n,A}(t^{(A)})$ are sums of independent and identically distributed terms.

4 Partition of Pearson chi-square

The A-dependence statistics provide a partition of Pearson chi-square. This partition is orthogonal in the same way as the sums of squares for factors and interactions in a balanced ANOVA model.

4.1 Estimated marginal probabilities

The inverse transformation gives

$$\sqrt{n}[p_n(t) - \prod_{k=1}^{d} p_n^{(k)}(t_k)] = \sum_{A \in \mathcal{I}_d} R_{n,A}(t^{(A)}) \prod_{k \notin A} p_n^{(k)}(t_k).$$

This yields

$$X^2_{1,\dots,d} = \sum_{A \in \mathcal{I}_d} \sum_{B \in \mathcal{I}_d} \sum_{t} R_{n,A}(t^{(A)}) R_{n,B}(t^{(B)}) \frac{\prod_{k \notin A} p_n^{(k)}(t_k) \prod_{k \notin B} p_n^{(k)}(t_k)}{\prod_{k=1}^{d} p_n^{(k)}(t_k)}.$$

The term corresponding to the case $A = B$ is

$$
\begin{aligned}
\sum_t [R_{n,A}(t^{(A)})]^2 \frac{\prod_{k\notin A}[p_n^{(k)}(t_k)]^2}{\prod_{k=1}^d p_n^{(k)}(t_k)} &= \sum_{t^{(A)}} \frac{[R_{n,A}(t^{(A)})]^2}{\prod_{k\in A} p_n^{(k)}(t_k)} \sum_{t^{(A^c)}} \prod_{k\notin A} p_n^{(k)}(t_k) \\
&= \sum_{t^{(A)}} \frac{[R_{n,A}(t^{(A)})]^2}{\prod_{k\in A} p_n^{(k)}(t_k)} \cdot 1 \\
&= T_A.
\end{aligned}
$$

Before considering terms such that $A \neq B$, observe that for an index $j \in A$,

$$
\begin{aligned}
\sum_{t_j} R_{n,A}(t^{(A)}) = &\frac{1}{\sqrt{n}} \sum_{i=1}^n \prod_{k\in A\setminus\{j\}} \left[\mathbb{I}\{X_i^{(k)} = t_k\} - p_n^{(k)}(t_k)\right] \\
&\cdot \sum_{t_j} \left[\mathbb{I}\{X_i^{(j)} = t_j\} - p_n^{(j)}(t_j)\right] \\
= &\frac{1}{\sqrt{n}} \sum_{i=1}^n \prod_{k\in A\setminus\{j\}} \left[\mathbb{I}\{X_i^{(k)} = t_k\} - p_n^{(k)}(t_k)\right] (1-1) \\
= &\, 0.
\end{aligned}
$$

For terms such that $A \neq B$, there is an index j in A but not in B (or the converse) for which

$$
\begin{aligned}
&\sum_t R_{n,A}(t^{(A)}) R_{n,B}(t^{(B)}) \frac{\prod_{k\notin A} p_n^{(k)}(t_k) \prod_{k\notin B} p_n^{(k)}(t_k)}{\prod_{k=1}^d p_n^{(k)}(t_k)} \\
= &\sum_{t^{(\{j\}^c)}} R_{n,B}(t^{(B)}) \frac{\prod_{k\notin A} p_n^{(k)}(t_k)}{\prod_{k\in B} p_n^{(k)}(t_k)} \sum_{t_j} R_{n,A}(t^{(A)}) \\
= &\, 0.
\end{aligned}
$$

Thus, the partition

$$
X^2_{1,\ldots,d} = \sum_{A\in\mathcal{I}_d} T_A
$$

is orthogonal. As a matter of computations, let

$$
X_A^2 = \sum_{t^{(A)}} \frac{n[p_n^{(A)}(t^{(A)}) - \prod_{k\in A} p_n^{(k)}(t_k)]^2}{\prod_{k\in A} p_n^{(k)}(t_k)} \tag{4}
$$

be the usual Pearson chi-square computed on the sub-table obtained by the selection of the variables in A. One computes X_A^2 for all $A \in \mathcal{I}_d$. Then, all the terms in the partition are given recursively as in Lancaster (1951) by

$$
\begin{aligned}
&T_A = X_A^2, && \text{if } |A| = 2;\\
&T_A = X_A^2 - \textstyle\sum_{\{B\subset A: 1<|B|<|A|\}} T_B, && \text{if } |A| > 2.
\end{aligned}
$$

4.2 Theoretical marginal probabilities

A similar partition holds for theoretical marginal probabilities. Lancaster (1951) gives a detailed analysis for tables of dimension three. For all dimensions,

$$\breve{X}^2_{1,\dots,d} = \sum_t \frac{n[p_n(t) - \prod_{k=1}^d p^{(k)}(t_k)]^2}{\prod_{k=1}^d p^{(k)}(t_k)} \tag{5}$$

has the following orthogonal partition

$$\breve{X}^2_{1,\dots,d} = \sum_{A\in\breve{\mathcal{I}}_d} \breve{T}_A,$$

where $\breve{\mathcal{I}}_d$ is the family of all subsets A of $\{1,\dots,d\}$ of size $|A| \geq 1$.

5 Distributions of T_A and $\breve{T}_A$

The asymptotic distribution is now derived under the hypothesis of mutual independence. The central limit theorem applied to $\breve{R}_{n,A}$ gives the following result. The limiting distribution of the variables $\breve{R}_{n,A}(t^{(A)})$, where the index $t^{(A)}$ is allowed to vary over all sub-cells, is gaussian with covariance given by

$$\operatorname{cov}\left(\breve{R}_{n,A}(s^{(A)}), \breve{R}_{n,A}(t^{(A)})\right) = \prod_{k\in A}\left[\mathbb{I}\{s_k = t_k\}p^{(k)}(s_k) - p^{(k)}(s_k)p^{(k)}(t_k)\right].$$

Moreover, since

$$E\left\{\prod_{k\in A}\left[\mathbb{I}\{X^{(k)} = s_k\} - p^{(k)}(s_k)\right]\prod_{k\in B}\left[\mathbb{I}\{X^{(k)} = t_k\} - p^{(k)}(t_k)\right]\right\} = 0,$$

for $A \neq B$, then $\breve{R}_{n,A}(s^{(A)})$ and $\breve{R}_{n,B}(t^{(B)})$ are asymptotically independent for all sub-cells $s^{(A)}$ and $t^{(B)}$.

The second result claims that the variables $R_{n,A}(t^{(A)})$ and $\breve{R}_{n,A}(t^{(A)})$ are asymptotically equivalent. This comes from the representation

$$R_{n,A}(t^{(A)}) - \breve{R}_{n,A}(t^{(A)}) = \sum_{B\subset A, B\neq\emptyset} (-1)^{|B|} \prod_{k\in B} \left[p_n^{(k)}(t_k) - p^{(k)}(t_k)\right] \cdot \breve{R}_{n,A\setminus B}(t^{(A\setminus B)})$$

which is a consequence of the identities

$$\begin{aligned} R_{n,A}(t^{(A)}) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \prod_{k\in A} \left[\left(\mathbb{I}\{X_i^{(k)} = t_k\} - p^{(k)}(t_k)\right) - \left(p_n^{(k)}(t_k) - p^{(k)}(t_k)\right)\right] \\ &= \sum_{B\subset A} (-1)^{|B|} \prod_{k\in B} \left[p_n^{(k)}(t_k) - p^{(k)}(t_k)\right] \\ &\quad \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \prod_{k\in A\setminus B} \left[\mathbb{I}\{X_i^{(k)} = t_k\} - p^{(k)}(t_k)\right] \\ &= \sum_{B\subset A} (-1)^{|B|} \prod_{k\in B} \left[p_n^{(k)}(t_k) - p^{(k)}(t_k)\right] \breve{R}_{n,A\setminus B}(t^{(A\setminus B)}). \end{aligned}$$

The term corresponding to $B = \emptyset$ in the last sum is $\breve{R}_{n,A}(t^{(A)})$. Thus, the limiting distribution of T_A is the same as that of $\breve{T}_A$. This limiting distribution is now derived. Let $k = |A|$. The product structure of the covariance function coupled with the quadratic nature of $\breve{T}_A$ imply that $\breve{T}_A$ is distributed asymptotically as

$$\sum_{j_1,\ldots,j_k} \lambda_{j_1,\ldots,j_k} Z_{j_1,\ldots,j_k}^2,$$

where $Z_{j_1,\ldots,j_k}$ are independent $N(0,1)$ variables. The eigenvalues $\lambda_{j_1,\ldots,j_k}$ and eigenfunctions $f_{j_1,\ldots,j_k}$ have a product form

$$\begin{aligned} \lambda_{j_1,\ldots,j_k} &= \prod_{l=1}^{k} \lambda_{j_l}, \\ f_{j_1,\ldots,j_k}(t) &= \prod_{l=1}^{k} f_{j_l}(t_{j_l}). \end{aligned}$$

The eigenvalue λ_{j_l} and associated function f_{j_l} are solutions of the equation

$$\frac{1}{p^{(j_l)}(s_{j_l})} \sum_{t_{j_l}} \left[\mathbb{I}\{s_{j_l} = t_{j_l}\} p^{(j_l)}(s_{j_l}) - p^{(j_l)}(s_{j_l}) p^{(j_l)}(t_{j_l})\right] f_{j_l}(t_{j_l}) = \lambda_{j_l} \cdot f_{j_l}(s_{j_l}).$$

It is readily checked that the value $\lambda_{j_l} = 0$ is an eigenvalue of multiplicity 1 with eigenfunction $f_{j_l}(t_{j_l}) \equiv 1$. The value $\lambda_{j_l} = 1$ is the other eigenvalue of multiplicity $I_{j_l} - 1$ with associated eigenfunctions satisfying

$$\sum_{t_{j_l}} p^{(j_l)}(t_{j_l}) f_{j_l}(t_{j_l}) = 0$$

which is the equation of a hyperplane in the euclidian space of dimension I_{j_l}. Since the only eigenvalues are 0 and 1, it follows that the limiting distribution of $\breve{T}_A$ is chi-square with

$$f_A = \prod_{k \in A} (I_k - 1)$$

degrees of freedom. The A-dependence statistics T_A are asymptotically distributed as independent chi-square variables and provide a partition of Pearson chi-square into all possible dependences of any order. The subsets A for which the A-dependence chi-square have a small p-value,

$$\mathrm{pr}\left(\chi^2_{f_A} > T_A\right),$$

can be flagged as the possible causes of dependence when the test $X^2_{1,\ldots,d}$ is significant. The total degrees of freedom is as it should

$$f = \sum_{A \in \mathcal{I}_d} \prod_{k \in A} (I_k - 1) = \prod_{k=1}^{d} I_k - 1 - \sum_{k=1}^{d} (I_k - 1).$$

In the partition of Pearson chi-square $\breve{X}^2_{1,\ldots,d}$ with theoretical marginal probabilities, the total degrees of freedom becomes

$$f = \sum_{A \in \breve{\mathcal{I}}_d} \prod_{k \in A} (I_k - 1) = \prod_{k=1}^{d} I_k - 1.$$

In this case, other possible sources for the significance of $\breve{X}^2_{1,\ldots,d}$ are the goodness-of-fit statistics $\breve{T}_k$ on anyone of the variables as measured by the p-values

$$\mathrm{pr}\left(\chi^2_{I_k - 1} > \breve{T}_k\right), \; k = 1, \ldots, d.$$

6 Test of serial independence

The A-dependence methodology is adapted to test for serial independence of a stationary time series Y_i, $i = 1, \dots, n$, where Y_i is a categorical variable taking I possible values. For a fixed value of d, consider the d-dimensional vectors formed of successive observations

$$X_i = (X_i^{(1)}, \dots, X_i^{(d)}) = (Y_i, \dots, Y_{i+d-1}),\ i = 1, \dots, n-d+1,$$

where $X_i^{(k)} = Y_{i+k-1}$. In this context, for $u = 1, \dots, I$ and $k = 1, \dots, d$, let

$$\begin{aligned} q(u) &= \mathrm{pr}(Y_1 = u), \\ p_n^{(k)}(u) &= \frac{1}{n-d+1} \sum_{i=1}^{n-d+1} \mathbb{I}\{Y_{i+k-1} = u\}. \end{aligned}$$

Let $\mathcal{A}_d \subset \mathcal{I}_d$ be the family of subsets $A \in \mathcal{I}_d$ such that $1 \in A$. There are $2^{d-1} - 1$ subsets in $\mathcal{A}_d$. For any subset $B \in \mathcal{A}_d$, define

$$p_n^{(B)}(t^{(B)}) = \frac{1}{n-d+1} \sum_{i=1}^{n-d+1} \prod_{k \in B} \mathbb{I}\{Y_{i+k-1} = t_k\}.$$

The restriction to subsets of $\mathcal{A}_d$ is used since a subset B and its translate $B + j$ lead essentially to the same sub-cell empirical probabilities. The ergodic theorem implies that

$$\begin{aligned} \mathrm{pr}\left[\lim_{n\to\infty} p_n^{(k)}(u) = q(u)\right] &= 1,\ k = 1, \dots, d, \\ \mathrm{pr}\left[\lim_{n\to\infty} p_n^{(B)}(t^{(B)}) = p^{(B)}(t^{(B)})\right] &= 1, \end{aligned}$$

where

$$p^{(B)}(t^{(B)}) = \mathrm{pr}(\cap_{k\in B}\{Y_k = t_k\}).$$

The variables resulting from the Möbius transformation are now

$$\begin{aligned} S_{n,A}(t^{(A)}) &= \frac{1}{\sqrt{n-d+1}} \sum_{i=1}^{n-d+1} \prod_{k\in A} \left[\mathbb{I}\{X_i^{(k)} = t_k\} - p_n^{(k)}(t_k)\right], \\ \breve{S}_{n,A}(t^{(A)}) &= \frac{1}{\sqrt{n-d+1}} \sum_{i=1}^{n-d+1} \prod_{k\in A} \left[\mathbb{I}\{X_i^{(k)} = t_k\} - q(t_k)\right]. \end{aligned}$$

It may be noted that both $S_{n,A}(t^{(A)})/\sqrt{n}$ and $\breve{S}_{n,A}(t^{(A)})/\sqrt{n}$ converge with probability one to

$$\nu_A(t^{(A)}) = \sum_{B \subset A} (-1)^{|A \setminus B|} p^{(B)}(t^{(B)}) \prod_{k \in A \setminus B} q(t_k)$$

and $E[\breve{S}_{n,A}(t^{(A)})/\sqrt{n}] = \nu_A(t^{(A)})$. The variable $\breve{S}_{n,A}(t^{(A)})$ is a sum of identically distributed terms forming an m-dependent sequence, where $m = d - 1$. The central limit theorem for m-dependent sequences, see Ferguson (1996, p. 69), can be used to establish, under the hypothesis of serial independence, that $\breve{S}_{n,A}(s^{(A)})$ and $\breve{S}_{n,B}(t^{(B)})$ are asymptotically and jointly normal with means 0 and asymptotic covariance $\sigma_{00} + 2\sigma_{01} + \cdots + 2\sigma_{0m}$, where

$$\sigma_{0u} = E\left\{\prod_{k \in A} \left[\mathbb{I}\{X_i^{(k)} = s_k\} - q(s_k)\right] \prod_{k \in B} \left[\mathbb{I}\{X_{i+u}^{(k)} = t_k\} - q(t_k)\right]\right\}.$$

Similar arguments as for $\breve{R}_{n,A}(t^{(A)})$ show that all σ_{0u} are null except when $A = B$ (both in $\mathcal{A}_d$) and $u = 0$. Thus, for $A \neq B$, $\breve{S}_{n,A}(s^{(A)})$ and $\breve{S}_{n,B}(t^{(B)})$ are asymptotically independent as in Section 5. Moreover, the limiting distribution of the variables $\breve{S}_{n,A}(t^{(A)})$, where the index $t^{(A)}$ is allowed to vary over all subcells, is gaussian with covariance given by

$$\operatorname{cov}\left(\breve{S}_{n,A}(s^{(A)}), \breve{S}_{n,A}(t^{(A)})\right) = \prod_{k \in A} \left[\mathbb{I}\{s_k = t_k\} q(s_k) - q(s_k) q(t_k)\right].$$

This same behavior holds for $S_{n,A}(t^{(A)})$ since

$$\begin{aligned} S_{n,A}(t^{(A)}) - \breve{S}_{n,A}(t^{(A)}) = & \sum_{B \subset A, B \neq \emptyset} (-1)^{|B|} \prod_{k \in B} \left[p_n^{(k)}(t_k) - q(t_k)\right] \\ & \cdot \breve{S}_{n,A \setminus B}(t^{(A \setminus B)}) \end{aligned}$$

which goes to 0 in probability as $n \to \infty$. The A-dependence statistics proposed for stationary sequences are

$$T_A = \sum_{t^{(A)}} \frac{[S_{n,A}(t^{(A)})]^2}{\prod_{k \in A} p_n^{(k)}(t_k)}, \quad A \in \mathcal{A}_d$$

which have the same asymptotic distribution as

$$\breve{T}_A = \sum_{t^{(A)}} \frac{[\breve{S}_{n,A}(t^{(A)})]^2}{\prod_{k \in A} q(t_k)}.$$

Again, as in Section 5, $\breve{T}_A$ is asymptotically distributed as chi-square with $g_A = (I-1)^{|A|}$ degrees of freedom. The analogue of Pearson chi-square for stationary sequences becomes

$$Y^2_{1,\ldots,d} = \sum_{A \in \mathcal{A}_d} T_A$$

which is also asymptotically chi-square with

$$g = \sum_{A \in \mathcal{A}_d} g_A = \sum_{k=2}^{d} \binom{d-1}{k-1} (I-1)^k = (I-1)(I^{d-1}-1)$$

total degrees of freedom.

The components T_A can be easily computed as in Section 4 with a statistical software capable of evaluating the usual Pearson chi-square for independence of a multi-way table. If only $Y^2_{1,\ldots,d}$ is needed, it is even easier. One computes X^2_A only for $A = \{1, \ldots, d\}$ and $A = \{2, \ldots, d\}$. Then,

$$Y^2_{1,\ldots,d} = \sum_{A \in \mathcal{A}_d} T_A = \sum_{A \in \mathcal{I}_d} T_A - \sum_{A \subset \{2,\ldots,d\}, |A| \geq 2} T_A = X^2_{1,\ldots,d} - X^2_{2,\ldots,d}.$$

With the objective of detecting large lagged dependences in a long sequence, such as a nucleotides sequence, the use of increasingly large values of d will yield a high-dimensional sparse contingency table. It is proposed, for such applications, to consider only subsets A of small sizes as in $\sum_{j=1}^{d} T_{1,j+1}$ which is asymptotically chi-square with $d(I-1)^2$ degrees of freedom. The term $T_{1,j+1}$ plays the role of an autocorrelation at lag j as in Box and Pierce (1970).

One can also conduct a test of serial independence with theoretical cell probabilities with the statistic

$$\breve{Y}^2_{1,\ldots,d} = \breve{T}_1 + \sum_{A \in \mathcal{A}_d} \breve{T}_A,$$

where, in the serial context,

$$\breve{T}_1 = \sum_{u=1}^{I} \frac{n[p_n^{(1)}(u) - q(u)]^2}{q(u)}.$$

The test $\breve{Y}^2_{1,\ldots,d}$ is asymptotically chi-square with

$$g = (I-1) + (I-1)(I^{d-1}-1) = I^d - I^{d-1}$$

degrees of freedom. Again, it can be computed easily by evaluating $\breve{X}_A$ for the same two subsets A as above

$$\breve{Y}^2_{1,\ldots,d} = \sum_{A\subset\{1,\ldots,d\},|A|\geq 1} \breve{T}_A - \sum_{A\subset\{2,\ldots,d\},|A|\geq 1} \breve{T}_A = \breve{X}^2_{1,\ldots,d} - \breve{X}^2_{2,\ldots,d}.$$

The test statistic $\breve{Y}^2_{1,\ldots,d}$ can actually be found in Good (1953) (his notation is $\nabla\psi^2_d = \psi^2_d - \psi^2_{d-1}$) for testing the randomness of a random digit generator. The proof in Good (1953) assumes that I is a prime number as it uses a partition derived from the discrete Fourier transform. His work corrected the belief that $\breve{X}^2_{1,2}$ is distributed as $\chi^2_{I^2-I}$, as it was advanced by Kendall and Smith Babington (1938, 1939), and later by Bartlett (1951). In fact, $\breve{X}^2_{1,2} = \breve{T}_1 + \breve{T}_2 + \breve{T}_{1,2}$, which is essentially $2\breve{T}_1 + \breve{T}_{1,2}$, is distributed as a linear combination of chi-square variables. In Kendall and Smith Babington (1938, pp. 158-159), the wrongly reported p-value for the serial test $\breve{X}^2_{1,2} = 110.44$ on 90 degrees of freedom is 0.07. The test $\breve{Y}^2_{12} = \breve{T}_1 + \breve{T}_{1,2}$ can be obtained by subtracting their frequency test, i.e. $\breve{T}_2 = 5.76$, from 110.44. This yields $\breve{Y}^2_{12} = 104.68$ on 90 degrees of freedom for a corrected p-value of 0.14. This latter p-value is more favorable to the randomness of the random digit generator tested by these authors.

7 Simulation

Some experiments were conducted to verify the convergence of A-dependence statistics to chi-square. The simulation of all situations covering the non-serial and serial cases, sampling size n, number of cells I and dimension d would have to be very extensive. Instead, it is preferable to simulate the situation at hand as in Section 7.4 before the real data analysis in Section 8.

7.1 Null distribution when testing for mutual independence

The first simulation consisted of 10000 replications of a sample of size $n = 100$ from the distribution of three independent bernoulli variables with probability 0.3 of success. Figure 1 shows the agreement between the histogram of $T_{1,2,3}$ and the χ^2_1 density. Similar histograms were obtained but are not reported for the other A-dependence statistics. The correlations between the components of the vector $(T_{1,2}, T_{1,3}, T_{2,3}, T_{1,2,3})$ were computed based on the 10000 simulated values. All the estimated correlations are close to zero; they are given in the

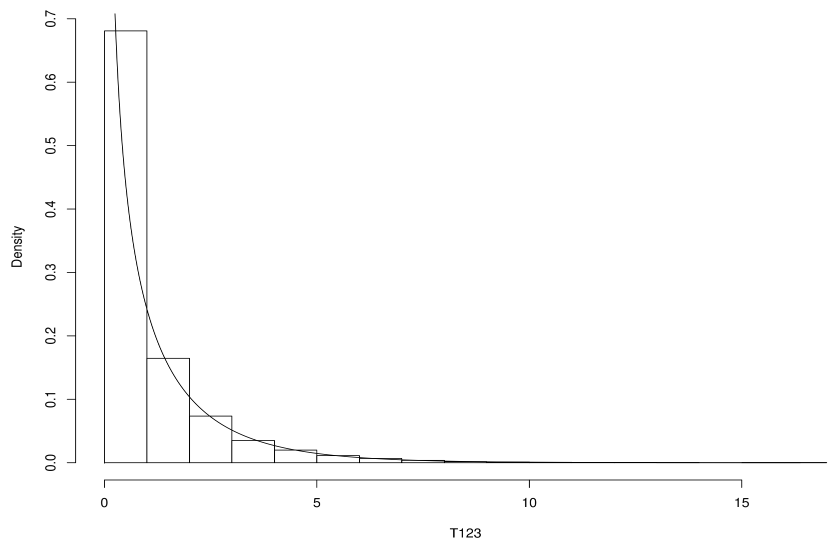


Fig. 1. Histogram of 10000 values of the A-dependence statistic $T_{1,2,3}$ evaluated from a sample of size $n = 100$ and the χ_1^2 density.

following correlation matrix

$$\begin{pmatrix} 1 & 0.002 & 0.010 & -0.011 \\ 0.002 & 1 & 0.009 & -0.015 \\ 0.010 & 0.009 & 1 & -0.017 \\ -0.011 & -0.015 & -0.017 & 1 \end{pmatrix}.$$

Table 1 gives the empirical quantiles based on the 10000 values, as well as the quantiles of the chi-square distributions. The agreement is satisfying for $d = 3$ and $n = 100$.

(TABLE 1 GOES HERE)

A second simulation investigated the convergence of $Y^2_{1,\dots,d}$ to chi-square for the test of serial independence with $d = 3$. This simulation consisted of 10000 replications of the test $Y^2_{1,2,3}$ for a sequence of length $n = 500$ made up of independent variables distributed on I=4 cells with associated probabilities $(.1, .2, .6, .1)$. The length required for the chi-squared approximation depends largely on the dimension d, the number of cells I, and the probabilities of the cells. The length required is large when either d or I is large, or when some cell probabilities are small. In this example with $n = 500$, the three dimensional cell with the smallest probability has an expected frequency of only 0.498, *i.e.* $.1^3 \times 498$. Figure 2 shows the agreement between the histogram and the chi-square density with 45 degrees of freedom. The correlation matrix of the

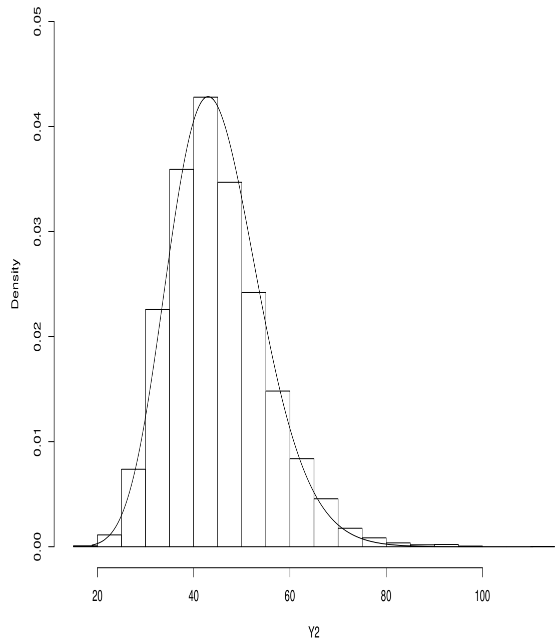


Fig. 2. Histogram of 10000 values of $Y^2_{1,2,3}$ evaluated from a sequence of length $n = 500$ and the χ^2_{45} density.

vector $(T_{1,2}, T_{1,3}, T_{1,2,3})$ is given by

$$\begin{pmatrix} 1 & 0.002 & 0.033 \\ 0.002 & 1 & 0.013 \\ 0.033 & 0.013 & 1 \end{pmatrix}.$$

Table 2 reports the empirical quantiles and the corresponding chi-square quantiles for each A-dependence statistics T_A.

(TABLE 2 GOES HERE)

7.3 Non-null distribution when testing for mutual independence

Define the uniform variable W on the set $\{1, 2, 3, 4, 5, 6, 7, 8\}$. The four-dimensional vector X is built from W through the indicator functions

$$X^{(1)} = \mathbb{I}\{W \in \{1, 2, 3, 5\}\},\ X^{(2)} = \mathbb{I}\{W \in \{1, 2, 4, 6\}\}$$
$$X^{(3)} = \mathbb{I}\{W \in \{1, 3, 4, 7\}\},\ X^{(4)} = \mathbb{I}\{W \in \{2, 3, 4, 8\}\}.$$

These four dependent binary variables are two-independent or pairwise independent; they are also three-independent. This is a case where anyone of the four variables is a deterministic function of the other three variables; for example, $X^{(4)} = \mathbb{I}\left\{X^{(1)} + X^{(2)} + X^{(3)} \in \{0, 2\}\right\}$. The simulation computes B=10000 replications of a sample of size $n = 100$ from the distribution above on X. For each subset A, let

$$\gamma = \mathrm{pr}\left[T_A > \chi^2_1(0.95)\right]$$

and

$$\hat{\gamma} = \frac{1}{B}\sum_{j=1}^{B} \mathbb{I}\left\{T_{A,j} > \chi^2_1(0.95)\right\},$$

where $T_{A,j}$ denotes the value of T_A for the replicate $j = 1, \ldots, B$.

(TABLE 3 GOES HERE)

The values of $\hat{\gamma}$ for subsets A such that $|A| \leq 3$ are close to 0.05 reflecting the fact that all subsets of three variables are mutually independent. However, the value $\hat{\gamma} = 1$ for $|A| = 4$ means that $T_{1,2,3,4}$ always detects the fourth order dependence.

7.4 Null distribution of the Box-Pierce type statistic

An independent binary sequence of length $n = 4156$ is generated such that $\text{pr}(Y = 1) = 0.44$. The statistics $T_{1,j+1}$, $j = 1, \ldots, d$, where $d = 3000$, were computed together with the number of these values greater than the 0.995 quantile of the χ_1^2 distribution:

$$\hat{\gamma} = \sum_{j=1}^{d} \mathbb{I}\{T_{1,j+1} > \chi_1^2(0.995)\}.$$

The expected value of $\hat{\gamma}$ should be approximately $.005d = 15$. The experiment just described was replicated 100 times. The 100 values of $\hat{\gamma}$ varied between 5 and 25 with an observed mean of 14.71.

8 Long nucleotides sequence

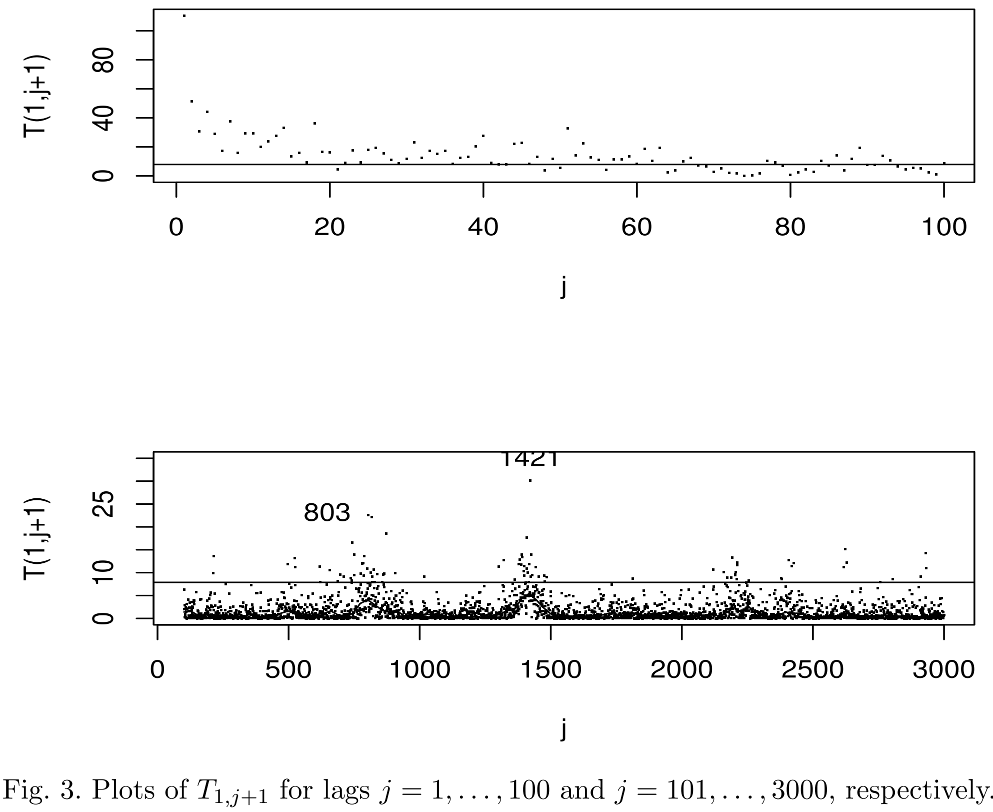


Fig. 3. Plots of $T_{1,j+1}$ for lags $j = 1, \ldots, 100$ and $j = 101, \ldots, 3000$, respectively.

The data in Whisenant *et al.* (1991) is a nucleotides sequence of 4156 base pairs (bp). A statistical analysis of this data set using the spectral envelope of a categorical time series can be found in Stoffer *et al.* (1993). The categorical variable represents the nucleotides, the purines, adenine (a) and guanine (g), the pyrimidines, cytosine (c) and thymine (t). A binary ($I = 2$) sequence is formed according to whether each bp is a purine (r) or a pyrimidine (y). The detection of long range dependences of such long sequences requires a large value of d for which the number of subsets becomes intractable. In that case the serial test of independence $Y^2_{1,\ldots,d}$ can be simplified to test only for (additive) interactions of low orders. For the statistical analysis of this sequence one could use a large value of d, for example $d = 3000$, in the Box-Pierce type statistic $\sum_{j=1}^{d} T_{1,j+1}$ which includes only the pairwise interactions at lags $j = 1, \ldots, d$. The fundamental fact remains: if the sequence is serially independent then, the statistics $T_{1,j+1}$ are asymptotically independent and distributed as $\chi^2_{(I-1)^2}$. One can plot $T_{1,j+1}$, $j = 1, \ldots, d$, with a threshold line corresponding to the 0.995 (say) quantile of the χ^2_1 distribution.

The observed number of points above the threshold value $\chi^2_1(0.995)$ is 157 whereas the expected number should be around 15. The first plot in Figure 3 reveals many significant lags between 1 and 60. The second plot reveals a bump around $j = 1421$. In terms of the frequency domain, this value corresponds to a frequency of approximately .0007 cycle per bp. Interestingly, the spectral envelope in Figure 1 of Stoffer *et al.* (1993) reveals a major peak at this same frequency.

Acknowledgments

We thank D.S. Stoffer and D.E. Tyler for sharing with us their data file of the nucleotides sequence in Whisenant *et al.* (1991). M. Bilodeau acknowledges the financial support provided by the Natural Sciences and Engineering Research Council of Canada.

References

Bartlett, M.S., 1951. The frequency goodness of fit test for probability chains. Proc. Cambridge Philos. Soc. 47, 86–95.

Beran, R., Bilodeau, M., Lafaye de Micheaux, P., 2007. Nonparametric tests of independence between random vectors. J. Multivariate Anal., In Press.

Bilodeau, M., Lafaye de Micheaux, P., 2005. A multivariate empirical characteristic function test of independence between random vectors. J. Multivariate Anal. 95, 345–369.

Blum, J.R., Kiefer, J., Rosenblatt, M., 1961. Distribution free tests of independence based on the sample distribution function. Ann. Math. Statistics 32, 485–498.

Box, G.E.P., Pierce, D.A., 1970. Distribution of residual correlations in autoregressive-integrated moving average time series models. J. Amer. Statist. Assoc. 65, 1509–1526.

Darroch, J.N., 1974. Multiplicative and additive interaction in contingency tables. Biometrika 61, 207–213.

Deheuvels, P., 1979. La fonction de dépendance empirique et ses propriétés: Un test non paramétrique d'indépendance. Acad. Roy. Belg. Bull. Cl. Sci. 5ième série 65, 274–292.

Deheuvels, P., 1981. An asymptotic decomposition for multivariate distribution-free tests of independence. J. Multivariate Anal. 11, 102–113.

Delgado, M.A., 1996. Testing serial independence using the sample distribution function. J. Time Ser. Anal. 17, 271–285.

Ferguson, T.S., 1996. A course in large sample theory. Chapman and Hall: London.

Genest, C., Rémillard, B., 2004. Tests of independence and randomness based on the empirical copula process, Test 13, 335–369.

Ghoudi, K., Kulperger, R.J., Rémillard, B., 2001. A nonparametric test of serial independence for time series and residuals. J. Multivariate Anal. 79, 191–218.

Good, I.J., 1953. The serial test for sampling numbers and other tests for randomness. Proc. Cambridge Philos. Soc. 49, 276–284.

Hong, Y., 1998. Testing for pairwise serial independence via the empirical distribution function. J. R. Statist. Soc. B 60, 429–453.

Kendall, M.G., Smith, B. Babington, 1938. Randomness and random sampling numbers. J. R. Statist. Soc. 101, 147–166.

Kendall, M.G., Smith, B. Babington, 1939. Second paper on random sampling numbers. Suppl. J. R. Statist. Soc. 6, 51–61.

Lancaster, H.O., 1951. Complex contingency tables treated by the partition of χ^2. J. Roy. Statist. Soc. 13, B 242–249.

Lancaster, H.O., 1969. The chi-squared distribution. Wiley, London.

Roy, S.N., Kastenbaum, M.A., 1956. On the hypothesis of no "interaction" in a multiway contingency table. Ann. Math. Statistics 27, 749–757.

Santner, T.J., Duffy, D.E., 1989. The statistical analysis of discrete data. Springer-Verlag, New York.

Stoffer, D.S., Tyler, D.E., McDougall, A.J., 1993. Spectral analysis for categorical time series: Scaling and the spectral envelope. Biometrika 80, 611–622.

Whisenant, E.C., Rasheed, B.K.A., Ostrer, H., Bhatnagar, Y.M., 1991. Evolution and sequence analysis of a human Y-chromosomal DNA fragment. J. Mol. Evol. 33, 133–141.

Figure 1

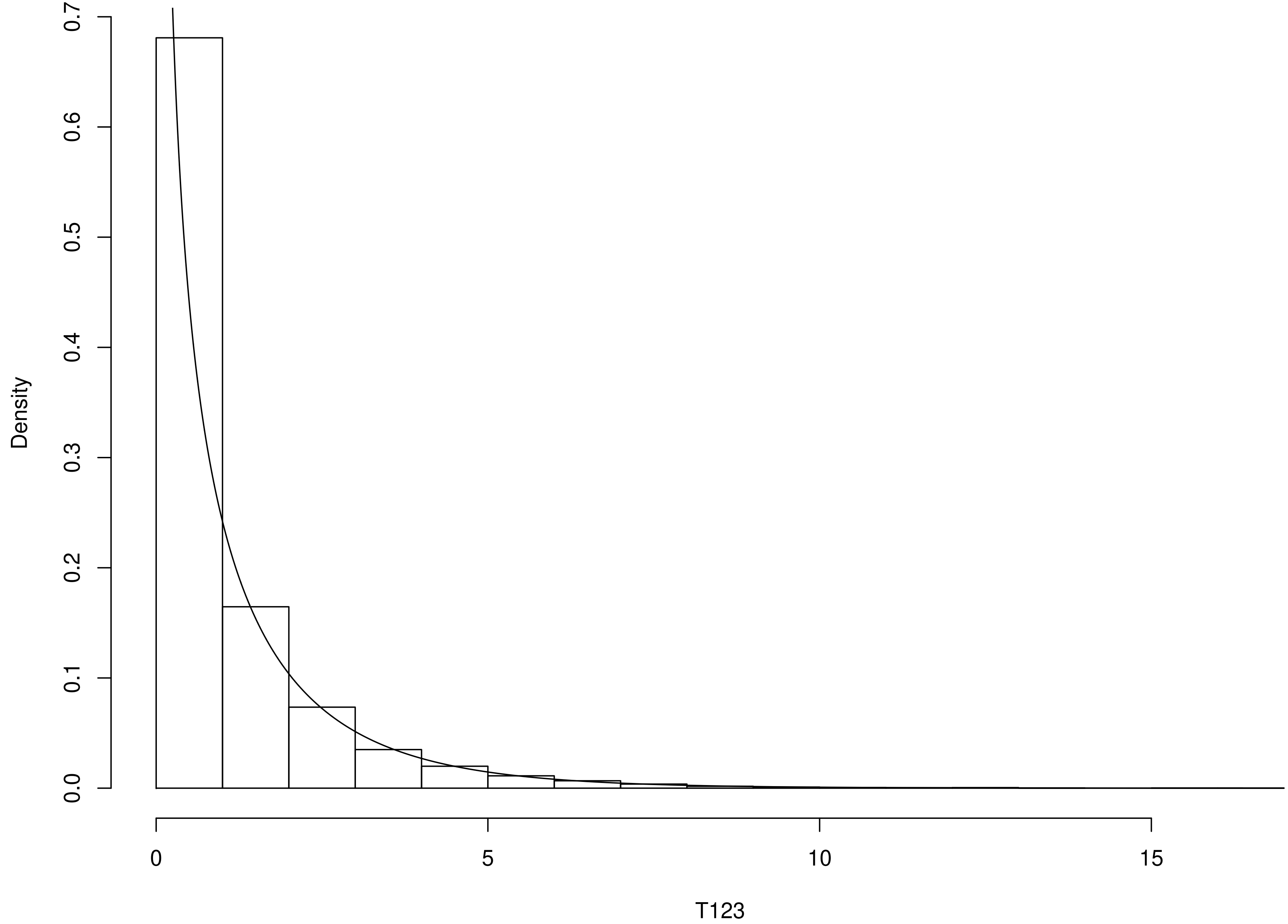


Figure 2

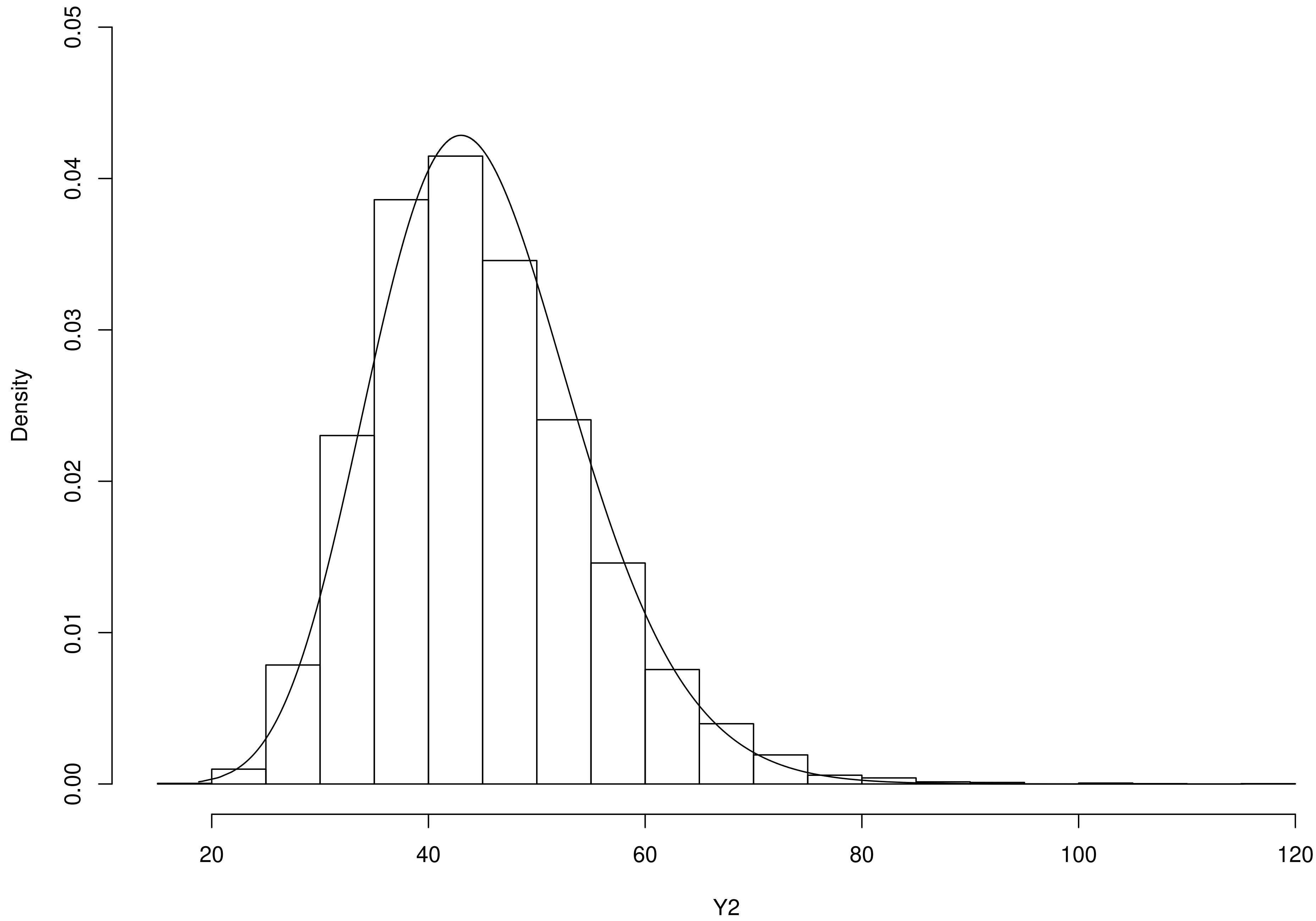
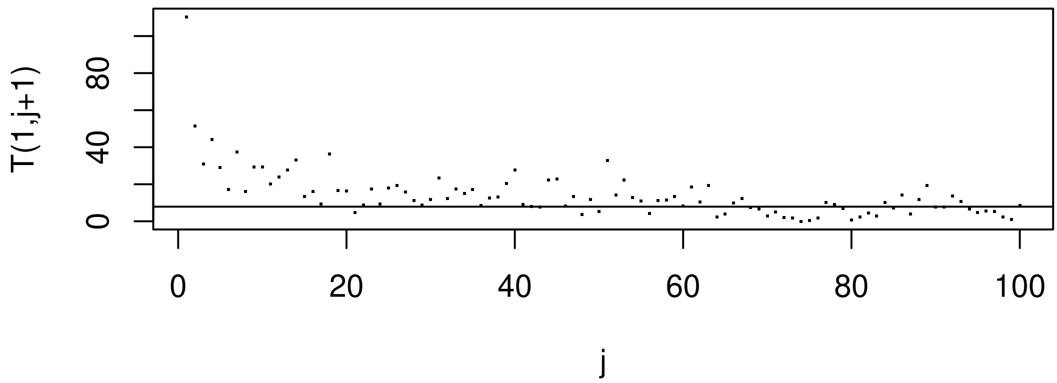
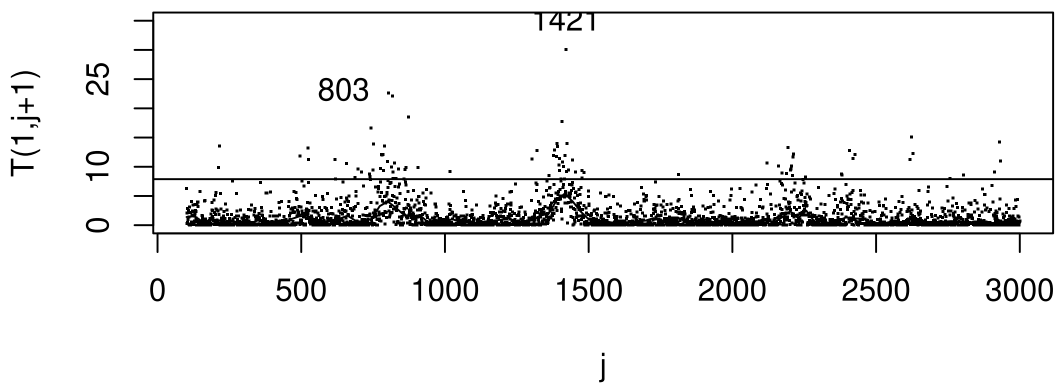


Figure 3


T(1,j+1)
0
40
80
0
20
40
60
80
100
j


T(1,j+1)
0
10
25
1421
803
0
500
1000
1500
2000
2500
3000
j

Table

	.90	.95	.99
$T_{1,2}$	2.73	3.84	6.55
$T_{1,3}$	2.74	3.84	6.87
$T_{2,3}$	2.63	3.69	6.53
$T_{1,2,3}$	2.64	3.83	6.70
χ_1^2	2.70	3.84	6.63
$X_{1,2,3}^2$	7.75	9.37	13.10
χ_4^2	7.77	9.48	13.27

Table 1: Empirical quantiles of the A-dependence statistics T_A and quantiles of chi-square distribution with f_A degrees of freedom.

	.90	.95	.99
$T_{1,2}$	14.53	16.73	21.63
$T_{1,3}$	14.50	16.57	21.61
χ_9^2	14.68	16.92	21.67
$T_{1,2,3}$	36.95	40.96	49.58
χ_{27}^2	36.74	40.11	46.96
$Y_{1,2,3}^2$	57.74	62.54	72.91
χ_{45}^2	57.51	61.66	69.96

Table 2: Empirical quantiles of the A-dependence statistics T_A and quantiles of chi-square distribution with g_A degrees of freedom.

A	1,2	1,3	1,4	2,3	2,4	3,4
$\hat{\gamma}$	0.049	0.050	0.053	0.055	0.051	0.052
A	1,2,3	1,2,4	1,3,4	2,3,4	1,2,3,4	
$\hat{\gamma}$	0.049	0.051	0.053	0.048	1.000	

Table 3: Empirical powers of the A-dependence statistics T_A.