AISM (2006) 58: 527–535
DOI 10.1007/s10463-005-0028-2

R.A. Al-Jarallah · A.R. Soltani · N.A. Al-Kandari

On continuity of the Pearson statistic and sample quantiles

Received: 28 October 2004 / Revised: 9 June 2005 / Published online: 17 June 2006

Abstract Convergence with probability one (in probability) of sequences of the sample quantiles and the Pearson statistic that are formed by columns of $N \times n$ arrays of random variables and bivariate random vectors, respectively, is established, $n \to \infty$. Two applications for the continuity of the Pearson statistics, when sampling is only possible along a sequence converging to an inaccessible targeting value, are presented.

Keywords Pearson statistic · Sample q-quantiles · Contingency tables · Array continuity · ARMA models · Stable random vectors

1 Introduction

In this work we consider with probability one or in probability convergence of sequences of the statistics sample q-quantiles and the Pearson statistics formed by the columns of sequences of arrays of random variables $\mathbf{X_n} = \{X_{n,i};\ i = 1, \ldots, N\}$, and bivariate random vectors, $(\mathbf{X}, \mathbf{Y})_\mathbf{n} = \{(X_{n,i}, Y_{n,i});\ i = 1, \ldots, N\}$ $n = 0, 1, 2, \ldots$ respectively. In sampling, for each n, the array $\mathbf{X_n}$ can be considered as a sample of size N from the random variable X_n, while, for each i, $\{X_{n,i};\ n = 1, 2, \ldots\}$ is a trajectory of the sequence $\{X_n\}_{n\geq 1}$. Thus, a sequence of

R.A. Al-Jarallah · A.R. Soltani (✉) · N.A. Al-Kandari
Department of Statistics and Operations Research,
Faculty of Science, Kuwait University,
P.O. Box 5969 Safat-13060, Kuwait
E-mail: reema@kuc01.kuniv.edu.kw
E-mail: soltani@kuc01.kuniv.edu.kw
E-mail: noriah@kuc01.kuniv.edu.kw

A.R. Soltani
Department of Statistics, College of Science,
Shiraz University, Shiraz 71454, Iran

arrays can represent a (independent) sample of the sample paths. Such a view has been realized in some areas of statistics. In analysis of longitudinal data, a repeated measurement (most often) takes place over time on a number of experimental units, Singer and Willett (2003). In time series parallel trajectories, rather than a single one, have appeared to be more informative for model buildings, see Pourahmadi (1999, 2001). This work is organized as follows.

The continuity for the sample q-quantiles and the Pearson statistic is established in Sect. 2, by showing that the row-wise convergence of the array leads to the convergence of the corresponding statistics; for the Pearson statistic we assume that the targeting variable is absolutely continuous. In Sect. 3, we provide two applications for the continuity of the Pearson statistic. In the first application, we propose a method to detect the tendency of the present and future in ARMA time series models towards independence. The second application provides a procedure to locate unknown atoms of the spectral measure of a multivariate symmetric stable distribution.

For other issues concerning convergence of arrays, see Kuczmaszewska (2004) and references therein. For a recent work on the asymptotic behavior of the Pearson statistic see Kruglov (2001).

2 Continuity for sample q-quantile function and Pearson statistic

Let us first define some terminologies. The array of bivariate random vectors $(\mathbf{X}, \mathbf{Y})_{\mathbf{0}}$ is said to be absolutely continuous if for each i, $(X_{0,i}, Y_{0,i})$ is absolutely continuous. The sequence $(\mathbf{X}, \mathbf{Y})_{\mathbf{n}}$ is said to converge to $(\mathbf{X}, \mathbf{Y})_{\mathbf{0}}$ with probability one if for each i, $(X_{n,i}, Y_{n,i}) \to (X_{0,i}, Y_{0,i})$ with probability one, as $n \to \infty$. The convergence in probability of a sequence of arrays is defined similarly. Same definitions are also for arrays of random variables.

For a sequence of independent and identically distributed (i.i.d) random variables $\xi_1, \ldots, \xi_N$, let

$$F_N(t) \equiv F_N(t, \omega) = \frac{1}{N} \sum_{i=1}^{N} 1_{(-\infty,t]}(\xi_i(\omega)),$$

be the corresponding empirical distribution function, Billingsley (1995). Also let F_N^{-1} be the generalized inverse of F_N, i.e.,

$$F_N^{-1}(q) = \inf\{t : F_N(t) \geq q\}, \quad 0 < q < 1.$$

The F_N^{-1} is called the sample quantile function, and for $0 < q < 1$, the $F_N^{-1}(q)$ is the smallest sample q-quantile. If ξ_1 is absolutely continous, then $F_N^{-1}(q) = \xi_{(\langle qN \rangle)}$ is the $\langle qN \rangle$th sample order statistic, $\langle qN \rangle = \min\{m : m \geq qN\}$, and is the unique sample q-quantile. The sample quantile function for an array $\mathbf{X} = \{X_1, \ldots, X_N\}$ is defined similarly. Corresponding to a sequence of arrays of random variables, $\mathbf{X_n} = \{X_{n,i}, i = 1, \ldots, N\}$, let $\{F_{n,N}^{-1}(q, \omega)\}_{n=1,2,\ldots}$ be the sequence of the sample quantile functions.

Lemma 2.1 *Assume the array* $\mathbf{X_n} \to \mathbf{X_0}$ *with probability one, then with probability one for each* $N \geq 1$,

$$F_{n,N}(t) \to F_{0,N}(t),$$

on continuity points of $F_{0,N}(t)$, *as* $n \to \infty$.

Proof The Lemma indeed asserts that there is a set A of probability one such that for every $\omega \in A$, $F_{n,N}(t,\omega) \to F_{0,N}(t,\omega)$ on continuity points of $F_{0,N}(.,\omega)$. This will follow since $1_{(-\infty,t]}(X_{n,i}(\omega)) = 1_{[X_{n,i}(\omega),\infty)}(t)$, are distribution functions corresponding to the unit masses at $X_{n,i}(\omega)$, $n = 0, 1, 2, \ldots$, respectively. □

Theorem 2.1 *Assume the array* $\mathbf{X_n}, n = 1, 2, \ldots$ *converges to the array* $\mathbf{X_0}$ *with probability one. Then with probability one*

$$F_n^{-1}(q) \to F_0^{-1}(q),$$

at every continuity point q *of* $F_0^{-1}(q)$, *as* $n \to \infty$.

Proof The result follows from Lemma 2.1 and the well-known result that the weak convergence of distribution function is equivalent to the weak convergence of corresponding quantile functions, see the proof of Theorem 25.6, Billingsley (1995). The proof is complete. □

Remark 2.1 In most of applications the discontinuity points of $F_0^{-1}(q)$ do not depend on ω and are $i/N, i = 1, \ldots, N$. If arrays $\mathbf{X_n}, n = 1, 2, \ldots$ are absolutely continuous then

$$F_n^{-1}(q) = \begin{cases} \mathbf{X}_{n,(qN)} & qN \text{ integer} \\ \mathbf{X}_{n,([qN]+1)} & \text{otherwise,} \end{cases}$$

where $X_{n,(r)}$ is the rth order statistic of the array X_n and $[qN]$ denotes the integer part of qN. Thus, it follows from Theorem 2.1 that $X_{n,(r)} \to X_{0,(r)}$ with probability one, $r = 1, \ldots, N$.

In order to investigate the continuity of the Pearson statistic let, for each $n = 0, 1, 2, \ldots$, the cells $\mathbf{I}_n^{t,s} = \mathbf{I}_n^t \times \mathbf{J}_n^s$, $t = 1, \ldots, T$, $s = 1, \ldots, S$ partition the range of $(\mathbf{X}, \mathbf{Y})_\mathbf{n}$, into rectangles, where for each $i = 1, \ldots, N$, the boundaries of intervals $\mathbf{I}_n^t = (a_n^t, a_n^{t+1}]$ and $\mathbf{J}_n^s = (b_n^s, b_n^{s+1}]$ could depend on $X_{n,i}$ and $Y_{n,i}$, respectively. Let

$$\nu_{n,N} = \sum_{t=1}^{T} \sum_{s=1}^{S} \frac{[N Z_{n,N}^{t,s} - Z_{n,N}^{t,\cdot} Z_{n,N}^{\cdot,s}]^2}{N Z_{n,N}^{t,\cdot} Z_{n,N}^{\cdot,s}} \tag{1}$$

be the Pearson statistic for the array $(\mathbf{X}, \mathbf{Y})_\mathbf{n}$, where

$$Z_{n,N}^{t,s} = \sum_{i=1}^{N} 1_{I_n^{t,s}}\left((X_{n,i}, Y_{n,i})\right), \quad Z_{n,N}^{t,\cdot} = \sum_{s=1}^{S} Z_{n,N}^{t,s}, \quad Z_{n,N}^{\cdot,s} = \sum_{t=1}^{T} Z_{n,N}^{t,s} \tag{2}$$

are the frequencies of the cell $\mathbf{I}_n^{t,s}$, the strips $\bigcup_{s=1}^{S} \mathbf{I}_n^{t,s}$ and $\bigcup_{t=1}^{T} \mathbf{I}_n^{t,s}$, respectively, $n = 0, 1, 2, \ldots$

The limiting behavior of $\nu_{n,N}$ as $n \to \infty$, naturally, depends on the limiting values of $Z_{n,N}^{t,s}$, $Z_{n,N}^{\cdot,s}$ and $Z_{n,N}^{t,\cdot}$. The following lemma concerns this point.

Lemma 2.2 *Assume* $(\mathbf{X}, \mathbf{Y})_{\mathbf{n}}$, $n = 1, 2, \ldots$, *converges to* $(\mathbf{X}, \mathbf{Y})_{\mathbf{0}}$ *with probability one, and* $(\mathbf{X}, \mathbf{Y})_{\mathbf{0}}$ *is absolutely continuous. Also, assume for every* t *and* s, *the boundaries* $a_n^t \to a_0^t$, $b_n^s \to b_0^s$ *with probability one, as* $n \to \infty$. *Then for every* t *and* s,

$$Z_{n,N}^{t,s} \to Z_{0,N}^{t,s}, \qquad Z_{n,N}^{t,\cdot} \to Z_{0,N}^{t,\cdot}, \qquad Z_{n,N}^{\cdot,s} \to Z_{0,N}^{\cdot,s},$$

with probability one.

Proof Since N is fixed, the result will follow if $1_{I_n^{t,s}}\big((X_{n,i}, Y_{n,i})\big) \to 1_{I_0^{t,s}}\big((X_{0,i}, Y_{0,i})\big)$ with probability one, as $n \to \infty$. A classical probability argument will imply this assertion whenever $(X_{0,i}(\omega), Y_{0,i}(\omega))$ is either an interior point of $\mathbf{I}_0^{t,s}$ or its complement. The assertion may not be true when $(X_{0,i}(\omega), Y_{0,i}(\omega))$ falls on the boundary of $\mathbf{I}_0^{t,s}$; but this event has a zero probability due to the absolute continuity assumption of the targeting vector. The proof of the lemma is complete. □

Lemma 2.2 provides the following with probability one continuity property for contingency tables, CT in short. Let $\mathbf{T_n}$ be the CT corresponding to $\big\{(\mathbf{X}, \mathbf{Y})_{\mathbf{n}},\ \mathbf{I_{n,N}}\big\}$ and $\mathbf{T_0}$ be the CT corresponding to $\big\{(\mathbf{X}, \mathbf{Y})_{\mathbf{0}},\ \mathbf{I_{0,N}}\big\}$, then we say $\mathbf{T_n} \longrightarrow \mathbf{T_0}$ with probability one (in probability) if for every t and s, $Z_{n,N}^{t,s} \longrightarrow Z_{0,N}^{t,s}$ with probability one (in probability). The following theorem immediately follows from Lemma 2.2.

Theorem 2.2 *Suppose partition sizes* T *and* S *are deterministic and do not depend on* n. *Then under the assumptions in Lemma 2.2, the* $\mathbf{T_n}$ *converges to* $\mathbf{T_0}$ *with probability one.*

The following theorem, which also follows from Lemma 2.2, provides with probability one continuity property of the Pearson statistic for contingency tables.

Theorem 2.3 *Assume the partition sizes* T *and* S *are deterministic and that for every* $t = 1, \ldots, T$, $s = 1, \ldots, S$ *and* $n = 0, 1, 2 \ldots$, $Z_{n,N}^{t,\cdot} \neq 0$, $\quad Z_{n,N}^{\cdot,s} \neq 0$, *with probability one. Then under the assumptions of Lemma 2.2,*

$$\nu_{n,N} \to \nu_{0,N}$$

with probability one, as $n \to \infty$, *for any fixed sample size* $N \geq 1$.

Remark 2.2 It can be deduced from Theorem 2.3 that $\nu \equiv \lim_{N\to\infty} \lim_{n\to\infty} \nu_{n,N}$ follows the Chi-square distribution. This allows one to apply the χ^2-test, for sufficiently large n in circumstances that the target $(X, Y)_0$ is not in access, but $(X, Y)_n$ can be sampled.

Remark 2.3 By using Theorem 2.1, Theorem 2.2, given above, and Theorem 20.5 (ii) in Billingsley (1995) we deduce that Theorem 2.2 and Theorem 2.3 are still valid if convergence with probability one is replaced by the convergence in probability.

3 Applications

The continuity property of the Pearson statistic presented in Sect. 3, appears to be useful when the target arrays $(\mathbf{X}, \mathbf{Y})_{\mathbf{0}}$ is not in access and its distribution is not known. We demonstrate such an occasion and provide two examples.

The underground streams from Zagros mountains, the major mountain system in west of Iran, has been, in part, a water sources for different underground water reservoir in different locations in Fars province. There is concern that these underground streams are becoming less significant, due to the increase in water production from intermediate regions. Different wells in one region can be considered as the first group of experimental units, where their water yields in a given year form a column, and successive yearly water yields of a well form a row, giving $\mathbf{X_n}$ values. Similar measurements for another region give $\mathbf{Y_n}$ values. The yearly water yields in the two regions, (X_n, Y_n), are expected to be dependent if they are supplied by the same underground streams. Upon the gradual shortage of the stream supply, (X_n, Y_n) approaches a steady, but unaccessible, state (X_0, Y_0) in future that are expected to have independent components.

Example 3.1 Rainfalls data are usually analyzed using time series models, see Madsen et al. (2000), and references therein. Let X_n denote the yearly rainfalls at year n in a given region. Also let $\{X_{n,N}, n = 1, \ldots, n_0\}$ denote n_0 measurements on successively rainfalls recorded in the station which is indexed by N. We demonstrate below that the continuity of the Pearson statistic can be applied to verify if the present X_1 and future X_n rainfalls have tendency to become independent. We use simulated data from the AR(1) model: $X_n = \alpha X_{n-1} + Z_n$, $n = 0, \pm 1, \pm 2, \ldots$. For rainfalls, each Z_n is an climate innovation affecting X_n. Define double arrays $(\mathbf{X}, \mathbf{Y})_{\mathbf{n}}$, $n = 0, 1, \ldots$, by $(\mathbf{X}, \mathbf{Y})_{\mathbf{0}} = \{(X_{1,i}, X_{n_0,i});\ i = 1, \ldots N\}$, and $(\mathbf{X}, \mathbf{Y})_{\mathbf{n}} = \{(X_{1,i}, X_{n,i});\ i = 1, \ldots N\}$, $n = 2, \ldots, n_0$. The Pearson statistic $\nu_{n,N}$ given in Eq. 1 was calculated for $n = 2, \ldots, 100$ and $N = 500$ realizations. Different values of α, the AR(1) coefficient, were examined and the percentiles were chosen to form partitions. (Figs. 1, 2)

Example 3.2 Let X be a bivariate symmetric α- stable random vector with atomic spectral measure Γ, namely,

$$E\mathrm{e}^{\mathrm{i}\nu.X} = \mathrm{e}^{-2\sum_{\mathrm{j}=1}^{k} |\nu.s_j|^{\alpha}\gamma_j}, \quad \nu \in R^2,$$

where $s_1, \ldots, s_k$ are on the upper part of the unit circle in R^2. The symmetric measure Γ on the unit circle with atoms $s_1, \ldots, s_k, -s_1, \ldots, -s_k$ and masses $\Gamma(s_j) = \Gamma(-s_j) = \gamma_j$, $j = 1, \ldots, k$, is called the spectral measure of X. An alternative representation for X is due to Modarres and Nolan (1994), namely, in distribution,

$$X = \gamma_1^{1/\alpha} Z_1 s_1 + \cdots + \gamma_k^{1/\alpha} Z_k s_k,$$

where $Z_1, \ldots, Z_k$ are independent one dimensional normalized α-stable random variables that are symmetric,

$$E\mathrm{e}^{\mathrm{i}zZ_j} = \mathrm{e}^{-|z|^{\alpha}}, \quad z \in R.$$

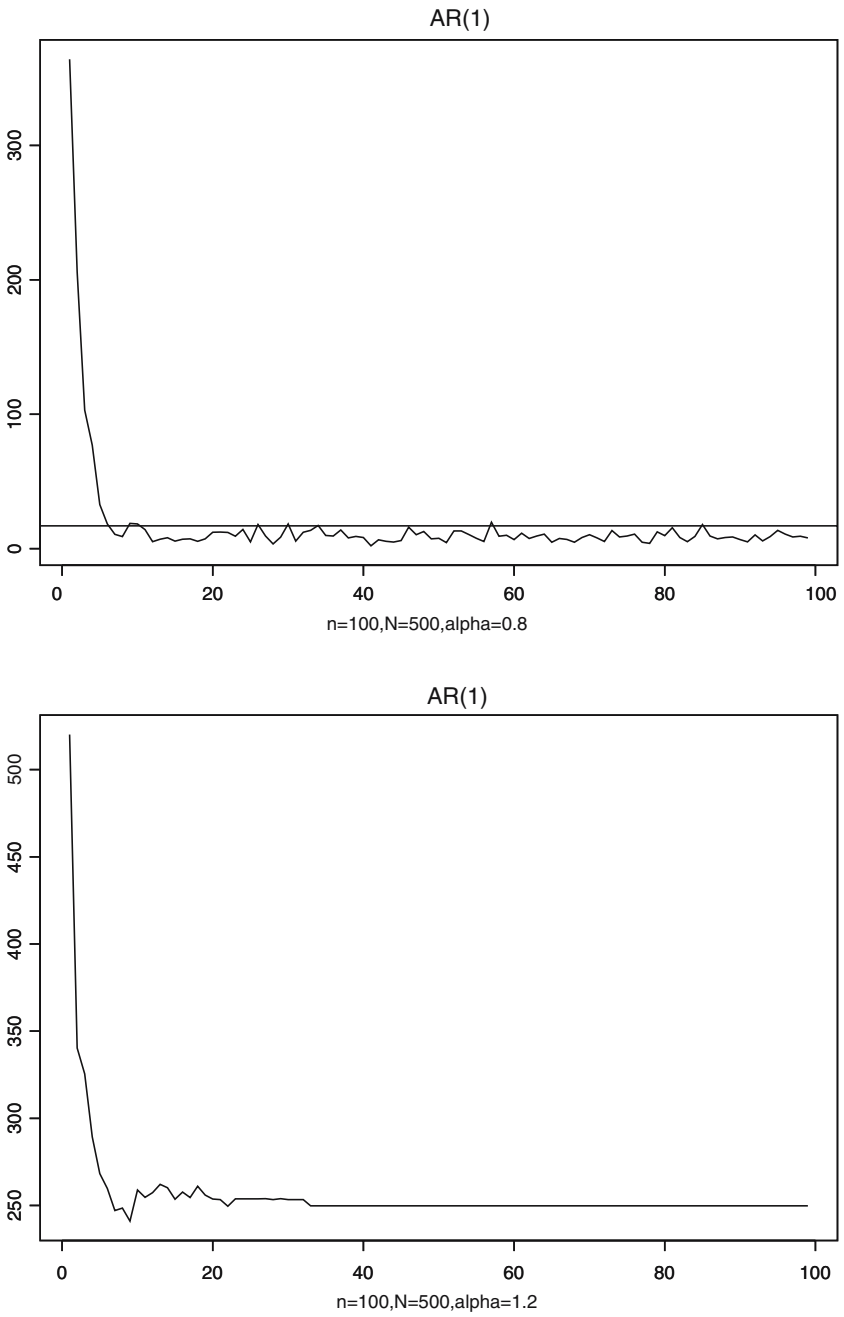


Fig. 1 *Top*: The values of the Pearson statistics against n in AR(1) with $\alpha = 0.8$. Clearly the Pearson statistic approaches zero rather fast, confirming that X_1 and X_n become nearly independent as n increases. *Bottom*: The values of the Pearson statistics against n in AR(1) with $\alpha = 1.2$. The Pearson statistic remains significantly large, indicating that X_1 and X_n remain dependent, no matter how large n is

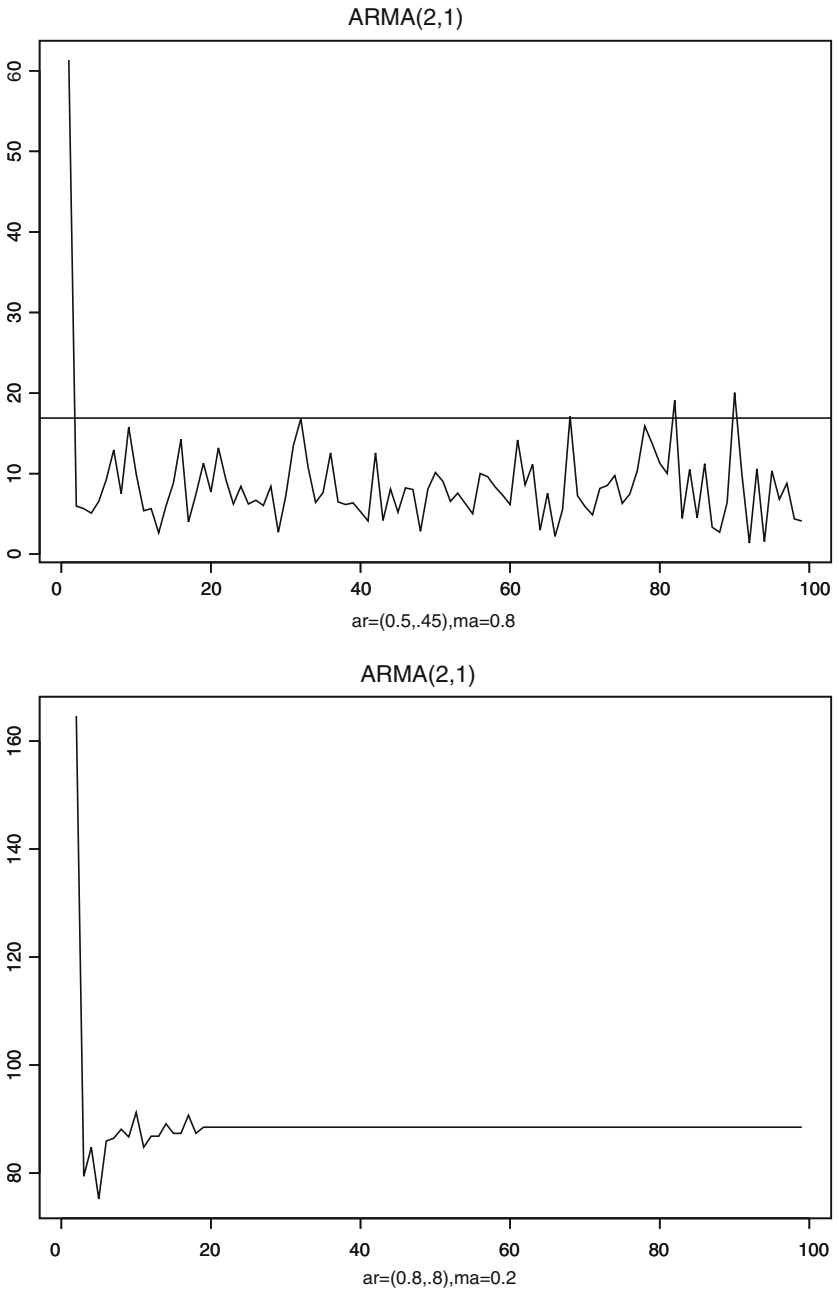
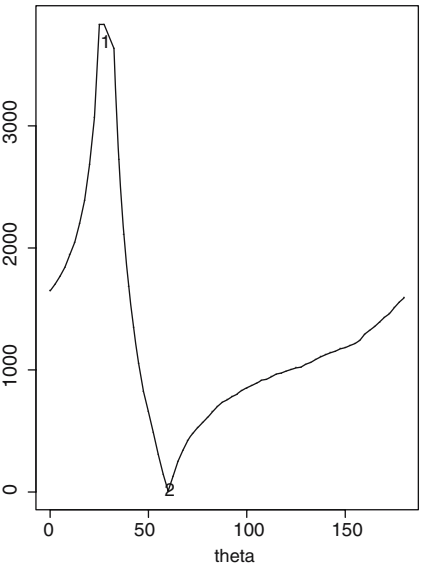
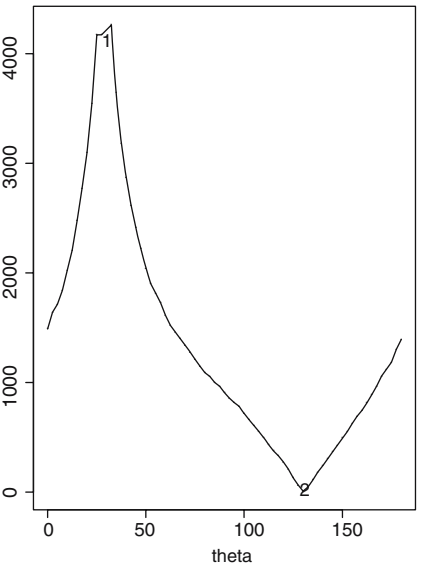


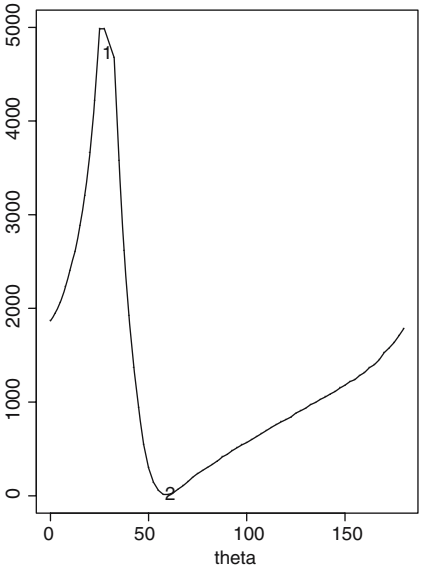
Fig. 2 *Top*: The values of the Pearson statistics against n in ARMA(2,1) with AR coefficients 0.5 and 0.45 and MA coefficient 0.8. Clearly the Pearson statistic approaches zero, confirming that X_1 and X_n become nearly independent as n increases. *Bottom*: The values of the Pearson statistics against n in ARMA(2,1) with AR coefficients 0.8, 0.8 and MA coefficient 0.2. The Pearson statistic remains significantly large, indicating that X_1 and X_n remain dependent, no matter how large n is



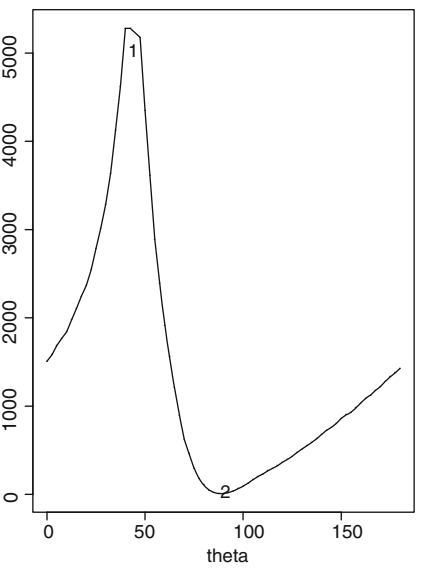
Pearson statistic; N=2000, alpha=.75, masses=30,60



Pearson statistic; N=2000, alpha=1.5, masses=30,130



Pearson statistic; N=2000, alpha=1.9, masses=30,60



Pearson statistic; N=2000, alpha=2, masses=45,90

Fig. 3 Pearson statistics values for bivariate stable random vectors. The spectral measure possesses only two atoms s_1 and s_2. The Pearson statistic corresponding to the N observations of (X_0, X_n), namely, $X_0 = \{(s_1^*, X_i), i = 1, \ldots, N\}$, $X_n = \{(\theta^*, X_i), i = 1, \ldots, N\}$, where $\theta = 0^\circ, 2.5^\circ, 5^\circ, \ldots, 180^\circ$, is the fixed grid in $[(0, \pi]$, is plotted against θ. We observe sharp decrease in values of the Pearson statistic as θ approaches the position of the second true atom; the minimum occurs in the vicinity of the true atom. *Top Left*: $N = 2,000$, $\alpha = 0.75$, $s_1 = 30$, $s_2 = 60$. *Top Right*: $N = 2,000$, $\alpha = 1.5$, $s_1 = 30$, $s_2 = 130$. *Bottom Left*: $N = 2,000$, $\alpha = 1.9$, $s_1 = 30$, $s_2 = 60$. *Bottom Right*: $N = 2,000$, $\alpha = 2$, $s_1 = 45$, $s_2 = 90$. In each graph the deviation of the point of the minimum from the true atom is at most 2.5°

The problem of estimating the atoms $s_1, \ldots, s_k$ and masses $\gamma_1, \ldots, \gamma_k$, using a sample $X_1, \ldots, X_N$ of X was considered by Rachev and Xin (1993), Cheng and Rachev (1994), Nolan et al. (2001). The approach in the first two works is nonparametric and an atom is estimated using samples that stay far away from the origin neighboring the direction of the atom. Other two methods, empirical characteristic function method and projection method are given in the third work cited above. For more on effectiveness of cited methods, their drawbacks and numerical complications, see Nolan et al. (2001).

The continuity of the Pearson statistic can be used to indicate if there are atoms in a given vicinity. Let $X_1, \ldots, X_N$ be a sample from X and $t_1, \ldots t_n$ be a fixed grid on the unit sphere. Let $t_1^*, \ldots, t_n^*$ annihilate $t_1, \ldots, t_n$, respectively. Let s be an atom and assume $t_n \to s$, then the array $X_n = \{X_{n,i} = (t_n^*, X_i), i = 1, \ldots, N\}$ converges pointwise to the array $X_0 = \{X_{0,i} = (s^*, X_i), i = 1, \ldots, N\}$. Consequently, (X_1, X_n) converges pointwise to (X_1, X_0). But components of (X_1, X_0) are less dependent compare to components of (X_1, X_n). So due to the continuity of the Pearson statistic and its sensitivity to dependent data, we expect a decrease in the value of ν_n as t_n approaches s. Thus, the plot of ν_n against n will indicate the existence of a hidden atom in a given vicinity. Using Nolan's simulation program, in //www.cas.american.edu/∼jpnolan, N realizations of a bivariate α-stable random vector X having certain atoms $s_1, \ldots, s_j$ were simulated; the ν_n was evaluated and plotted in Fig. 3.

Acknowledgements The authors are thankful to referees for their comments. This research was supported by Kuwait University, Research Grant No.[SS01/03].

References

Billingsley, P. (1995). *Probability and Measure* (3rd ed.) New York: Wiley.

Cheng, B.N., & Rachev, S.T. (1994). Multivariate stable future prices. *Mathematical Finance*, *2,* 133–153.

Kruglov, V.M. (2001). The asymptotic behavior of the Pearson statistic. *Theory of Probability and its Applications*, *45,* 69–92.

Kuczmaszewska, A. (2004). On some conditions for complete convergence for arrays. *Statistics and Probability Letters*, *66,* 399–405 .

Madsen, H., Butts, M.B., Khu, S.T., & Liong, S.Y. (2000). Data assimilation in rainfall-runoff forecasting. In: *Proceedings 4th International Conference on Hydroinformatics*, Iowa, USA, July 2000.

Modarres, R., & Nolan, J.P. (1994). A method for simulating stable random vectors. *Computational Statistics*, *9,* 11–19.

Nolan, J.P., Panoraska, A.K., & McCulloch, J.H. (2001). Estimation of stable spectral measures, stable non-Gaussian models in finance and econometrics. *Mathematical and Computer Modelling*, *34,* 1113–1122.

Pourahmadi, M. (1999). Joint mean covariance models with application to longitudinal data: unconstrained parameterisation. *Biometrika*, *86,* 677–690.

Pourahmadi, M. (2001). *Foundation of Time Series Analysis and Prediction Theory*. New York: Wiley.

Rachev, S.T., & Xin, H. (1993). Test for association of random variables in the domain of attraction of multivariate stable law. *Probability and Mathematical Statistics*, *14,* 125–141.

Singer, J.D., & Willett, J.B. (2003). *Applied longitudinal Data Analysis: Modelling Change and Event occurrence*. New York: Oxford University Press.