dCovTS: Distance Covariance/Correlation for Time Series

Maria Pitsillou and Konstantinos Fokianos

Abstract The distance covariance function is a new measure of dependence between random vectors. We drop the assumption of i.i.d data to introduce distance covariance for time series. The R package **dCovTS** provides functions that compute and plot distance covariance and correlation functions for both univariate and multivariate time series. Additionally it includes functions for testing serial independence based on distance covariance. This paper describes the theoretical background of distance covariance methodology in time series and discusses in detail the implementation of these methods with the R package **dCovTS**.

Introduction

There has been a considerable recent interest in measuring dependence by employing the concept of *distance covariance function*. Székely et al. (2007) initially introduced the distance covariance as a new measure of dependence defined as the weighted L_2-norm between the joint characteristic function of two random vectors of arbitrary, but not necessarily of equal dimensions, and their marginal characteristic functions. However, the idea of using distance covariance for detecting independence can be also found in some early work by Feuerverger (1993). He considered measures of this form, with the main differences being the restriction to the univariate case and the choice of the weight function. Since Székely et al.'s (2007) work, there has been a wide range of studies extending the distance covariance definition and methodology in various topics; see Gretton et al. (2009) and Josse and Holmes (2014) and the references therein for a nice review.

Székely et al.'s (2007) distance covariance methodology is based on the assumption that the underlying data are i.i.d. However, this assumption is often violated in many practical problems. Remillard (2009) proposed to extend the distance covariance methodology to a time series context in order to measure serial dependence. There have been few works on how to develop a distance covariance methodology in the context of time series (Zhou, 2012; Dueck et al., 2014; Davis et al., 2016). Motivated by the work of Székely et al. (2007), Zhou (2012) recently defined the so-called *auto-distance covariance function* (ADCV) - and its rescaled version, the so-called *auto-distance correlation function* (ADCF), for a strictly stationary multivariate time series. Compared to the classical Pearson autocorrelation function (ACF) which measures the strength of linear dependencies and can be equal to zero even when the variables are related, ADCF vanishes only in the case where the observations are independent. However, Zhou (2012) studied the asymptotic behavior of ADCV at a fixed lag order. Fokianos and Pitsillou (2016a) relaxed this assumption and constructed a univariate test of independence by considering an increasing number of lags following Hong's (1999) generalized spectral domain methodology. Although the proposed methodology is for univariate processes, it can be extended for multivariate processes.

Zhou (2012) developed a distance covariance methodology for multivariate time series, but he did not explore the interrelationships between the various time series components. Fokianos and Pitsillou (2016b) made this possible by defining the matrix version of pairwise auto-distance covariance and correlation functions. In particular, they construct multivariate tests of independence based on these new measures in order to identify whether there is some inherent nonlinear interdependence between the component series.

The **energy** (Rizzo and Székely, 2013) package for R (R Core Team, 2014) is a package that involves a wide range of functions for the existing distance covariance methodology. However, there is no package for the aforementioned distance covariance methodology in time series. Thus, we aim at filling this gap by publishing an R-package named **dCovTS**. In this first version of the package, we provide functions that compute and plot ADCV and ADCF using the functions `dcov` and `dcor` respectively from **energy** (Rizzo and Székely, 2013) package. The new testing methodology proposed by Fokianos and Pitsillou (2016a,b) is also included in the package.

The structure of the paper is as follows. In the first two sections we introduce the theoretical background of distance covariance function for both univariate and multivariate time series respectively. In the next section, we briefly state the main results about the asymptotic properties of distance covariance function. The proposed testing methodology for both univariate and multivariate time series are also described. Empirical p-values of the tests and empirical critical values for the distance correlation plots are computed via the wild bootstrap methodology (Dehling and Mikosch, 1994; Shao, 2010; Leucht and Neumann, 2013b) which is explained in the corresponding section. The implementation

section demonstrates the usage of the package with two real data examples. Lastly, we give some concluding remarks and some further points for future extensions of the **dCovTS** package.

Distance covariance function

Denote a univariate strictly stationary time series by $\{X_t, t \in \mathbb{Z}\}$. Motivated by Székely et al. (2007) and Zhou (2012), we define the distance covariance function as a function of the joint and marginal characteristic functions of the pair $\left(X_t, X_{t+j}\right)$. Denote by $\phi_j(u,v)$ the joint characteristic function of X_t and X_{t+j}; that is

$$\phi_j(u,v) \quad = \quad E\Big[\exp\Big(i(uX_t + vX_{t+j})\Big)\Big], \quad j = 0, \pm 1, \pm 2, \ldots,$$

and the marginal characteristic functions of X_t and X_{t+j} as $\phi(u) := \phi_j(u,0)$ and $\phi(v) := \phi_j(0,v)$ respectively, where $(u,v) \in \mathbb{R}^2$, and $i^2 = -1$. For a strictly stationary α-mixing univariate time series, Hong (1999) defined a new measure of dependence between the joint characteristic function of X_t and its lagged observation X_{t+j} and the product of their marginals, namely

$$\sigma_j(u,v) = \phi_j(u,v) - \phi(u)\phi(v), \quad j = 0, \pm 1, \pm 2, \ldots, \tag{1}$$

where $(u,v) \in \mathbb{R}^2$. Considering the property that the joint characteristic function factorizes under independence of X_t and X_{t+j}, $\sigma_j(u,v)$ equals 0 if and only if X_t and X_{t+j} are independent. Thus, compared to the classical autocorrelation function (ACF), $\sigma_j(\cdot,\cdot)$ can capture all pairwise dependencies including those with zero autocorrelation. The auto-distance covariance function (ADCV), $V_X(j)$, between X_t and X_{t+j} is then defined as the square root of

$$V_X^2(j) \quad = \quad \int_{\mathbb{R}^2} \left|\sigma_j(u,v)\right|^2 d\mathcal{W}(u,v), \quad j = 0, \pm 1, \pm 2, \ldots \tag{2}$$

where $\mathcal{W}(\cdot,\cdot)$ is a positive weight function for which the above integral exists.

Although Hong (1999) suggests the use of an arbitrary integrable weight function, $\mathcal{W}(\cdot,\cdot)$, we propose the use of a non-integrable weight function, i.e.

$$\mathcal{W}(u,v) = \mathcal{W}_0(u)\mathcal{W}_0(v) = \frac{1}{\pi\,|u|^2}\frac{1}{\pi\,|v|^2}, \quad (u,v) \in \mathbb{R}^2 \tag{3}$$

which avoids missing any potential dependence among observations (Székely et al., 2007, p. 2771). Rescaling (2), one can define the auto-distance correlation function (ADCF) as the positive square root of

$$R_X^2(j) \quad = \quad \frac{V_X^2(j)}{V_X^2(0)}, \quad j = 0, \pm 1, \pm 2, \ldots \tag{4}$$

for $V_X^2(0) \neq 0$ and zero otherwise. Székely et al. (2007) showed that by applying a non-integrable weight function, like (3), ADCF is scale invariant and is not zero under dependence.

The empirical ADCV, $\widehat{V}_X(\cdot)$, is the non-negative square root of

$$\widehat{V}_X^2(j) \quad = \quad \frac{1}{(n-j)^2}\sum_{r,l=1+j}^{n} A_{rl}B_{rl}, \quad 0 \leq j \leq (n-1) \tag{5}$$

and $\widehat{V}_X^2(-j) = \widehat{V}_X^2(j)$, for $-(n-1) \leq j < 0$, where $A = A_{rl}$ and $B = B_{rl}$ are Euclidean distance matrices given by

$$A_{rl} \quad = \quad a_{rl} - \bar{a}_{r\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot},$$

with $a_{rl} = |X_r - X_l|$, $\bar{a}_{r\cdot} = \left(\sum_{l=1+j}^{n} a_{rl}\right)/(n-j)$, $\bar{a}_{\cdot l} = \left(\sum_{r=1+j}^{n} a_{rl}\right)/(n-j)$, $\bar{a}_{\cdot\cdot} = \left(\sum_{r,l=1+j}^{n} a_{rl}\right)/(n-j)^2$. B_{rl} is defined analogously in terms of $b_{rl} = |Y_r - Y_l|$, where $Y_t \equiv X_{t+j}$. Székely and Rizzo (2014) proposed an unbiased version of the sample distance covariance. In the context of time series data this is given by

$$\widetilde{V}_X^2(j) \quad = \quad \frac{1}{(n-j)(n-j-3)}\sum_{r\neq l}\widetilde{A}_{rl}\widetilde{B}_{rl}, \tag{6}$$

for $n > 3$, where $\widetilde{A}_{rl}$ is the (r,l) element of the so-called $\mathcal{U}$-centered matrix $\widetilde{A}$, defined by

$$\widetilde{A}_{rl} = \begin{cases} a_{rl} - \frac{1}{n-j-2}\sum_{t=1+j}^{n} a_{rt} - \frac{1}{n-j-2}\sum_{s=1+j}^{n} a_{sl} + \frac{1}{(n-j-1)(n-j-2)}\sum_{t,s=1+j}^{n} a_{ts}, & r \neq l; \\ 0, & r = l. \end{cases}$$

The empirical ADCF, $\widehat{R}_X(j)$ (or its unbiased version, $\widetilde{R}_X(j)$), can be obtained by replacing (5) (or (6)) into (4). The functions `ADCV` and `ADCF` in **dCovTS** return the empirical quantities $\widehat{V}_X(\cdot)$ and $\widehat{R}_X(\cdot)$ respectively. Using the same functions with argument `unbiased=TRUE`, the results correspond to the unbiased sqaured quantities $\widetilde{V}^2_X(\cdot)$ and $\widetilde{R}^2_X(\cdot)$. Note that the default option has been set to `unbiased=FALSE` (corresponding to (5)).

Distance covariance matrix

We denote by $\{\mathbf{X}_t : t = 0, \pm1, \pm2, \dots\}$ a d-dimensional time series process, with components $\{X_{t;i}\}_{i=1}^{d}$. The characteristic functions can be defined in analogous way as in the univariate case. In particular, the joint characteristic function of $X_{t;r}$ and $X_{t+j;m}$ is given by

$$\phi_j^{(r,m)}(u,v) = E\Big[\exp\Big(i(uX_{t;r} + vX_{t+j;m})\Big)\Big], \quad j = 0, \pm1, \pm2, \dots$$

and the marginal characteristic functions of $X_{t;r}$ and $X_{t+j;m}$ by $\phi^{(r)}(u) := \phi_j^{(r,m)}(u,0)$ and $\phi^{(m)}(v) := \phi_j^{(r,m)}(0,v)$ respectively, with $(u,v) \in \mathbb{R}^2$, $r, m = 1, \dots d$ and $i^2 = -1$. The pairwise ADCV between $X_{t;r}$ and $X_{t+j;m}$ is denoted by $V_{rm}(j)$ and it is defined as the non-negative square root of

$$V^2_{rm}(j) = \int_{\mathbb{R}^2} \left|\sigma_j^{(r,m)}(u,v)\right|^2 \mathcal{W}(u,v)dudv, \quad j = 0, \pm1, \pm2, \dots$$

where $\mathcal{W}(\cdot,\cdot)$ is given by (3) and $\sigma_j^{(r,m)}(u,v)$ is similarly defined as in the univariate case, namely

$$\sigma_j^{(r,m)}(u,v) = \phi_j^{(r,m)}(u,v) - \phi^{(r)}(u)\phi^{(m)}(v).$$

Clearly, $V^2_{rm}(j) \geq 0, \forall\ j$ and $X_{t;r}$ and $X_{t+j;m}$ are independent if and only if $V^2_{rm}(j) = 0$. The ADCV matrix, $V(j)$, is then defined by

$$V(j) = \Big[V_{rm}(j)\Big]_{r,m=1}^{d}, \quad j = 0, \pm1, \pm2, \dots \tag{7}$$

The pairwise ADCF between $X_{t;r}$ and $X_{t+j;m}$, $R_{rm}(j)$, is a coefficient that lies in the interval $[0,1]$ and also measures dependence and is defined as the positive square root of

$$R^2_{rm}(j) = \frac{V^2_{rm}(j)}{\sqrt{V^2_{rr}(0)}\sqrt{V^2_{mm}(0)}}, \tag{8}$$

for $V_{rr}(0)V_{mm}(0) \neq 0$ and zero otherwise. The ADCF matrix of $\mathbf{X}_t$, is then defined as

$$R(j) = \Big[R_{rm}(j)\Big]_{r,m=1}^{d}, \quad j = 0, \pm1, \pm2, \dots$$

$V_{rm}(j)$ measures the dependence of $X_{t;r}$ on $X_{t+j;m}$. In general, $V_{rm}(j) \neq V_{mr}(j)$ for $r \neq m$, since they measure different dependence structure between the series $\{X_{t;r}\}$ and $\{X_{t;m}\}$ for all $r, m = 1, 2, \dots, d$. Thus, $V(j)$ and $R(j)$ are non-symmetric matrices. Moreover, because of the assumed stationarity and relation $\mathrm{Cov}(x,y) = \mathrm{Cov}(y,x)$, $V(j) = V'(-j)$ and consequently $R(j) = R'(-j)$. More properties of these new defined functions can be found in Fokianos and Pitsillou (2016b).

Estimation of $V^2_{rm}(\cdot)$ can be dealt in a similar way as in the univariate case. In particular, let first $Y_{t;m} \equiv X_{t+j;m}$. Based on the sample $\{(X_{t;r}, Y_{t;m}) : t = 1+j, \dots, n\}$, we define the Euclidean distance matrices by $(a^r_{ts}) = |X_{t;r} - X_{s;r}|$ and $(b^m_{ts}) = |Y_{t;m} - Y_{s;m}|$ and the centered distance matrices by

$$\begin{aligned} A^r_{ts} &= a^r_{ts} - \bar{a}^r_{t.} - \bar{a}^r_{.s} + \bar{a}^r_{..}, \\ B^m_{ts} &= b^m_{ts} - \bar{b}^m_{t.} - \bar{b}^m_{.s} + \bar{b}^m_{..}, \end{aligned}$$

where the quantities in the right hand side are defined analogously as those defined in the univariate case. The biased estimator of $V^2_{rm}(\cdot)$ is then given by

$$\widehat{V}^2_{rm}(j) = \begin{cases} \dfrac{1}{(n-j)^2} \displaystyle\sum_{t,s=1+j}^{n} A^r_{ts} B^m_{ts}, & 0 \le j \le (n-1); \\ \dfrac{1}{(n+j)^2} \displaystyle\sum_{t,s=1}^{n+j} A^r_{ts} B^m_{ts}, & -(n-1) \le j < 0. \end{cases} \tag{9}$$

Analogously to (6), an unbiased estimator of $\widehat{V}^2_{rm}(\cdot)$ is given by

$$\widetilde{V}^2_{rm}(j) = \begin{cases} \dfrac{1}{(n-j)(n-j-3)} \displaystyle\sum_{t,s=1+j}^{n} \widetilde{A}^r_{ts} \widetilde{B}^m_{ts}, & 0 \le j \le (n-1); \\ \dfrac{1}{(n+j)(n+j-3)} \displaystyle\sum_{t,s=1}^{n+j} \widetilde{A}^r_{ts} \widetilde{B}^m_{ts}, & -(n-1) \le j < 0, \end{cases} \tag{10}$$

where $\widetilde{A}^r_{ts}$ are computed appropriately.

The sample ADCV matrix, $\widehat{V}(\cdot)$, is then obtained by replacing its elements by the positive square root of (9) and can be calculated from **dCovTS** using the `mADCV` function. The unbiased estimator of ADCV matrix, $\widetilde{V}(\cdot)$, is obtained from **dCovTS** using the argument `unbiased=TRUE`. The package also gives the sample ADCF matrix $\widehat{R}(\cdot)$ (function `mADCF`) which is obtained after replacing (9) (or (10)) into (8). The distance correlation plots for both univariate and multivariate time series are obtained by the `ADCFplot` and `mADCFplot` functions respectively, where the shown critical values (blue dotted horizontal line) are computed by employing bootstrap methodology described in the appropriate section. Recall that these are computed by using the biased definition of distance covariance and correlation.

Consistency and asymptotic distribution of distance covariance

Consider first the univariate case. For a strictly stationary and α-mixing process X_t, with $E|X_t| < \infty$, then for all $j = 0, \pm 1, \pm 2, \ldots$

$$\widehat{V}^2_X(j) \quad \to \quad V^2_X(j)$$

almost surely, as $n \to \infty$. A detailed proof of this result can be found in Fokianos and Pitsillou (2016a). Under mild conditions, Zhou (2012) obtained the weak consistency of $\widehat{V}^2_X(\cdot)$ and its asymptotic distribution at a fixed lag, but under alternative mixing conditions.

In addition, Fokianos and Pitsillou (2016b) showed that for a d-dimensional strictly stationary and ergodic time series process $\{\mathbf{X}_t\}$ with $E|X_{t;r}| < \infty$ for $r = 1, \ldots, d$, then for all $j = 0, \pm 1, \pm 2, \ldots$

$$\widehat{V}(j) \quad \to \quad V(j)$$

almost surely as $n \to \infty$. Under pairwise independence, the empirical pairwise ADCV is a degenerate V-statistic of order two with a measurable kernel function that is symmetric, continuous and positive semidefinite Then

$$(n-j)\widehat{V}^2_X(j) \quad \to \quad Z := \sum_k \lambda_k Z^2_k \tag{11}$$

in distribution, as $n \to \infty$, where $\{Z_k\}$ is an i.i.d sequence of $N(0,1)$ random variables, and (λ_k) is a sequence of nonzero eigenvalues. A similar result showing the limiting distribution of $\widehat{V}_{rm}(\cdot)$ can be obtained by replacing $\widehat{V}_X(\cdot)$ by $\widehat{V}_{rm}(\cdot)$ in (11).

Testing for pairwise dependence in univariate time series

As shown in the previous section, the asymptotic distribution of distance covariance is derived at a fixed lag, for both univariate and multivariate time series. Fokianos and Pitsillou (2016a,b) constructed the asymptotic behavior of distance covariance considering an increasing number of lags by employing Hong's (1999) generalized spectral domain methodology. Hong (1999) highlighted that standard spectral density approaches become inappropriate for non-Gaussian and nonlinear processes with zero autocorrelation. Considering a univariate strictly stationary α-mixing process, he proposed

the generalized spectral density, which is the Fourier transform of $\sigma_j(u,v)$ defined in (1), given by

$$f(\omega,u,v) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\sigma_j(u,v)e^{-ij\omega}.$$

Under the null hypothesis of independence, the corresponding null density is given by

$$f_0(\omega,u,v) = \frac{1}{2\pi}\sigma_0(u,v), \quad \omega\in[-\pi,\pi].$$

Any deviation of f from f_0 is a strong evidence of pairwise dependence. Thus, Hong (1999) compares the Parzen's (1957) kernel-type estimators $\hat{f}(\omega,u,v)$ and $\hat{f}_0(\omega,u,v)$ via an L_2-norm resulting in a test statistic of the form

$$T_n^{(2)} = \int_{\mathbb{R}^2}\sum_{j=1}^{n-1}(n-j)k^2(j/p)\left|\hat{\sigma}_j(u,v)\right|d\mathcal{W}(u,v), \tag{12}$$

where $\mathcal{W}(\cdot,\cdot):\mathbb{R}^2\to\mathbb{R}$ is an arbitrary nondecreasing function with bounded total variation, p is a bandwidth of the form $p=cn^{\lambda}$ for $c>0$ $\lambda\in(0,1)$ and $k(\cdot)$ is a Lipschitz-continuous kernel function satisfying the following assumption:

Assumption 1 $k:\mathbb{R}\to[-1,1]$ is symmetric and is continuous at 0 and at all but a finite number of points, with $k(0)=1$, $\int_{-\infty}^{\infty}k^2(z)dz<\infty$ and $|k(z)|\leq C\,|z|^{-b}$ for large z and $b>1/2$.

`kernelFun` in **dCovTS** computes a number of such kernel functions including the truncated (default option), Bartlett, Daniell, QS and Parzen kernels.

Fokianos and Pitsillou (2016a) proposed a portmanteau type statistic based on ADCV

$$T_n = \sum_{j=1}^{n-1}(n-j)k^2(j/p)\widehat{V}_X^2(j). \tag{13}$$

Under the null hypothesis that the data are i.i.d and some further assumptions about the kernel function $k(\cdot)$, the standardized version of T_n follows a $N(0,1)$ asymptotically and it is consistent. The authors also considered a similar test statistic based on ADCF

$$\sum_{j=1}^{n-1}(n-j)k^2(j/p)\widehat{R}_X^2(j). \tag{14}$$

The function `UnivTest` from **dCovTS** package performs univariate tests of independence based on (13) and its rescaled version (14), using the arguments `testType="covariance"` and `testType="correlation"` respectively.

Testing for pairwise dependence in multivariate time series

Following a similar methodology described in the previous section, Fokianos and Pitsillou (2016b) suggested a test statistic suitable for testing pairwise independence in a multivariate time series framework. The proposed test statistic is based on the ADCV matrix (7) and it is given by

$$\widetilde{T}_n = \sum_{j=1}^{n-1}(n-j)k^2(j/p)\mathrm{tr}\{\widehat{V}^*(j)\widehat{V}(j)\}. \tag{15}$$

where $k(\cdot)$ is a univariate kernel function satisfying Assumption 1, p is a bandwidth as described before. Moreover, $\widehat{V}^*(\cdot)$ denotes the complex conjugate matrix of $\widehat{V}(\cdot)$ and $\mathrm{tr}(A)$ denotes the trace of the matrix A. The authors formed the statistic (15) in terms of the ADCF matrix as follows

$$\overline{T}_n = \sum_{j=1}^{n-1}(n-j)k^2(j/p)\mathrm{tr}\{\widehat{V}^*(j)\widehat{D}^{-1}\widehat{V}(j)\widehat{D}^{-1}\}, \tag{16}$$

where $D=diag\{V_{rr}(0), r=1,2,\ldots,d\}$. Under the null hypothesis of independence and some further assumptions about the kernel function $k(\cdot)$, the standardized version of the test statistics $\widetilde{T}_n$ and $\overline{T}_n$ given in (15) and (16) were proved to follow $N(0,1)$ asymptotically and they are consistent. The multivariate tests of independence based on $\widetilde{T}_n$ and $\overline{T}_n$ are performed via `mADCVtest` and `mADCFtest` respectively in **dCovTS** package.

Table 1: Functions in **dCovTS**

Function	Description
`ADCF,mADCF`	Estimates distance correlation for a univariate and multivariate time series respectively
`ADCV,mADCV`	Estimates distance covariance for a univariate and multivariate time series respectively
`ADCFplot,mADCFplot`	Plots sample distance correlation in a univariate and multivariate time series framework respectively
`kernelFun`	Gives a range of univariate kernel function, $k(\cdot)$, that satisfy Assumption 1
`UnivTest`	Performs a univariate test of independence based on T_n
`mADCFtest,mADCVtest`	Perform multivariate tests of independence based on $\overline{T}_n$ and $\widetilde{T}_n$ respectively

Bootstrap methodology

To examine the asymptotic behavior of the proposed test statistics, a resampling method is proposed. First, recall that all test statistics T_n, $\widetilde{T}_n$ and $\overline{T}_n$ of equations (13), (15) and (16) respectively, are functions of degenerate V- statistics of order two. Dehling and Mikosch (1994) proposed wild bootstrap techniques to approximate the distribution of degenerate U-statistics for the case of i.i.d data. Recently, Leucht and Neumann (2013a,b) suggested the use of a new variant of dependent wild bootstrap (Shao, 2010) to approximate the limit distribution of degenerate U- and V-statistics for dependent data. The method relies on generating auxiliary random variables $(W_{tn}^*)_{t=1}^{n-j}$. Shao (2010) highlighted that the methodology of wild bootstrap for time series extends that of Wu (1986) by allowing the auxiliary random variables W_{tn}^* to be dependent. In particular, Leucht and Neumann (2013b) proposed to generate the sequence W_{tn}^* by a first order autoregressive model. In the case of independent data, Dehling and Mikosch (1994) studied the wild bootstrap methodology by employing independent auxiliary variables W_{tn}^*. Because our focus is on testing independence we implement the calculation of the test statistics by using W_{tn}^* i.i.d standard normal random variables. Thus, the empirical p-values of the tests are derived based on this methodology.

We also suggest the use of independent wild bootstrap for obtaining simultaneous 95% empirical critical values for the distance correlation plots. In the case of a univariate time series, we additionally propose the subsampling approach suggested by Zhou (2012, Section 5.1) for computing the pairwise 95% critical values (argument `method="Subsampling"`). The choice of the subsampling block size is based on the minimum volatility method proposed by Politis et al. (1999, Section 9.4.2). In addition, the package provides the ordinary independent bootstrap methodology to derive empirical p-values of the tests and simultaneous 95% critical values for the ADCF plots (argument `method="Independent Bootstrap"`). The default bootstrap method provided to the user is the independent wild bootstrap technique.

The computation of the bootstrap replications, and thus the empirical p-values and the critical values, can be distributed to multiple cores simultaneously (argument `parallel=TRUE`). To do this, the **doParallel** (Analytics and Weston, 2014) package needs to be installed first, in order to register a computing cluster.

Implementation of dCovTS package

The current version of **dCovTS** package (version number 1.1) is available from CRAN and can be downloaded via `https://cran.r-project.org/web/packages/dCovTS/`. The aim of the **dCovTS** package is to provide a set of functions that compute and plot distance covariance and correlation functions in both univariate and multivariate time series. As we mentioned, the package supports both versions of biased and unbiased estimators of distance covariance and correlation functions. Moreover, it offers functions that perform univariate and multivariate tests of independence based on distance covariance function using the biased estimator (corresponding to (5) and (9)). All these functions are provided in Table 1. Apart from these functions, the package also provides two real datasets listed in Table 2. A more detailed description of the functions and datasets can be found in the help files. We apply **dCovTS** to two real data examples.

Table 2: Datasets in **dCovTS**

Data	Description
`ibmSp500`	Monthly returns of IBM and S&P 500 composite index from January 1926 to December 2011
`MortTempPart`	Mortality, temperature and pollution data measured daily in Los Angeles County over the period 1970-1979

Regression with autocorrelated errors

We first consider the pollution, temperature and mortality data measured daily in Los Angeles County over the 10 year period 1970-1979 (Shumway et al., 1988). The data are available in our package by the argument `MortTempPart` and contain 508 observations and 3 variables representing the mortality (cmort), temperature (tempr) and pollutant particulates (part) data.

```
> library(dCovTS)
> data(MortTempPart)
> MortTempPart[1:10,] # the first ten observations
    cmort tempr  part
1   97.85 72.38 72.72
2  104.64 67.19 49.60
3   94.36 62.94 55.68
4   98.05 72.49 55.16
5   95.85 74.25 66.02
6   95.98 67.88 44.01
7   88.63 74.20 47.83
8   90.85 74.88 43.60
9   92.06 64.17 24.99
10  88.75 67.09 40.41
> attach(MortTempPart)
```

Following the analysis of Shumway and Stoffer (2011), the possible effects of temperature (T_t) and pollutant particulates (P_t) on daily cardiovascular mortality (M_t) are examined via regression. In particular, once the temperature is adjusted for its mean ($T_\cdot = 74.3$), we fit the following regression model using the function `lm`

$$\begin{aligned} \widehat{M}_t &= 2831.49 - 1.396_{(0.101)}t - 0.472_{(0.032)}(T_t - T_\cdot) \\ &+ 0.023_{(0.003)}(T_t - T_\cdot)^2 + 0.255_{(0.019)}P_t, \end{aligned} \tag{17}$$

where the standard errors of the estimators are given in parentheses. Figure 1 provides the sample autocorrelation (ACF), partial correlation (PACF) and ADCF plots of the residuals of model (17). The plots shown in Figure 1 suggest an AR(2) process for the residuals. The new fit is

$$\begin{aligned} \widehat{M}_t &= 3075.15 - 1.517_{(0.423)}t - 0.019_{(0.050)}(T_t - T_\cdot) \\ &+ 0.015_{(0.002)}(T_t - T_\cdot)^2 + 0.155_{(0.027)}P_t, \end{aligned} \tag{18}$$

where the standard errors of the estimators are given in parentheses. The above model fit was derived by using the `arima` function of R. The correlation plots for the residuals from the new model (18) are shown in Figure 2 indicating that there is no serial dependence. The calls for both model fits and their diagnostic plots are given below. ADCF plots (lower plots of Figures 1 and 2) are constructed using both resampling schemes explained in the previous section: independent wild bootstrap (with $b = 499$ replications) and Subsampling.

```
> temp  = tempr-mean(tempr)  # center temperature
> temp2 = temp^2
> trend = time(cmort)
> fit = lm(cmort~ trend + temp + temp2 + part, na.action=NULL)
> Residuals <- as.numeric(resid(fit))
> ##Correlation plots
> acf(Residuals,lag.max=18,main="")
> pacf(Residuals,lag.max=18,main="")
> ADCFplot(Residuals,MaxLag=18,main="Wild Bootstrap",method="Wild")
> ADCFplot(Residuals,MaxLag=18,main="Subsampling",method="Subsampling")
```

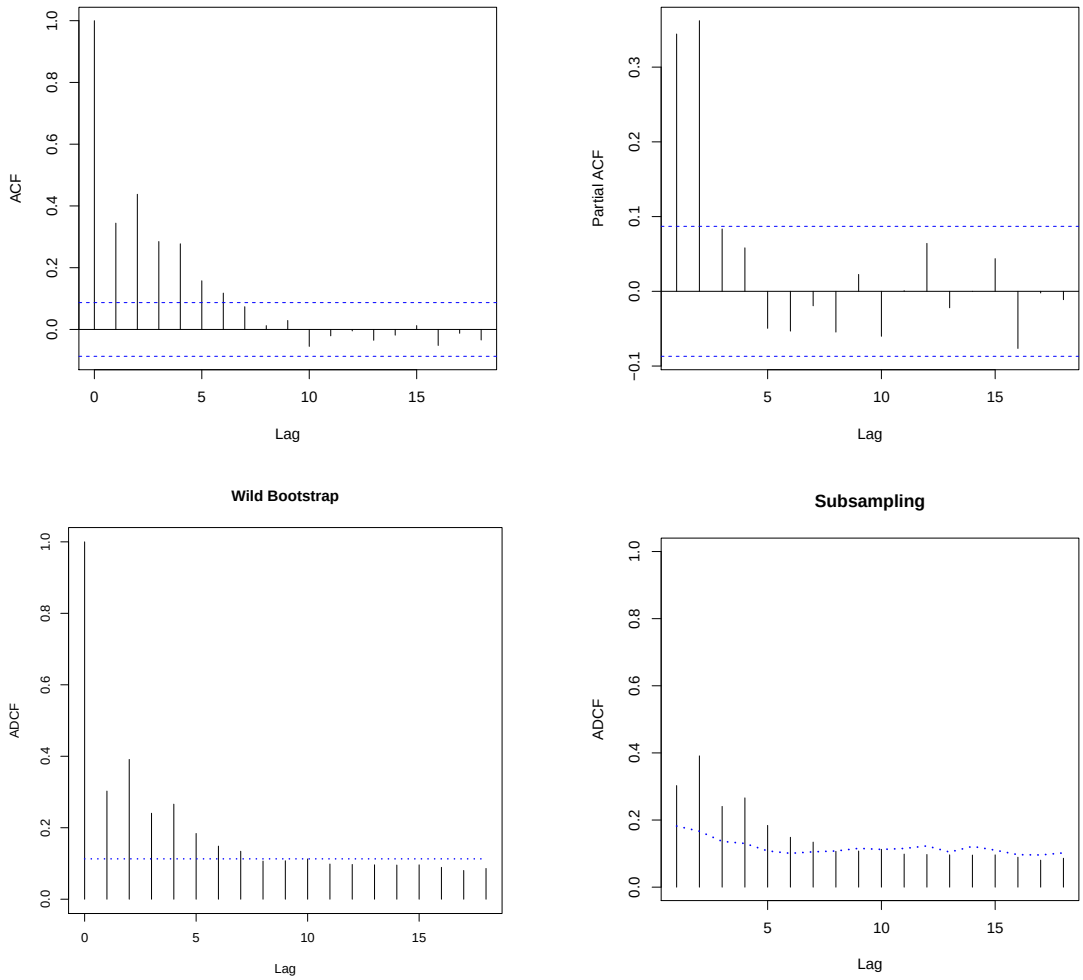

Figure 1: Sample ACF, PACF and ADCF plots of the mortality residuals of model (17).

```
> fit2 <- arima(cmort, order=c(2,0,0), xreg=cbind(trend,temp,temp2,part))
> Residuals2 <- as.numeric(residuals(fit2))
> ##Correlation plots
> acf(Residuals2,lag.max=18,main="")
> pacf(Residuals2,lag.max=18,main="")
> ADCFplot(Residuals2,MaxLag=18,main="Wild Bootstrap",method="Wild")
> ADCFplot(Residuals2,MaxLag=18,main="Subsampling",method="Subsampling")
```

To formally confirm the absence of any serial dependence among the new residuals of model (18), as shown in Figure 2, we perform univariate tests of independence based on the test statistic T_n given in (13). We use `UnivTest` function from our package with argument `testType="covariance"` (default option). In order to examine the effect of using different bandwidths, we choose $p = [3n^{\lambda}]$ for λ=0.1, 0.2 and 0.3, that is $p = 6$, 11, and 20 and we apply Bartlett kernel. The resulting p-values are 0.118, 0.170 and 0.208 respectively suggesting acceptance of independence. P-values are calculated for $b = 499$ independent wild bootstrap replications. Bootstrap procedure can be computed on multiple cores simultaneously by using the argument `parallel=TRUE` (they take about 10, 14 and 23 seconds respectively on a standard laptop with Intel Core i5 system and CPU 2.30 GHz):

```
> UnivTest(Residuals2, type="bartlett", p=6, b=499, parallel=TRUE)

        Univariate test of independence based on distance covariance

data:  Residuals2, kernel type: bartlett, bandwidth=6, boot replicates 499
Tn = 67.7344, p-value = 0.118
```

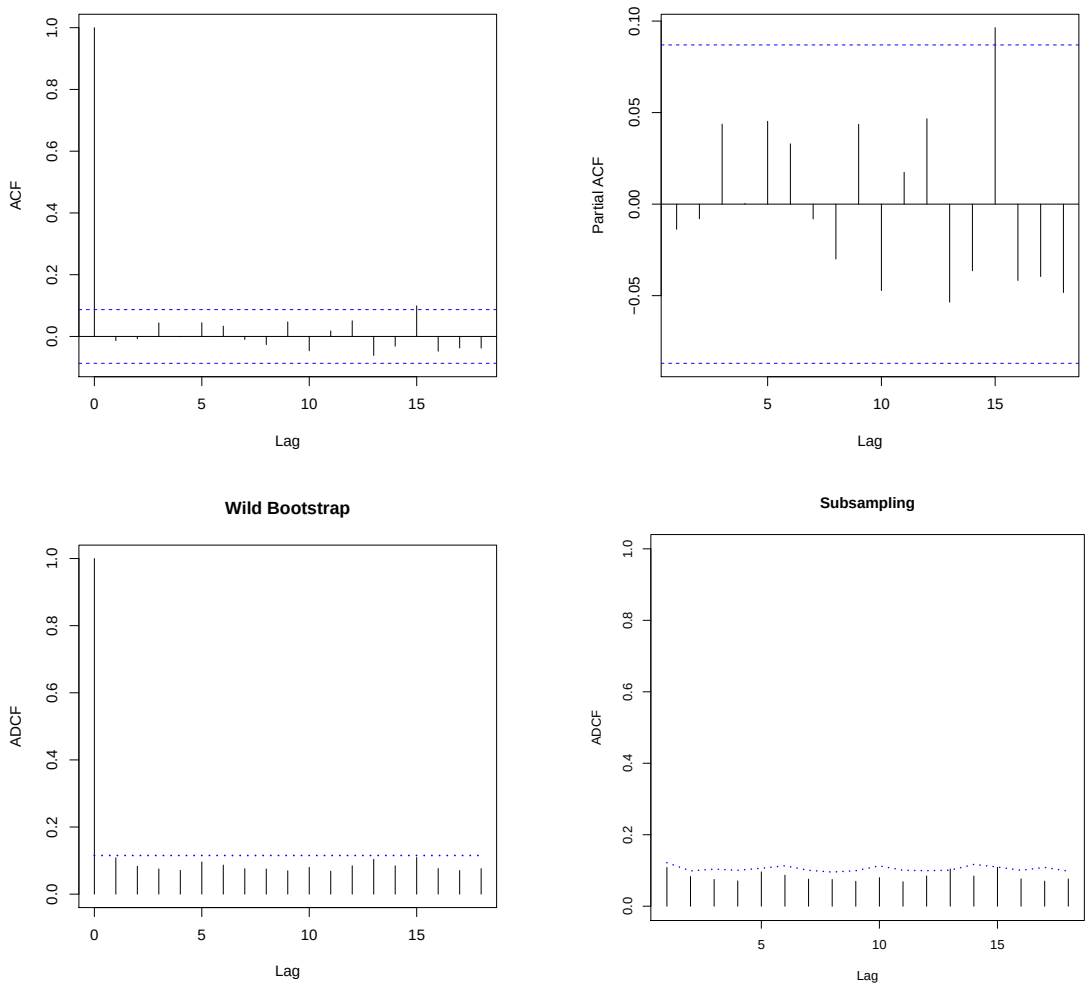

Figure 2: Sample ACF, PACF and ADCF plots of the mortality residuals of model (18) indicating that the new residuals can be taken as white noise.

```
> UnivTest(Residuals2, type="bartlett", p=11, b=499, parallel=TRUE)

        Univariate test of independence based on distance covariance

data:  Residuals2, kernel type: bartlett, bandwidth=11, boot replicates 499
Tn = 125.6674, p-value = 0.170

> UnivTest(Residuals2, type="bartlett", p=20, b=499, parallel=TRUE)

        Univariate test of independence based on distance covariance

data:  Residuals2, kernel type: bartlett, bandwidth=20, boot replicates 499
Tn = 225.9266, p-value = 0.208
```

We compare the proposed test statistic with other test statistics to check its performance. In particular, we consider the Box-Pierce (Box and Pierce, 1970) test statistic

$$BP = n \sum_{j=1}^{p} \hat{\rho}^2(j),$$

the Ljung-Box (Ljung and Box, 1978) test statistic

$$LB = n(n+2)\sum_{j=1}^{p}(n-j)^{-1}\hat{\rho}^2(j),$$

the test statistic proposed by Hong (1996)

$$T_n^{(1)} = n\sum_{j=1}^{n-1}k^2(j/p)\hat{\rho}^2(j)$$

and the test statistic $T_n^{(2)}$ proposed by Hong (1999) defined in (12) with $\mathcal{W}(u,v) = \Phi(u)\Phi(v)$, $\Phi(\cdot)$ being the cumulative distribution function of standard normal. For the aforestated bandwidth values, all these alternative test statistic give large p-values indicating the absence of any serial dependence among the new residuals. More precisely, BP and LB give 0.848, 0.906, 0.170 and 0.844, 0.901, 0.142 respectively. BP and LB based tests are performed in R by the function `Box.test` as follows:

```
> box1 <-  Box.test(Residuals2,lag=6)
> box2 <-  Box.test(Residuals2,lag=11)
> box3 <-  Box.test(Residuals2,lag=20)
> ljung1 <-  Box.test(Residuals2,lag=6,type="Ljung")
> ljung2 <-  Box.test(Residuals2,lag=11,type="Ljung")
> ljung3 <-  Box.test(Residuals2,lag=20,type="Ljung")
```

The p-values obtained by $T_n^{(1)}$ are 0.896, 0.930 and 0.870 respectively. $T_n^{(2)}$ gives the following p-values: 0.854, 0.752 and 0.504 respectively. $T_n^{(1)}$ and $T_n^{(2)}$ are calculated by employing the Bartlett kernel. These p-values are calculated for $b = 499$ ordinary bootstrap replications. The R functions for constructing these test statistics are beyond the scope of this paper and are available from the authors upon request.

Bivariate financial time series

We now analyze the monthly log returns of the stocks of International Business Machines (IBM) and the S&P 500 composite index starting from 30 September 1953 to 30 December 2011 for 700 observations. A larger dataset is available in our package by the object `ibmSp500` starting from January 1926 for 1032 observations. It is actually a combination of two smaller datasets: the first one was first reported by Tsay (2010) and the second one was first reported by Tsay (2014). ACF and ADCF plots of the original series are provided in Figure 3, whereas Figure 4 shows the ACF and ADCF plots of the squared series.

The R commands for constructing these plots are as follows:

```
> data(ibmSp500)
> new_data <- tail(ibmSp500[,2:3],700)
> lseries <- log(new_data+1)
> at=scale(lseries,center=T,scale=F)
> at2 <- at^2
> colnames(at) <- c("IBM","SP")
> colnames(at2) <- c("IBM_sq","SP_sq")
> acf(at,lag.max=18)
> acf(at2,lag.max=18)
> mADCFplot(at,MaxLag=18,ylim=c(0,0.2))
> mADCFplot(at2,MaxLag=18,ylim=c(0,0.2))
```

The ACF plots of the original series (upper panel of Figure 3) suggest no serial correlation among observations, while the ACF plots of the squared series (upper panel of Figure 4) imply strong dependence. This confirm the conditional heteroscedasticity in the monthly log returns. However, the ADCF plots for both original and squared series (lower panels of Figures 3 and 4) suggest dependence. Indeed, choosing $p = [3n^{\lambda}]$ for λ=0.1, 0.2 and 0.3, that is $p = 6$, 12 and 22, and employing Bartlett kernel, $\overline{T}_n$ gives low p-values (0.022, 0.014 and 0.020 respectively). The calls for these three multivariate tests of independence can be found below (they take about 2, 3 and 6 minutes respectively for $b = 499$ bootstrap replications on a standard laptop with Intel Core i5 system and CPU 2.30 GHz):

```
> mADCFtest(at,"bartlett",p=6,b=499,parallel=TRUE)

 Multivariate test of independence based on distance correlation
```

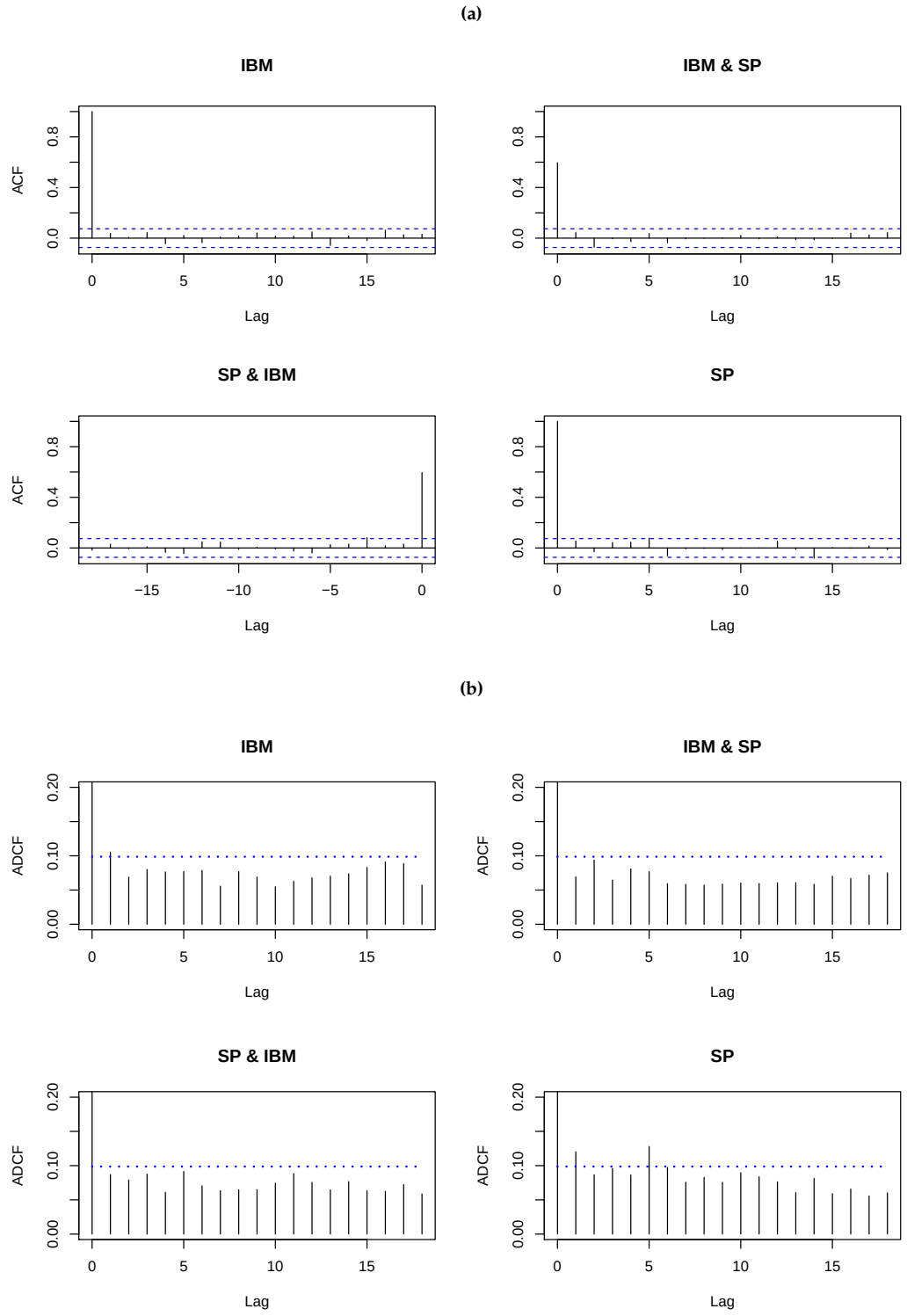

Figure 3: (a) The sample ACF of the original series. (b) The sample ADCF of the original series.

```
data:  at, kernel type: bartlett, bandwidth=6, boot replicates 499
Tnbar = 34.1743, p-value = 0.022

> mADCFtest(at,"bartlett",p=12,b=499,parallel=TRUE)

        Multivariate test of independence based on distance correlation

data:  at, kernel type: bartlett, bandwidth=12, boot replicates 499
```

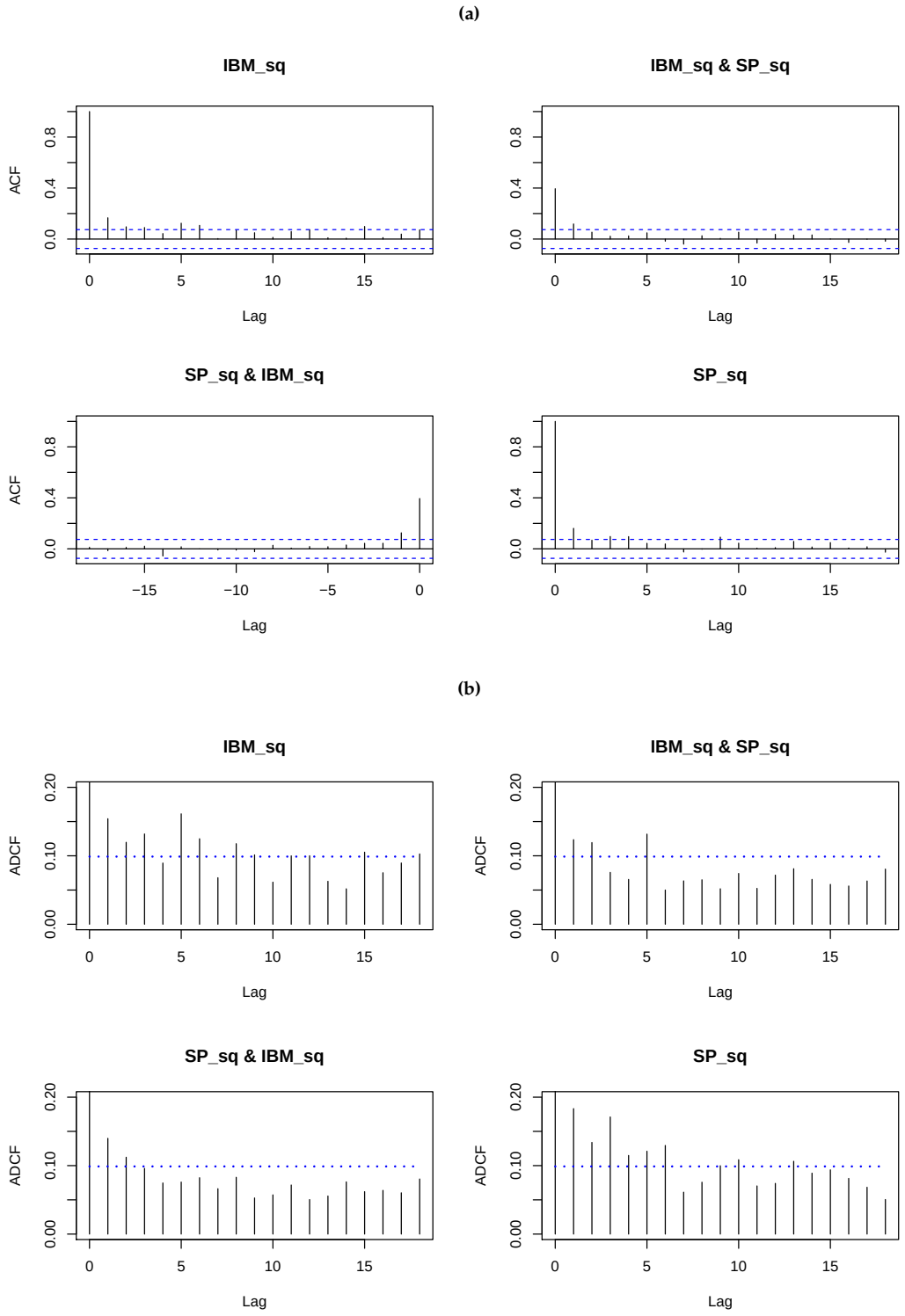

Figure 4: (a) The sample ACF of the squared series. (b) The sample ADCF of the squared series.

```
Tnbar = 71.1713, p-value = 0.014

> mADCFtest(at,"bartlett",p=22,b=499,parallel=TRUE)

        Multivariate test of independence based on distance correlation

data:  at, kernel type: bartlett, bandwidth=22, boot replicates 499
Tnbar = 122.9424, p-value = 0.02
```

To compare the performance of the proposed test statistic $\overline{T}_n$, we consider the multivariate Ljung-Box statistic (Hosking, 1980) defined by:

$$mLB = n^2 \sum_{j=1}^{p} (n-j)^{-1} \text{tr}\{\widehat{\Gamma}'(j)\widehat{\Gamma}^{-1}(0)\widehat{\Gamma}(j)\widehat{\Gamma}^{-1}(0)\}$$

where $\widehat{\Gamma}(\cdot)$ is the ordinary covariance matrix. In contrast to the $\overline{M}_n$'s results, mLB gives large p-values (0.218, 0.731 and 0.525) respectively. The **portes** (Mahdi and McLeod, 2012) package needs to be installed in order to perform tests of independence based on mLB statistic:

```
> library(portes)
> LjungBox(at,c(6,12,22))
```

Assuming that the bivariate log returns follows a VAR model and employing the AIC to choose its best order, we obtain that a VAR(2) model fits well the data. Figure 5 shows the ACF plots (upper panel) and ADCF plots (lower panel) of the residuals after fitting a VAR(2) model to the original bivariate log return series using the function `VAR` from the **MTS** (Tsay, 2015) package. In contrast to the ACF plot, the ADCF plot still indicates some dependence among the residuals. Constructing tests of independence based on $\overline{T}_n$ and mLB for the same choices of bandwidth, p =6, 12, 22, we confirm this visual result. In particular, employing a Bartlett kernel, $\overline{T}_n$ statistic gives low p-values (0.036, 0.018 and 0.034 respectively) whereas the mLB statistic yields large p-values (0.669, 0.958 and 0.806 respectively). The calls for the plots of Figure 5 and the corresponding tests of independence are as follows:

```
> library(MTS)
> model <- VAR(at,2)
> resids <- residuals(model)
> colnames(resids) <- c("IBM_res","SP_res")
> windows(9,6)
> acf(resids,lag.max=18)
> mADCFplot(resids,MaxLag=18,ylim=c(0,0.13))

> ## Tests of independence based on \overline{T}_n
> mADCFtest(resids,"bartlett",p=6,b=499,parallel=TRUE)

        Multivariate test of independence based on distance correlation

data:  resids, kernel type: bartlett, bandwidth=6, boot replicates 499
Tnbar = 29.9114, p-value = 0.036

> mADCFtest(resids,"bartlett",p=12,b=499,parallel=TRUE)

        Multivariate test of independence based on distance correlation

data:  resids, kernel type: bartlett, bandwidth=12, boot replicates 499
Tnbar = 64.7754, p-value = 0.018

> mADCFtest(resids,"bartlett",p=22,b=499,parallel=TRUE)

        Multivariate test of independence based on distance correlation

data:  resids, kernel type: bartlett, bandwidth=22, boot replicates 499
Tnbar = 115.3462, p-value = 0.034

> ## Tests of independence based on mLB
> LjungBox(resids,c(6,12,22))
```

Summary and further research

There have been many works in the literature based on Székely et al.'s (2007) distance covariance methodology. The R package **energy** (Rizzo and Székely, 2013), provides functions that cover this methodology. However, there is no published package that includes functions about distance covariance for time series data. **dCovTS** contributes to filling this gap by providing functions that compute

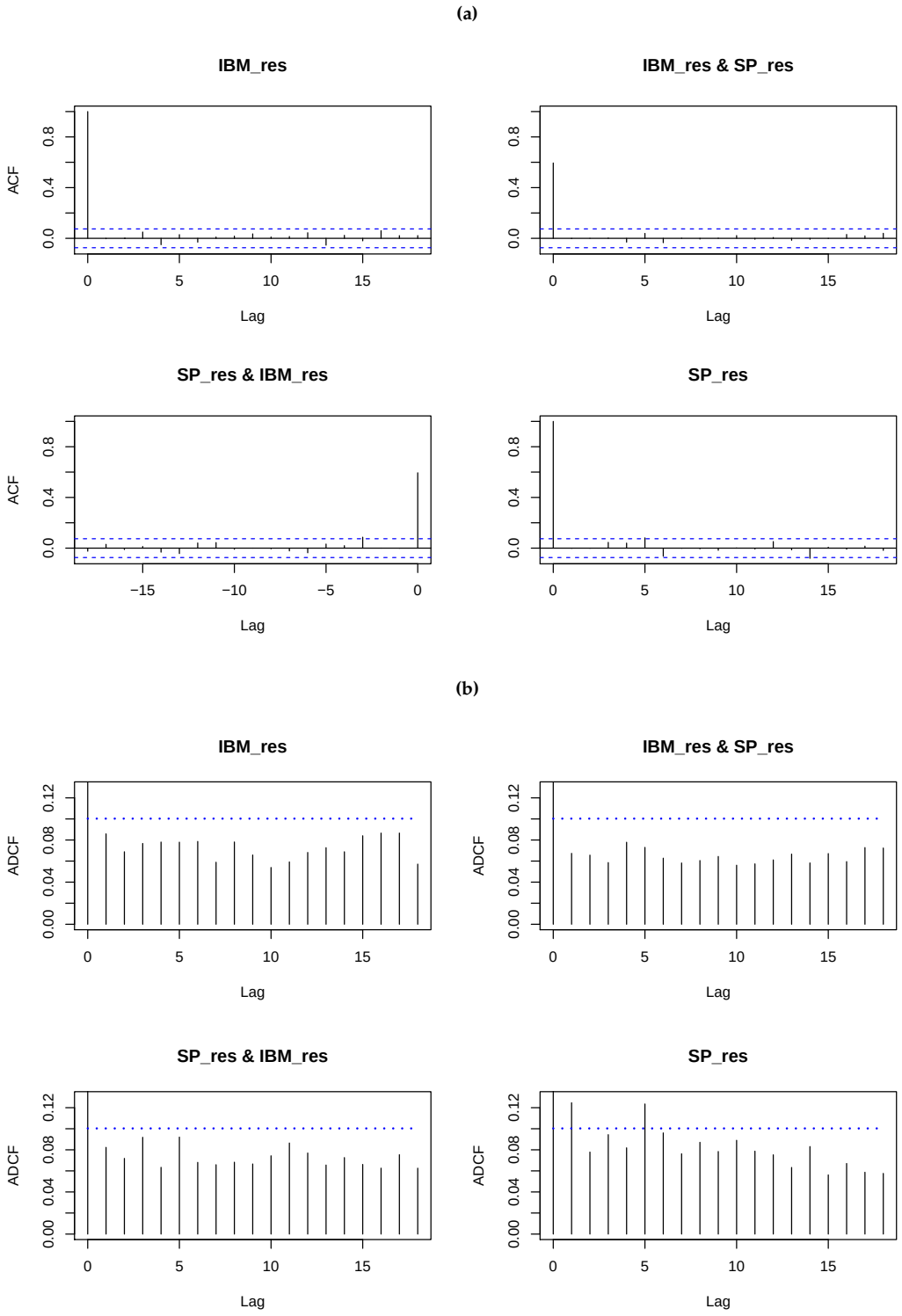

Figure 5: The sample ACF (upper panel) and sample ADCF (lower panel) of the residuals after fitting VAR(2) model to the original series.

distance covariance and correlation functions for both univariate and multivariate time series. We also include functions that develop univariate and multivariate tests of serial dependence based on distance covariance and correlation functions.

There is a number of possible extensions of this package that some of them are not covered by existing theory and can be seen as further research. One possible direction is to develop a theory based on partial ADCV or conditional ADCV and a related testing methodology to identify possible dependencies among time series (see Székely and Rizzo (2014) for partial distance covariance methodology

and Poczos and Schneider (2012), Wang et al. (2015) for conditional distance covariance methodology; all three works deal with independent random variables). Among the many applications of partial correlation are graphical models. Thus, a graphical modeling theory based on partial ADCV could be carried out and this methodology can be included for a future version of this package.

Acknowledgments

The authors thank Tobias Liboschik for his considerable help on the development of this package. The authors would also like to thank Dominic Edelmann for carefully checking the package and making helpful comments and suggestions for its improvement. In addition, we would like to extend our gratitude to R. Bivand and to an anonymous reviewer whose comments improved our original submission.

Bibliography

R. Analytics and S. Weston. *doParallel: Foreach parallel adaptor for the parallel package*, 2014. URL http://CRAN.R-project.org/package=doParallel. R package version 1.0.8. [p6]

G. E. P. Box and D. A. Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65:1509–1526, 1970. [p9]

A. Feuerverger. A consistent test for bivariate dependence. *International Statistical Review*, 61:419–433, 1993. [p1]

K. Fokianos and M. Pitsillou. Consistent testing for pairwise dependence in time series. *Technometrics*, 2016a. http://dx.doi.org/10.1080/00401706.2016.1156024. [p1, 4, 5]

K. Fokianos and M. Pitsillou. On multivariate auto-distance covariance and correlation functions. Submitted for publication, 2016b. [p1, 3, 4, 5]

A. Gretton, K. Fukumizu, and B. K. Sriperumbudur. Discussion of: Brownian distance covariance. *The Annals of Applied Statistics*, 3:1285–1294, 2009. [p1]

J. Dueck, D. Edelmann, T. Gneiting, and D. Richards. The affinely invariant distance correlation. *Bernoulli*, 20:2305–2330, 2014. [p1]

R. A. Davis, M. Matsui, T. Mikosch, and P. Wan. Applications of distance correlation to time series. http://arxiv.org/abs/1606.05481, 2016. [p1]

H. Dehling and T. Mikosch. Random quadratic forms and the boostrap for U-Statistics. *Journal of Multivariate Analysis*, 51:392–413, 1994. [p1, 6]

C. F. J. Wu. Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, 14:1261–1295, 1986. [p10]

Y. Hong. Consistent testing for serial correlation of unkown form. *Econometrica*, 64:837–864, 1996. [p6]

Y. Hong. Hypothesis testing in time series via the emprical characteristic function: A generalized spectral density approach. *Journal of the American Statistical Association*, 94:1201–1220, 1999. [p1, 2, 4, 5, 10]

J. R. M. Hosking. Multivariate portmanteau statistic. *Journal of the American Statistical Association*, 75: 349–386, 1980. [p13]

J. Josse and S. Holmes. Tests of independence and beyond. 2014. URL http://arxiv.org/pdf/1307.7383.pdf. [p1]

A. Leucht and M. H. Neumann. Degenerate U- and V-statistics under ergodicity: asymptotics, bootstrap and applications in statistics. *Annals of the Institute of Statistical Mathematics*, 65:349–386, 2013a. [p6]

A. Leucht and M. H. Neumann. Dependent wild bootstrap for degenerate U- and V- statistics. *Journal of Multivariate Analysis*, 117:257–280, 2013b. [p1, 6]

G. M. Ljung and G. E. P. Box. On a measure of lack of fit in time series models. *Biometrika*, 62:297–303, 1978. [p10]

E. Mahdi and A. I. McLeod. Improved multivariate portmanteau diagnostic test. *Journal of Time Series Analysis*, 33, 2012. [p13]

E. Parzen. On consistent estimates of the spectrum of a stationary time series. *Annals of Mathematical Statistics*, 28:329–348, 1957. [p5]

B. Poczos and J. Schneider. Conditional distance variance and correlation. 2012. URL http://www.cs.cmu.edu/~bapoczos/articles/poczos12distancecorr.pdf. [p15]

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/. [p1]

B. Remillard. Discussion of: Brownian distane covariance. *The Annals of Applied Statistics*, 3:1295–1298, 2009. [p1]

M. L. Rizzo and G. J. Székely. *energy: E-statistics (energy statistics)*, 2013. URL http://CRAN.R-project.org/package=energy. R package version 1.5.0. [p1, 13]

X. Shao. The dependent wild bootstrap. *Journal of the American Statistical Association*, 105:218–235, 2010. [p1, 6]

R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications With R Examples*. Springer, 2011. Third Edition. [p7]

N. P. Politis, J. P. Romano and M. Wolf. *Subsampling*. Springer, 1999. [p6]

R. H. Shumway, R. S. Azari, and Y. Pawitan. Modeling mortality fluctuations in los angeles as functions of pollution and weather effects. *Environmetnal research*, 45:224–241, 1988. [p7]

G. J. Székely and M. L. Rizzo. Energy statistics:a class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143:1249–1272, 2013. [p]

G. J. Székely and M. L. Rizzo. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42:2382–2412, 2014. [p2, 14]

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35:2769–2794, 2007. [p1, 2, 13]

R. S. Tsay. *Analysis of Financial Time Series*. Wiley Series in Probability and Statistics, Hoboken, NJ, 2010. Third Edition. [p10]

R. S. Tsay. *Multivariate Time Series Analysis With R and Financial Applications*. Wiley, Hoboken, NJ, 2014. [p10]

R. S. Tsay. *MTS: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models*, 2015. URL http://CRAN.R-project.org/package=MTS. R package version 0.33. [p13]

X. Wang, W. Pan, W. Hu, Y. Tian, and H. Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 2015. DOI: 10.1080/01621459.2014.993081. [p15]

Z. Zhou. Measuring nonlinear dependence in time-series, a distance correlation approach. *Journal of Time Series Analysis*, 33:438–457, 2012. [p1, 2, 6]

Maria Pitsillou
Department of Mathematics & Statistics
University of Cyprus
Cyprus
pitsillou.maria@ucy.ac.cy

Konstantinos Fokianos
Department of Mathematics & Statistics
University of Cyprus
Cyprus
fokianos@ucy.ac.cy