Methodol Comput Appl Probab (2010) 12:749–762
DOI 10.1007/s11009-010-9180-0

Spectrum-Based Comparison of Stationary Multivariate Time Series

Nalini Ravishanker · J. R. M. Hosking · Jaydip Mukhopadhyay

Received: 27 April 2009 / Revised: 29 March 2010 /
Accepted: 8 April 2010 / Published online: 28 April 2010

Abstract The problem of comparison of several multivariate time series via their spectral properties is discussed. A pairwise comparison between two independent multivariate stationary time series via a likelihood ratio test based on the estimated cross-spectra of the series yields a quasi-distance between the series. A hierarchical clustering algorithm is then employed to compare several time series given the quasi-distance matrix. For use in situations where components of the multivariate time series are measured in different units of scale, a modified quasi-distance based on a profile likelihood based estimation of the scale parameter is described. The approach is illustrated using simulated data and data on daily temperatures and precipitations at multiple locations. A comparison between hierarchical clustering based on the likelihood ratio test quasi-distance and a quasi-distance described in Kakizawa et al. (J Am Stat Assoc 93:328–340, 1998) is interesting.

Keywords Hierarchical clustering · Likelihood ratio · Periodogram matrix · Quasi-distance · Spectral density matrix

AMS 2000 Subject Classification 37M10 · 62M15 · 62H30

N. Ravishanker (✉)
Department of Statistics, University of Connecticut, Storrs, CT 06269, USA
e-mail: nalini.ravishanker@uconn.edu

J. R. M. Hosking
IBM Research Division, Yorktown Heights, NY 10598, USA

J. Mukhopadhyay
Bristol Myers Squibb, Wallingford, CT 06492, USA

1 Introduction

This article describes a spectrum-based method for comparing linear, stationary multivariate time series. Coates and Diggle (1986) employed the maximum and minimum periodogram ordinate tests of equality of the underlying spectra for two independent univariate Gaussian stationary time series. Kakizawa et al. (1998) used the Kullback–Leibler and Chernoff information as disparity measures for comparing several stationary multivariate time series, with an application to earthquake-explosion discrimination. In this paper, an approach based on a likelihood ratio test on spectral matrices is investigated in order to cluster several continuous-valued stationary multivariate time series.

In general, clustering can use alternate dissimilarity measures, and the outcome of the clustering will reveal groupings of the time series that reflect the dissimilarity measure. Clustering is often the first step in model building for multivariate time series (see, for example, Pai et al. (1994), for a fully Bayesian modeling of multiple time series). For instance, suppose we wish to build a vector autoregressive fractionally integrated moving average (VARFIMA) model to the daily temperatures observed at L locations. In a post-cluster analysis, knowledge of which time series belongs to the jth cluster, for $j = 1, \cdots, J$, could enable us to achieve parsimony in the model building by imposing a simpler structure on the $L \times L$ AR or MA coefficient matrices, by specifying similar coefficients for similar time series. Such simplification would afford considerable saving in computational time and would ensure better convergence of parameter estimates, especially when L is large. Some simplicity in the structure of the $L \times L$ error covariance matrix may also be achieved. For instance, the $L \times L$ error covariance matrix, $\mathbf{\Sigma}_\varepsilon$ may be viewed as a block matrix $\{\mathbf{\Sigma}_{ab}\}$ for $a, b = 1, \cdots, J$. From post-cluster analyses, we can obtain information to set similar elements in the ath diagonal sub-blocks, corresponding to series within the ath cluster. Also, by carrying out a post-cluster cross-spectral analysis between series in different clusters, we may get information for simplifying the off-diagonal blocks of $\mathbf{\Sigma}_\varepsilon$, such as setting them to $\mathbf{0}$.

Our approach uses only the spectrum, a second-order property of time series, and ignores first-order moments. It is therefore appropriate for applications where second-order properties are of prime importance. The earthquake-explosion discrimination of Kakizawa et al. (1998) is an example. Other fields also provide motivating examples. In financial portfolio management, an analyst might wish to cluster thousands of stocks into different groups based on time series data on different characteristics which describe features such as depth of coverage and liquidity, value-growth status, etc. It would be useful to compare such a statistical grouping with the financial analyst's grouping based on certain market definers. Another example in marketing would be for clustering of products based on time series of volume of sales and prices of consumer products sold in a particular region. In this case, spectral clustering would take into account the frequency and depth of price reductions as well as the speed of reactions to sales volume to price shocks, in a way applies equally to high-volume and low-volume products. Finally, spectral clustering of environmental data is appropriate when the focus of interest is on the similarity of periodicities in the data, e.g. relative strengths of annual, subannual, and long-term variations, rather than on the overall level of the measurements at different sites.

In Section 2, a quasi-distance based on a likelihood ratio test for the comparison of pairwise spectra is derived, using smoothed periodogram matrices as estimates. Section 3 discusses use of a hierarchical clustering algorithm using such quasi-distances as input. In Section 4, the approach is illustrated using simulated time series and daily temperatures and precipitations at several locations in the Pacific Northwest. Section 5 contains some concluding remarks.

2 Likelihood and Profile Likelihood Ratio Tests for Spectral Density Matrices

A likelihood ratio test statistic to compare the spectra of pairwise time series is derived. In Section 2.1, this approach is described using the smoothed periodogram estimates of the spectral matrices for time series whose components are measured on the same scale. In Section 2.2, a profile likelihood is used to extend the testing scenario to time series whose components are possibly measured on different scales.

2.1 Likelihood Ratio Test Based on Smoothed Periodogram

Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T$ and $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T$ denote T observations from two independent p-dimensional stationary processes $\{\boldsymbol{X}_t\}$ and $\{\boldsymbol{Y}_t\}$, let $\boldsymbol{\mu}_{\boldsymbol{x}}$ and $\boldsymbol{\mu}_{\boldsymbol{y}}$ denote their respective means, let $\boldsymbol{\gamma}_{\boldsymbol{x}}(r)$ and $\boldsymbol{\gamma}_{\boldsymbol{y}}(r)$ denote their respective autocovariance matrices at lag r, and let $\boldsymbol{\lambda}_{\boldsymbol{x}}(\omega_l) = \{\lambda_{\boldsymbol{x},jk}(\omega_l)\}$ and $\boldsymbol{\lambda}_{\boldsymbol{y}}(\omega_l) = \{\lambda_{\boldsymbol{y},jk}(\omega_l)\}$ denote their respective spectral density matrices at Fourier frequency ω_l where $\omega_l = 2\pi l/T$, $l = -[(T-1)/2], \cdots, [T/2]$. The components of the spectral matrices are defined as

$$\lambda_{\boldsymbol{x},jk}(\omega) = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} e^{-ir\omega} \gamma_{\boldsymbol{x},jk}(r), \quad j,k = 1, \cdots, p, \tag{1}$$

and

$$\lambda_{\boldsymbol{y},jk}(\omega) = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} e^{-ir\omega} \gamma_{\boldsymbol{y},jk}(r), \quad j,k = 1, \cdots, p, \tag{2}$$

with $i = \sqrt{-1}$.

An approach to assess whether the two series are similar is based on a test of the equality of their spectral density matrices, i.e., a test of

$$H_0 : \boldsymbol{\lambda}_{\boldsymbol{x}}(\omega) = \boldsymbol{\lambda}_{\boldsymbol{y}}(\omega) \qquad \forall\, \omega \in (-\pi, \pi] \setminus \{0\}. \tag{3}$$

The $p \times p$ raw periodogram matrix of $\{\boldsymbol{X}_t\}$ at the Fourier frequency ω_l, $l = -[(T-1)/2], \cdots, [T/2]$, is $\boldsymbol{I}_{\boldsymbol{x}}(\omega_l) = \{I_{\boldsymbol{x},jk}(\omega_l)\}$, where

$$I_{\boldsymbol{x},jk}(\omega_l) = \frac{1}{2\pi T} \left(\sum_{t=1}^{T} x_{tj} \exp(-it\omega_l) \right) \left(\sum_{t=1}^{T} x_{tk} \exp(it\omega_l) \right). \tag{4}$$

$I_{\boldsymbol{x},jj}(\omega_l)$ is the estimate of the direct spectrum of jth component of $\{\boldsymbol{X}_t\}$ and $I_{\boldsymbol{x},jk}(\omega_l)$ is the estimate of cross spectrum of the jth and kth components. The raw periodogram of $\{\boldsymbol{Y}_t\}$ is similarly defined.

It is well known that although the raw periodogram matrix is asymptotically unbiased, it is not a consistent estimator of the spectral density matrix. The smoothed

periodogram estimate using the Daniell window at ω_l, the average of the raw periodogram ordinates in a neighborhood of ω_l, is (Brockwell and Davis 1991)

$$\boldsymbol{I}_{\boldsymbol{x}}^{\star}(\omega_l) = (2m+1)^{-1} \sum_{|k|\le m} \boldsymbol{I}_{\boldsymbol{x}}(\omega_{l+k}) \tag{5}$$

for $\omega_l \in (-\pi, \pi) \setminus \{0\}$, and m is a fixed positive integer. The limiting distribution of the smoothed periodogram at Fourier frequency ω_j is related to a complex Wishart distribution with $(2m+1)$ degrees of freedom and scale parameter $\boldsymbol{\lambda}(\omega_j)$ (Brillinger 1986, Theorem 7.3.3, p. 245), provided the time series are strictly stationary with all moments finite. Specifically, the asymptotic distributions of the smoothed periodograms of $\{\boldsymbol{X}_t\}$ and $\{\boldsymbol{Y}_t\}$ are $\boldsymbol{I}_{\boldsymbol{x}}^{\star}(\omega) \xrightarrow{d} (2m+1)^{-1} W_c(2m+1, p, \boldsymbol{\lambda}_x(\omega))$ and $\boldsymbol{I}_{\boldsymbol{y}}^{\star}(\omega) \xrightarrow{d} (2m+1)^{-1} W_c(2m+1, p, \boldsymbol{\lambda}_y(\omega))$. We have assumed that T is the same for both series, so it seems reasonable to assume that m is too. Recall that a Hermitian random $p \times p$ matrix $\boldsymbol{X} \sim W_c(2m+1, p, \boldsymbol{\Sigma})$ if its frequency function can be written as (Hannan 1970, p. 295)

$$\lambda(\boldsymbol{X}) = \frac{1}{\Gamma_p(2m+1)|\boldsymbol{\Sigma}|^{(2m+1)}} |\boldsymbol{X}|^{(2m+1)-p} \exp\left(-\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{X})\right), \tag{6}$$

with

$$\Gamma_p(s) = \pi^{p(p-1)/2} \prod_{j=1}^{p} \Gamma(s-j+1), \tag{7}$$

where $\Gamma(\,.\,)$ is the gamma function.

A likelihood ratio test is constructed for $H_0 : \boldsymbol{\lambda}_{\boldsymbol{x}}(\omega) = \boldsymbol{\lambda}_{\boldsymbol{y}}(\omega) = \boldsymbol{\lambda}(\omega)$, say, based on the independently distributed periodograms of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$, applying results derived in Conradsen et al. (2003) for testing equality of two complex Wishart matrices. The likelihood function at a fixed Fourier frequency ω is

$$\begin{aligned} L(\boldsymbol{\lambda}_{\boldsymbol{x}}(\omega), \boldsymbol{\lambda}_{\boldsymbol{y}}(\omega)) = {} & \frac{1}{\{\Gamma_p(2m+1)\}^2} \left|\boldsymbol{I}_{\boldsymbol{x}}^{\star}(\omega)\boldsymbol{I}_{\boldsymbol{y}}^{\star}(\omega)\right|^{(2m+1)-p} \left|\boldsymbol{\lambda}_{\boldsymbol{x}}(\omega)\boldsymbol{\lambda}_{\boldsymbol{y}}(\omega)\right|^{-(2m+1)} \\ & \times \exp\left[-\operatorname{tr}\left\{\boldsymbol{\lambda}_{\boldsymbol{x}}{}^{-1}(\omega)\boldsymbol{I}_{\boldsymbol{x}}^{\star}(\omega) + \boldsymbol{\lambda}_{\boldsymbol{y}}{}^{-1}(\omega)\boldsymbol{I}_{\boldsymbol{y}}^{\star}(\omega)\right\}\right]. \end{aligned} \tag{8}$$

Under (3), the likelihood function becomes

$$\begin{aligned} L_0(\boldsymbol{\lambda}(\omega)) = {} & \frac{1}{\{\Gamma_p(2m+1)\}^2} \left|\boldsymbol{I}_{\boldsymbol{x}}^{\star}(\omega)\boldsymbol{I}_{\boldsymbol{y}}^{\star}(\omega)\right|^{(2m+1)-p} |\boldsymbol{\lambda}(\omega)|^{-2(2m+1)} \\ & \times \exp\left(-\operatorname{tr}\left[\boldsymbol{\lambda}^{-1}(\omega)\left\{\boldsymbol{I}_{\boldsymbol{x}}^{\star}(\omega) + \boldsymbol{I}_{\boldsymbol{y}}^{\star}(\omega)\right\}\right]\right). \end{aligned} \tag{9}$$

Define

$$Q(\omega) = \frac{L_0(\widehat{\boldsymbol{\lambda}}(\omega))}{L(\widehat{\boldsymbol{\lambda}}_x(\omega))\, L(\widehat{\boldsymbol{\lambda}}_y(\omega))} \tag{10}$$

$$= 2^{2p(2m+1)} \frac{|\boldsymbol{I}_{\boldsymbol{x}}^{\star}(\omega)|^{2m+1} |\boldsymbol{I}_{\boldsymbol{y}}^{\star}(\omega)|^{2m+1}}{|\boldsymbol{I}_{\boldsymbol{x}}^{\star}(\omega) + \boldsymbol{I}_{\boldsymbol{y}}^{\star}(\omega)|^{2(2m+1)}}. \tag{11}$$

An effective overall test statistic for H_0 is the unweighted average of the M smallest $Q(\omega_j)$. We use this statistic, denoted by $Q^\star$, in the rest of this paper. It might be useful in some situations, such as when the time series are periodic with known period, or when the time series exhibit long memory, to use some other linear combination of the M smallest $Q(\omega_j)$, such as a weighted average with weights reflecting the effects of periodicity or long memory at different Fourier frequencies. We give a subjective guideline for selecting the value of M in the definition of $Q^\star$. The idea is to choose M sufficiently large such that local variations in the ratios $Q(\omega_j)$ over the Fourier frequencies should be smoothed out; but also to keep it small enough so that we can uncover sufficient detail in the overall spectra, and be sensitive to relatively narrow peaks in the spectra. In practice, a choice of M in the range $\sqrt{n}$ to $n/10$ seems adequate. The critical region will be $Q^\star < k_1$ where the value of k_1 may be determined from the null distribution of $Q^\star$, using simulation. For the simulation, we may take advantage of the facts that (1) the distribution of the test statistic does not depend on the true spectrum (and therefore does not depend on any parameters of the data-generating process, such as the AR or MA parameters of an ARMA process), so that we may generate white noise time series, and (2) the periodogram ordinates at distinct Fourier frequencies are independent. However, it is difficult to derive the explicit distribution of $Q^\star$ and thus to compute a critical value. More investigation of this as well as alternate forms of the test statistic would be useful.

2.2 Profile Likelihood Ratio Test

Situations arise in practice where components of one multivariate time series are scalar multiples of corresponding components of another multivariate time series, but nevertheless both would have similar second-moment properties (apart from scale). A test statistic based on a profile likelihood function is derived. Suppose that there are T observations $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ for $t = 1, \cdots, T$, generated from two independent p-dimensional stationary processes $\{\boldsymbol{X}_t\}$ and $\{\boldsymbol{Y}_t\}$, with spectral density matrices $\boldsymbol{\lambda}_{\boldsymbol{x}}(\omega)$ and $\boldsymbol{\lambda}_{\boldsymbol{y}}(\omega)$. We consider a test of the hypothesis

$$H_0[\boldsymbol{C}] : \boldsymbol{\lambda}_{\boldsymbol{x}}(\omega) = \boldsymbol{C}\boldsymbol{\lambda}_{\boldsymbol{y}}(\omega)\boldsymbol{C} \qquad \forall\, \omega \in (-\pi, \pi) \setminus \{0\} \tag{12}$$

where $\boldsymbol{C} = \mathrm{diag}(c_1, \cdots, c_p)$, for positive constants $c_1, \cdots, c_p$. Although the theory may apply for any arbitrary, nonsingular, constant matrix $\boldsymbol{C}$, the diagonal case gives the physical interpretation in our problem of the time series being similar except for scale, with interest centering on whether $\boldsymbol{X}_t$ and $\boldsymbol{C}\boldsymbol{Y}_t$ have identical spectra. For a given $\boldsymbol{C} = \mathrm{diag}(c_1, \cdots, c_p)$, the likelihood function under $H_0[\boldsymbol{C}]$ is

$$\begin{aligned} L(\boldsymbol{\lambda}_{\boldsymbol{y}}(\omega)|\boldsymbol{C}, \boldsymbol{I}^\star_{\boldsymbol{x}}(\omega), \boldsymbol{I}^\star_{\boldsymbol{y}}(\omega)) = {} & \frac{1}{\{\Gamma_p(2m+1)\}^2} |\boldsymbol{I}^\star_{\boldsymbol{x}}(\omega)\boldsymbol{I}^\star_{\boldsymbol{y}}(\omega)|^{2m+1-p} \\ & \times |\boldsymbol{C}\boldsymbol{\lambda}_{\boldsymbol{y}}(\omega)\boldsymbol{C}|^{-(2m+1)} |\boldsymbol{\lambda}_{\boldsymbol{y}}(\omega)|^{-(2m+1)} \\ & \times \exp[-\mathrm{tr}\{(\boldsymbol{C}\boldsymbol{\lambda}_{\boldsymbol{y}}(\omega)\boldsymbol{C})^{-1}\boldsymbol{I}^\star_{\boldsymbol{x}}(\omega) + \boldsymbol{\lambda}_{\boldsymbol{y}}(\omega)^{-1}\boldsymbol{I}^\star_{\boldsymbol{y}}(\omega)\}]\,. \end{aligned} \tag{13}$$

Using Giri (1965), the maximum likelihood estimate of $\boldsymbol{\lambda}_y(\omega)$, given $\boldsymbol{C}$, at a fixed ω is $\widehat{\boldsymbol{\lambda}}_y(\omega) = (\boldsymbol{I}^{\star}_y(\omega) + \boldsymbol{C}^{-1}\boldsymbol{I}^{\star}_x(\omega)\boldsymbol{C}^{-1})/\{2(2m+1)\}$. Substituting this estimate of $\widehat{\boldsymbol{\lambda}}_y(\omega)$ into Eq. refscalelrt, the profile likelihood function for $\boldsymbol{C}$ is:

$$L(\boldsymbol{C}|\hat{\boldsymbol{\lambda}}, \boldsymbol{I}^{\star}_x(\omega), \boldsymbol{I}^{\star}_y(\omega)) = \frac{1}{\{\Gamma_p(2m+1)\}^2}|\boldsymbol{I}^{\star}_x(\omega)\boldsymbol{I}^{\star}_y(\omega)|^{(2m+1)-p} \times \left|\frac{\boldsymbol{I}^{\star}_y(\omega) + \boldsymbol{C}^{-1}\boldsymbol{I}^{\star}_x(\omega)\boldsymbol{C}^{-1}}{2(2m+1)}\right|^{-2(2m+1)} |\boldsymbol{C}|^{-2(2m+1)} e^{-2(2m+1)p} \tag{14}$$

The estimation of $\boldsymbol{C}$ in closed form is very cumbersome, so we use numerical optimization of the profile likelihood function 14.

Under $H_0[\boldsymbol{C}]$, the likelihood function is given by Eq. 14 with $\boldsymbol{C}$ replaced by its estimate $\hat{\boldsymbol{C}}$. The profile likelihood ratio test statistic is a linear combination of the M smallest values of

$$Q(\omega)|_p = 2^{2(2m+1)p}\frac{|\boldsymbol{I}^{\star}_x(\omega)|^{2m+1}|\boldsymbol{I}^{\star}_y(\omega)|^{2m+1}|\hat{\boldsymbol{C}}|^{-2(2m+1)}}{|\boldsymbol{I}^{\star}_y(\omega) + \hat{\boldsymbol{C}}^{-1}\boldsymbol{I}^{\star}_x(\omega)\hat{\boldsymbol{C}}^{-1}|^{2(2m+1)}}\,. \tag{15}$$

The critical region for the above LRT statistic, say $Q^{\star}_P$, will be $Q^{\star}_P < k$.

In the special case where $\boldsymbol{C} = \mathrm{diag}(c, \cdots, c)$, the likelihood function becomes

$$L(\boldsymbol{C}, \hat{\boldsymbol{\lambda}}_y(\omega)|\boldsymbol{I}^{\star}_x(\omega), \boldsymbol{I}^{\star}_y(\omega)) = \frac{1}{\{\Gamma_p(2m+1)\}^2}|\boldsymbol{I}^{\star}_x(\omega)\boldsymbol{I}^{\star}_y(\omega)|^{2m+1-p} \times |\boldsymbol{\lambda}_x(\omega)\boldsymbol{\lambda}_y(\omega)|^{-(2m+1)} \exp\Big[-\mathrm{tr}\left\{c^{-2}\boldsymbol{\lambda}_y(\omega)^{-1}\boldsymbol{I}^{\star}_x(\omega) + \boldsymbol{\lambda}_y(\omega)^{-1}\boldsymbol{I}^{\star}_y(\omega)\right\}\Big] \tag{16}$$

and the maximum likelihood estimate of c follows from

$$\frac{d(\log L)}{dc} = -\frac{2p(2m+1)}{c} + 2c\,\mathrm{tr}\left(\boldsymbol{I}^{\star -1}_y\boldsymbol{I}^{\star}_x\right), \tag{17}$$

giving

$$\hat{c}^{-2} = \frac{\mathrm{tr}(\boldsymbol{I}^{\star -1}_y\boldsymbol{I}^{\star}_x)}{p(2m+1)}\,. \tag{18}$$

Although this is of some interest, it may be sufficient in practice to use a normalized spectral estimate, obtained by dividing each series by its standard error and then computing the spectrum.

3 Quasi-distances and Hierarchical Clustering

In order to cluster L stationary p-dimensional time series, a pairwise quasi-distance between two series $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ based on the likelihood ratio test statistic is defined as

$$d^{\omega}(\boldsymbol{x}_t, \boldsymbol{y}_t) = 1 - [Q^{\star}]^{1/(2m+1)}, \tag{19}$$

and $d^{\omega}(\boldsymbol{x}_t, \boldsymbol{x}_t) = 0$. For arguments $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$, this quasi-distance function satisfies $d(\boldsymbol{x}_t, \boldsymbol{y}_t) \geq 0$ and $d(\boldsymbol{x}_t, \boldsymbol{x}_t) = 0$, but does not satisfy the triangle inequality of the distance function. The $L \times L$ quasi-distance matrix is input into a hierarchical clustering algorithm (such as complete linkage, which is used here) and enables grouping of the L time series (Johnson and Wichern 2002). Clustering similar series for a set of multivariate time series in the scaled data situation discussed in Section 2.2 is based on a quasi-distance which is a function of the scale matrix when the estimated spectrum is not normalized:

$$d^{\omega}(\boldsymbol{x}_t, \boldsymbol{y}_t)|_P = 1 - [Q_P^{\star}]^{1/(2m+1)}, \tag{20}$$

These are illustrated in Section 4.

Kakizawa et al. (1998) described the use of the Kullback–Leibler and Chernoff information measures as disparity measures for comparing several multivariate time series, with application to earthquake-explosion discrimination. The Chernoff divergence measure (referred to here as KST quasi-distance) is defined as

$$J(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2T} \sum_{\omega} \left(\log \frac{|\alpha \widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}(\omega) + (1-\alpha)\widehat{\boldsymbol{\lambda}}_{\boldsymbol{y}}(\omega)|}{|\widehat{\boldsymbol{\lambda}}_{\boldsymbol{y}}(\omega)|} + \log \frac{|\alpha \widehat{\boldsymbol{\lambda}}_{\boldsymbol{y}}(\omega) + (1-\alpha)\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}(\omega)|}{|\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}(\omega)|} \right) \tag{21}$$

where $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}(\omega)$ and $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{y}}(\omega)$ represent the estimates of the spectral densities at Fourier frequency ω, and $0 < \alpha < 1$. When working with non-normalized spectra a modified KST quasi-distance is useful. We define it by

$$J(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{C}) = J(\boldsymbol{x}, \boldsymbol{C}\boldsymbol{y}). \tag{22}$$

It is straightforward to show that $J(.\,, .\,; \boldsymbol{C})$ is symmetric and that $J(\boldsymbol{x}, \boldsymbol{C}\boldsymbol{y}) = J(\boldsymbol{C}^{-1}\boldsymbol{x}, \boldsymbol{y})$, implying that the quasi-distance measure is invariant to whether $\boldsymbol{x}_t$ is rescaled to match the scale of $\boldsymbol{y}_t$ or vice versa. The scale matrix $\boldsymbol{C}$ in Eq. 22 can be estimated using the profile likelihood function given in Eq. 14.

4 Illustrations

4.1 Simulated Time Series

The approach is illustrated on simulated bivariate VAR time series under three different scenarios. Case I and Case II each consist of $L = 200$ zero-mean, bivariate VAR(1) stationary Gaussian time series each of length $n = 250$, with 100 series generated from each of two different populations defined by different sets of $\boldsymbol{\Phi}_1 = \{\phi_{1\,jk}\}$ parameters. The true parameters are shown in Table 1, and are selected such that the eigenvalues of $\boldsymbol{\Phi}_1$ under Population 1 and under Population 2 in Case II are closer to each other compared to the eigenvalues $\boldsymbol{\Phi}_1$ under the two populations in Case I. Case III consists of 100 zero-mean, bivariate VAR(1) stationary Gaussian time series, and 100 zero-mean, bivariate VAR(2) stationary Gaussian time series, each of length $n = 250$. The performance of the LRT based clustering method is tested by generating VAR(1) stationary Gaussian time series, and VAR(2) stationary Gaussian time series.

Calculating all possible pairwise quasi-distances between series, a complete linkage hierarchical clustering algorithm is used to cluster the series into two groups

Table 1 Parameters of simulated bivariate time series

	Case I: VAR(1)		Case II: VAR(1)		Case III	
	Pop 1	Pop 2	Pop 1	Pop 2	VAR(1)	VAR(2)
ϕ_{111}	1.90	0.60	1.90	0.70	1.90	0.50
ϕ_{112}	1.90	−0.70	1.90	−0.30	1.90	0.90
ϕ_{121}	−0.50	1.20	−0.50	0.30	−0.50	0.00
ϕ_{122}	−0.05	−0.50	−0.05	0.60	−0.05	0.30
ϕ_{211}						0.00
ϕ_{212}						0.00
ϕ_{221}						0.00
ϕ_{222}						0.40
σ_{11}	1.00	1.00	1.00	1.00	1.00	1.00
σ_{12}	0.50	0.50	0.50	0.50	0.50	0.50
σ_{22}	2.00	2.00	2.00	2.00	2.00	2.00

under each case. The comparative performance of the $Q^\star$ and KST (with $\alpha = 0.5$) based clustering using the smoothed periodogram was carried out. In Case I, separation of the simulated populations based on the $Q^\star$ and KST based clustering (with $\alpha = 0.5$) yielded misclassification rates of 0.01 and 0 respectively. Under Case II, the rates were 0.1 and 0.01 respectively, while under Case III, they were respectively 0.03 and 0.01. In general, the spectral based clustering based on $Q^\star$ is accurate.

The performance of Eq. 20 using non-normalized spectra is equivalent to the use of a distance function based on normalized spectra for separating dissimilar series when the different series are measured on different scales. To illustrate, time series from two different populations are generated. Population 1 consists of 50 series from a bivariate VAR(1) population, the parameters are those listed under Pop 1 under Case I in Table 1. Population 2 consists of 50 bivariate time series obtained by multiplying each series from Population 1 by a diagonal matrix, $\boldsymbol{C}$=diag(15,10). The quasi-distance function 19 based on the non-normalized smoothed periodogram treats the scaled time series in Population 2 as distinct from the series in Population 1, so that the 50 series from Population 1 are clustered into one group, while the 50 series from Population 2 are clustered into the second group. However, the quasi-distance function 20 , based on the non-normalized smoothed periodogram accurately discriminates between the populations, so that all 100 series are clustered into a single group. The KST based quasi-distance function 21 exhibits a similar behavior. The quasi-distance 22 is again able to discriminate the populations comparably with the clustering based on the normalized smoothed periodogram.

4.2 Illustration on Daily Temperatures and Precipitation

We illustrate our approach using climate data for the state of Washington, in the Pacific Northwest of the USA. We cluster locations based on stochastic properties of time series of maximum daily temperatures (TMAX), minimum daily temperatures (TMIN), and precipitation from January 1, 2005 through December 31, 2007. Data were obtained from the site www.ncdc.noaa.gov and were cleaned to exclude locations with missing values. Clean data were obtained for 43 locations. The time series of daily precipitation amounts was converted for simplicity into a binary time series (PRCP) assuming the value one if the precipitation was positive and zero if

the precipitation was zero. All variables were scaled to have range 1, giving them approximately equal importance in the clustering procedures described below.

The National Climatic Data Center has defined 10 climate divisions for the state of Washington: 1, West Olympic Coast; 2, North East Olympic San Juan; 3, Puget Sound Lowlands; 4, East Olympic Cascade Foothills; 5, Cascade Mountains West; 6, East Slope Cascades; 7, Okanogan Big Bend; 8, Central Basin; 9, Northeastern; and 10, Plouse Blue Mountains. The climate divisions to which each of the sites belong are indicated in Fig. 1. Note, however, that "the divisional boundaries may not delineate areas of climatological homogeneity" (Guttman and Quayle 1996).

We chose the number of clusters to be 10, the number of NCDC climate divisions for the state of Washington. In general, as in any cluster analysis, the number of clusters could be chosen according to some objective criterion, such as the "average silhouette criterion" of Kaufman and Rousseeuw (1990).

Let us suppose that the time series may differ either in their first order properties (means) or in their second order properties (spectra), or both. The following framework helps in clustering $N = 43$ such series.

First order criterion The quasi-distances are based on pairwise tests of whether the time series $\{\mathbf{X}_t\}$ and $\{\mathbf{Y}_t\}$ have same means, i.e., test $H_{01} : \boldsymbol{\mu}_x = \boldsymbol{\mu}_y$. Let $\overline{\mathbf{X}}$ and $\overline{\mathbf{Y}}$ denote the sample means of the two time series. The Hotelling T_H^2 statistic for testing H_{01} has the form

$$T_H^2 = \left(\overline{\mathbf{X}} - \overline{\mathbf{Y}}\right)' \mathbf{S}^{-1} \left(\overline{\mathbf{X}} - \overline{\mathbf{Y}}\right) \tag{23}$$

where

$$\begin{aligned} \mathbf{S} = \operatorname{cov}\left(\overline{\mathbf{X}} - \overline{\mathbf{Y}}\right) &= \operatorname{cov}\left(\overline{\mathbf{X}}\right) + \operatorname{cov}(\overline{\mathbf{Y}}) \\ &= \frac{2\pi}{T}[\boldsymbol{\lambda}_x(0) + \boldsymbol{\lambda}_y(0)] \end{aligned} \tag{24}$$

and serves as a quasi-distance in a hierarchical clustering step. Although it is well known that $T_H^2 \sim [\{2(T-1)p\}/(2T-p-1)]F_{p,2T-p-1}$ under H_{01}, rather than employing formal decision rules based on a critical value from this F distribution, we treat the value of T_H^2 as a quasi-distance which we can use as a measure of dissimilarity between time series. However, the quadratic form 25, although it adjusts for correlation between variables at individual sites, can still cause variables to have

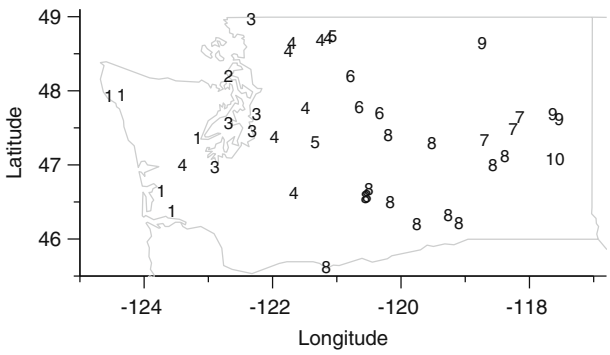

Fig. 1 Climate divisions for Washington State

greatly different degrees of influence on the final clustering. To assign approximately equal importance to the variables in the Washington state climate data, we rescale the means in Eq. 25 so that they have the same range across all the sites. Letting $\mathbf{R}$ be a 3×3 diagonal matrix whose diagonal elements are the ranges, across all the sites in the Washington climate data set, of the three variables TMAX, TMIN, and PRCP, we define a quasi-distance

$$(T_H^*)^2 = \left(\overline{\mathbf{X}} - \overline{\mathbf{Y}}\right)' \mathbf{R}^{-1} \mathbf{S}^{-1} \mathbf{R}^{-1} \left(\overline{\mathbf{X}} - \overline{\mathbf{Y}}\right) \tag{25}$$

and use it as a measure of the dissimilarity between sites. We store the dissimilarities between each pair of sites in an $N \times N$ matrix $\mathbf{D}_1$.

Second order criterion Here, the quasi-distances are based on pairwise tests of whether the time series $\{\mathbf{X}_t\}$ and $\{\mathbf{Y}_t\}$ have the same spectra, i.e., test H_{02} : $\boldsymbol{\lambda}_x(\omega) = \boldsymbol{\lambda}_y(\omega)\ \forall\ \omega \in (-\pi, \pi] \setminus \{0\}$. The values of $-Q^*$ defined in Section 2.1 for each pair of series are input into another $N \times N$ matrix, $\mathbf{D}_2$.

Combined criterion We use ideas from Woznica et al. (2007) to combine the first and second order properties. For a given 2×2 weight matrix $\mathbf{A} = \{a_{h,k}\}$, we define a combined distance matrix $\mathbf{D}_c$ whose (i, j)th element $D_{c,i,j}$ is given by

$$D_{c,i,j}^2 = \sum_{h=1}^{2} \sum_{k=1}^{2} a_{h,k} D_{h,i,j} D_{k,i,j}\,. \tag{26}$$

In order to give equal importance to both first and second order criteria in clustering the time series, we choose $\mathbf{A} = \mathrm{diag}(a_{1,1}, a_{2,2})$ with $a_{k,k}$ defined as the reciprocal of the range over all pairs of quasi-distances in $\mathbf{D}_k$. Woznica *et al.* discuss other optimal choices for $\mathbf{A}$, although it is not immediately obvious that these would be useful in clustering problems. The final clustering based on $\mathbf{D}_c$ then reflects the groupings in the N time series based on their first and second moment properties, both equally weighted.

Results from clustering the locations in the state of Washington are given in Fig. 2. The figure shows the results of the first-order, second-order and combined clusterings described above, and, for comparison with the second-order clustering, the clustering based on the KST quasi-distance 21 with $\alpha = 1/2$.

All the clustering methods agree with the climate division map in clearly separating five regions in the humid western part of the state from another five in the relatively arid eastern part. In the western part, corresponding to climate divisions 1–5, all the clustering methods find a cluster of sites around Puget Sound. All except the first-order method find a cluster of sites in the northern part of climate division 4 (clusters 6, 9, and 1 for the second-order, combined, and KST methods, respectively) that is distinct from the other sites in that division. Also, all the clustering methods assign the site in the Olympic Peninsula that is in climate division 4 (Elma, at approx. lat. 47, long. −123.4) to a different region from the other sites in climate division 4. There is therefore a clear indication that climate division 4 may not be climatologically homogeneous.

In the eastern part of the state neither the first-order nor the second-order method yields clearly delineated clusters, but the "combined" clusters have a fairly high degree of geographic coherence, apart from a single isolated site in each of clusters 3 and 7.

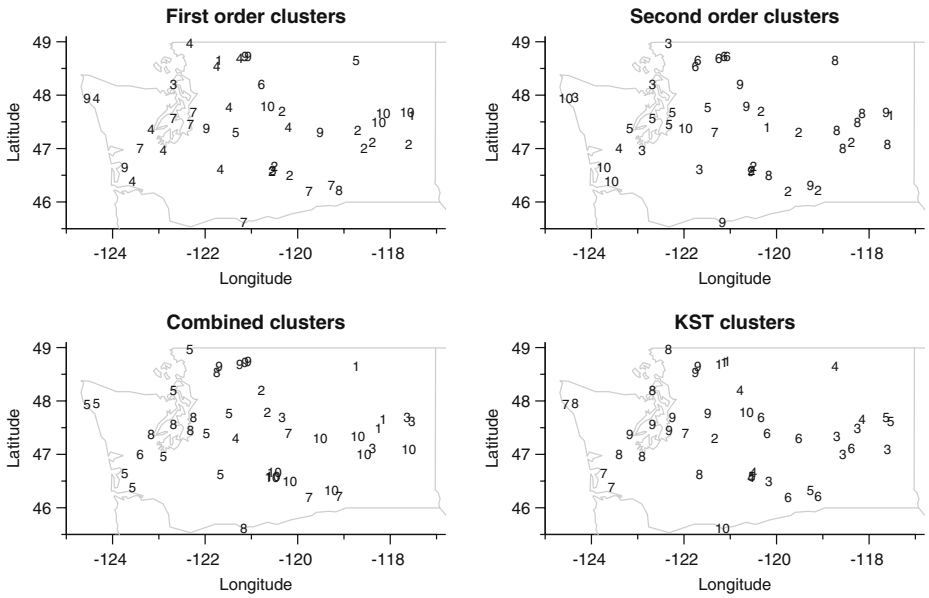

Fig. 2 Clustering for Washington State, based on first order, second order (with $M = 100$) and combined criteria, and on the KST criterion with $\alpha = 0.5$

We consider the "combined" clusters in a little more detail. The number of locations in the 10 clusters are 3, 2, 4, 1, 10, 1, 3, 6, 4, 9. Table 2 shows the minimum, mean, and maximum of PRCP, TMAX, and TMIN across each cluster. Figure 3 shows the average, minimum and maximum direct spectra corresponding to PRCP for each of the 10 clusters, while Fig. 4 shows the coherency between the spectra of PRCP and TMAX. The table shows some clear differences between the clusters. For example, clusters 4, 5, 6, and 9 have high average values of PRCP, while clusters 1, 3, 7, and 10 have low values of PRCP; clusters 8 and 2 have intermediate values. The plots also show differences in spectral behavior in the different clusters. For example,

Table 2 Minima, averages and maxima for mean characteristics (PRCP, TMAX, TMIN) across clusters, using the "combined" criterion

Cluster	No. of sites	Minimum	Average	Maximum
1	3	(0.21, 56.6, 32.2)	(0.24, 57.8, 33.4)	(0.26, 59.3, 34.1)
2	2	(0.33, 52.4, 30.2)	(0.36, 54.9, 32.2)	(0.39, 57.4, 34.2)
3	4	(0.24, 57.6, 35.7)	(0.27, 59.5, 37.3)	(0.31, 61.3, 38.7)
4	1	(0.52, 49.4, 36.4)	(0.52, 49.4, 36.4)	(0.52, 49.4, 36.4)
5	0	(0.38, 56.9, 40.0)	(0.50, 59.5, 41.3)	(0.58, 61.5, 43.9)
6	1	(0.56, 63.1, 41.3)	(0.56, 63.1, 41.3)	(0.56, 63.1, 41.3)
7	3	(0.15, 61.4, 42.0)	(0.17, 64.2, 43.1)	(0.20, 66.2, 44.6)
8	6	(0.25, 58.9, 42.3)	(0.42, 60.6, 43.7)	(0.51, 64.5, 45.3)
9	4	(0.48, 55.6, 39.0)	(0.50, 57.2, 40.8)	(0.52, 58.2, 42.1)
10	9	(0.16, 61.7, 35.3)	(0.22, 63.1, 38.2)	(0.35, 64.7, 42.5)

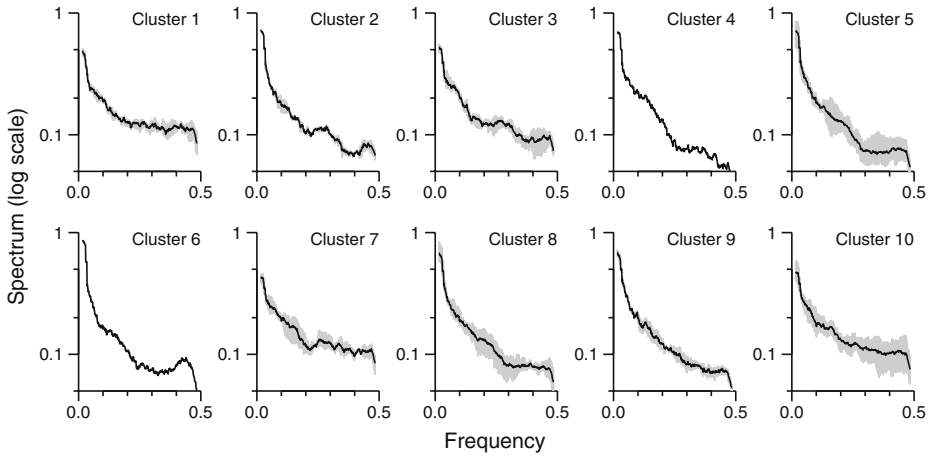

Fig. 3 Average (*solid line*) and range (*shaded area*) of direct spectra for PRCP across clusters

the spectra for PRCP for humid-region clusters 4, 5, 6, 8, and 9 have relatively high amplitudes at low frequencies and relatively low amplitudes, less than 0.1, at frequencies greater than 0.25. The spectra for arid-region clusters, 1, 3, 7, and 10, have lower amplitudes at low frequencies and less difference between amplitudes at high and low frequencies. Cluster 2 is intermediate between the two larger groupings of clusters but its spectrum more closely resembles those of the humid-area group. The coherency graphs Clusters 4 and 9, both in the Cascades mountains in the center of the state, have high coherency, compared to the other clusters, in the frequency range 0.1–0.2.

The clustering methods, particularly the "combined" method, bring out some clear spatial patterns in the climate data, but are by no means definitive. There is scope

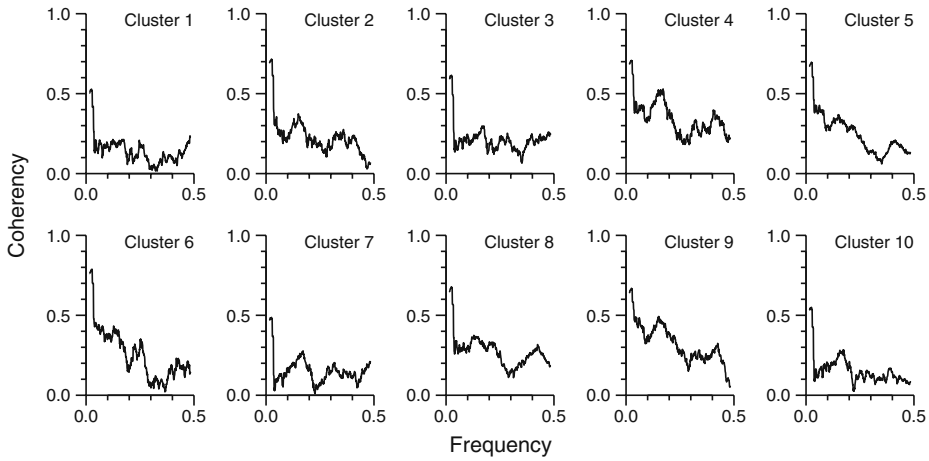

Fig. 4 Average coherency for (PRCP, TMAX) across clusters

for further development of this approach—for instance, to see whether there are clusters that share similar values of other weather-related characteristics such as altitude, wind speed, etc. More extensive analysis based on a more comprehensive set of climate data could answer such questions.

5 Concluding Remarks

We have shown that likelihood ratio test on spectral matrices can be used as the basis of an effective clustering method for continuous-valued stationary multivariate time series. Spectral clustering creates dissimilarity measures based on (second-order) spectra for mean-subtracted data. It is of course possible that time series clustered into a group might have different mean levels. If we wish to use the clustering outcomes to simplify the error covariance and/or coefficient matrices in the subsequent modeling, this should not be an issue. If, on the other hand, the scientist wishes to include the differential mean behavior in the clustering, then it is possible that a nested sequence of hypotheses may be envisaged, with one level of clustering based on the first moment, and the other level carrying out spectral clustering. We do not explore this here.

Other clustering methods such as quasi-ML-ratio tests in a VAR(p) setting (perhaps as $p \to \infty$) may be used for grouping, by testing whether the fit in a combined estimation is worse than in separated models. The advantage of such a parametric framework to assess similarity between time series is that if the assumed model is the true model, then the parametric test on coefficients should be in general more powerful. However, in practice, the true model is rarely known, and in such cases, the testing approach might lead to incorrect grouping, and a nonparametric approach such as the spectral approach may be more useful, especially for a first level analysis. Future work could address a careful numerical comparison of the different approaches to clustering vector time series.

Acknowledgements The authors are grateful to anonymous reviewers whose suggestions enhanced the paper.

References

Brillinger DR (1986) Time series: data analysis and theory. SIAM, Philadelphia
Brockwell PJ, Davis RA (1991) Time series: theory and methods. Springer, New York
Coates DS, Diggle PJ (1986) Tests for comparing two estimated spectral densities. J Time Ser Anal 7:7–20
Conradsen K, Nielsen AA, Schou J, Skriver H (2003) A test statistic in the complex Wishart distribution and its application to change detection in polarimetric SAR data. IEEE Trans Geosci Remote Sens 41:4–19
Giri N (1965) On the complex analogues of T^2 and R^2-tests. Ann Math Stat 36:664–670
Goodman NR (1963) The distribution of the determinant of a complex Wishart distributed matrix. Ann Math Stat 34:178–180
Guttman NB, Quayle RG (1996) A historical perspective of U.S. climate divisions. Bull Am Meteorol Soc 77:293–303
Hannan EJ (1970) Multiple time series. Wiley, New York
Johnson AJ, Wichern DW (2002) Applied multivariate statistical analysis. Prentice Hall, New Jersey
Kakizawa Y, Shumway RH, Taniguchi M (1998) Discrimination and clustering for multivariate time series. J Am Stat Assoc 93:328–340

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York

Pai JS, Ravishanker N, Gelfand AE (1994) Bayesian analysis of concurrent time series with application to regional IBM revenue data. J Forecast 13:463–479

Woznica A, Kalousis A, Hilario M (2007) Learning to combine distances for complex representations. In ICML '07: Proceedings of the 24th international conference on machine learning. ACM, Corvallis, pp 1031–1038