Behav Res (2017) 49:988–1005
DOI 10.3758/s13428-016-0754-9

Detecting correlation changes in multivariate time series: A comparison of four non-parametric change point detection methods

Jedelyn Cabrieto[1] · Francis Tuerlinckx[1] · Peter Kuppens[1] · Mariel Grassmann[2,3] · Eva Ceulemans[1]

Published online: 6 July 2016

Abstract Change point detection in multivariate time series is a complex task since next to the mean, the correlation structure of the monitored variables may also alter when change occurs. DeCon was recently developed to detect such changes in mean and\or correlation by combining a moving windows approach and robust PCA. However, in the literature, several other methods have been proposed that employ other non-parametric tools: E-divisive, Multirank, and KCP. Since these methods use different statistical approaches, two issues need to be tackled. First, applied researchers may find it hard to appraise the differences between the methods. Second, a direct comparison of the relative performance of all these methods for capturing change points signaling correlation changes is still lacking. Therefore, we present the basic principles behind DeCon, E-divisive, Multirank, and KCP and the corresponding algorithms, to make them more accessible to readers. We further compared their performance through extensive simulations using the settings of Bulteel et al. (Biological Psychology, 98 (1), 29-42, 2014) implying changes in mean and in correlation structure and those of Matteson and James (Journal of the American Statistical Association, 109 (505), 334-345, 2014) implying different numbers of (noise) variables. KCP emerged as the best method in almost all settings. However, in case of more than two noise variables, only DeCon performed adequately in detecting correlation changes.

Keywords Change point detection · Correlation changes · Multivariate time series · DeCon · ROBPCA

Change point detection is an old and important problem in time series analysis (Basseville & Nikiforov, 1993; Bhattacharya & Johnson, 1968; Kander & Zacks, 1966; Page, 1954). As indicated by its name, the goal of change point detection is to detect whether and when abrupt distributional changes take place in a time series, which is crucial in a diverse set of fields such as climate science, economy, medicine, etc. (see Chen & Gupta, 2012). Current applications in the field of behavioral sciences include detection of workload changes using heart rate variability (Hoover, Singh, Fishel-Brown, & Muth, 2011), capturing active state transition in fMRI activity (Lindquist, Waugh, & Wager, 2007) and revealing cardio-respiratory changes preceding the occurrence of panic attacks (Rosenfield, Zhou, Wilhelm, Conrad, Roth, & Meuret, 2010). Until recently, research on this topic focused almost exclusively on univariate time series, yielding approaches to detect changes in mean, and, in some cases, variance and/or autocorrelation.

With the advance of technology, more and more studies generate multivariate time series. For example, climate studies

✉ Jedelyn Cabrieto
Jed.Cabrieto@ppw.kuleuven.be

Francis Tuerlinckx
francis.tuerlinckx@kuleuven.be

Peter Kuppens
peter.kuppens@kuleuven.be

Mariel Grassmann
Mariel.Grassmann@kienbaum.de

Eva Ceulemans
eva.ceulemans@kuleuven.be

[1] Quantitative Psychology and Individual Differences Research Group, KU Leuven – University of Leuven, Tiensestraat 102, Leuven B-3000, Belgium

[2] Health Psychology Research Group, KU Leuven – University of Leuven, Tiensestraat 102, Leuven B-3000, Belgium

[3] Department of Aviation and Space Psychology, German Aerospace Center (DLR), Sportallee 54a, Hamburg 22335, Germany

monitor several environmental factors such as temperature, precipitation and water discharges (Jarusikova, 1997). In neurophysiology (Terien, Germain, Marque, & Karlsson, 2013), analysis of biological functions entails following numerous physiological signals. Turning to examples from the behavioral sciences, in emotion psychology, experiential, behavioral, and physiological reactions to emotional stimuli are tracked across time (Christie & Friedman, 2004; Mauss, Levenson, McCarter, Wilhelm, & Gross, 2005), and in developmental psychology, performance on several Piagetian tasks (Piaget, 1972) is examined over time to assess how cognitive abilities of children develop (Amsel & Renninger, 1997; Klausmeier & Sipple, 1982; van der Maas & Molenaar, 1992).

Given multivariate data, change point detection involves more than changes in single variables because the system characteristics seldom react in an isolated way to change. Indeed, in many cases, theory prescribes that, next to the mean, also the correlation structure of (a subset of) the system characteristics alters when change occurs. In emotion psychology, researchers postulate that physiological, experiential, and behavioral reactions synchronize in emotion-inducing situations to enable the organism to quickly and efficiently cope with environmental threats or opportunities (Mauss et al., 2005). In developmental psychology, one conjectures that before a sudden developmental jump – the mastery of a specific ability - the correlation structure of a set of tasks changes (Amsel & Renninger, 1997; van der Maas & Molenaar, 1992). Outside psychology, one can think of the strengthened correlation between economic growth rates of countries implementing a common monetary policy (Crowley & Schultz, 2011), parts of the brain exhibiting excessive neuronal synchronization during a seizure (Terien et al., 2013), or climactic indices demonstrating higher correlations during cool seasons (Wright & Wallace, 1988). As a consequence, detecting correlation changes becomes an integral aspect of the change point analysis problem in the case of multivariate data (see Aue, H rmann, Horváth, & Reimherr, 2009; Müller, Baier, Galka, Stephani, & Muhle, 2005; Terien, Marque, Germain, & Karlsson, 2009).

Recently, a number of non-parametric multivariate change point detection methods have been proposed that can be used to detect changes in both correlation structure and means: DeCon (Bulteel, Ceulemans, Thompson, Waugh, Gotlib, Tuerlinckx, & Kuppens, 2014), E-divisive (Matteson & James, 2014), Multirank (Lung-Yut-Fong, Lévy-Leduc, & Cappé, 2012), and KCP (Arlot, Celisse, & Harchaoui, 2012). However, we see two problems when an applied researcher wants to apply these methods. First, the methods are based on different statistical approaches: Robust methods for DeCon, rank information for Multirank, the kernel trick for KCP and Euclidean distances for E-divisive. This diversity makes it difficult for the applied researcher to appraise the methods. Second, because they are based on different statistical approaches, it is still unknown which of these four methods should be preferred in which circumstances. More specifically, a direct comparison between these methods for the detection of correlational changes is lacking (although Matteson & James, 2014, conducted a partial comparison).

Given these two problems, this paper fulfills two goals. The first goal is to introduce the basic principles behind each method and the corresponding algorithms in easy to follow steps to make them more accessible to readers. The second goal is to study the relative performance by means of extensive simulations. Note that we focus on non-parametric methods, given their wide applicability.

In the remainder of this paper, we first introduce each of the four methods, using an illustrative hypothetical example. Next, we apply the methods to two sets of simulated data based on Bulteel et al. (2014) and on Matteson and James (2014) and to an empirical data set on pilot reactions. In the final section, the results are discussed and future research directions are enumerated.

Method

Before discussing the four methods in detail, we introduce an illustrative hypothetical data set that will be used throughout this section. Let $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{50}\}$ denote the whole time series, composed of 50 time points at which three variables are measured. This time series is shown in Fig. 1. A change point occurs between the 25th and 26th time point, segmenting the time series into two phases of 25 subsequent time points. The 25 observations in Phase 1 are randomly sampled from a multivariate normal distribution with mean, $\mu_{1:25} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$, and covariance matrix, $\Sigma_{1:25} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, implying that all variables are independent. In Phase 2, the observations are also drawn from a multivariate normal distribution but the means $\mu_{26:50} = \begin{bmatrix} 3 \\ 6 \\ 9 \end{bmatrix}$ are higher and the variables become strongly correlated, as indicated by the covariance matrix $\Sigma_{1:25} = \begin{bmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{bmatrix}$. It should be emphasized that the actual change occurs between the 25th and 26th time points. However, this true change point is unobserved. Hence, in the remainder of this paper, we will use the first observation after the change in distribution as the change point. Thus, for the hypothetical data, the change point is $T = 26$.

DeCon

DeCon bases change point detection on outlier identification using robust statistics (Bulteel et al., 2014). The method slides a time window of size W across the time series by sequentially deleting the first time point in the window, and adding one new observation as the last time point. Per window, it is determined whether the last time point is an outlier with respect to the distribution of the other time points in the window. If the latter is the

Phase 1 Observations

X_1	X_2	X_3	…	X_{25}
2.35	2.04	2.52	…	2.27
1.98	2.82	2.07	…	2.10
3.77	3.62	4.29	…	3.36

Phase 2 Observations

X_{26}	X_{27}	X_{28}	…	X_{50}
1.08	4.50	2.01	…	3.77
4.37	7.61	5.26	…	6.72
7.36	10.44	7.88	…	9.67

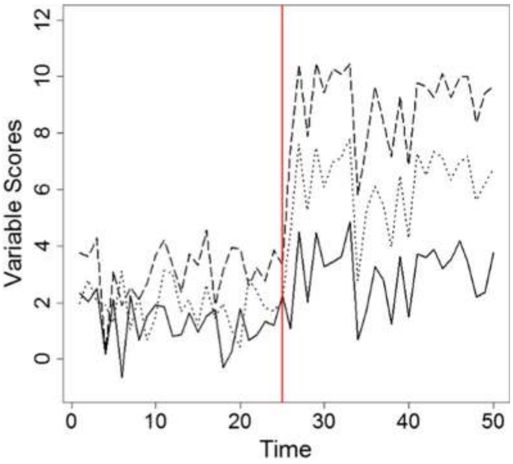

Fig. 1 Illustrative hypothetical data set with three variables, 50 time points, and one change point between the 25th and 26th time point, segmenting the series into two phases. Phase 1 observations, $\boldsymbol{X}_{1:25}$, were drawn from $MVN\left(\begin{bmatrix}1\\2\\3\end{bmatrix},\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}\right)$ and Phase 2 observations, $\boldsymbol{X}_{26:50}$, were drawn from $MVN\left(\begin{bmatrix}3\\6\\9\end{bmatrix},\begin{bmatrix}1&0.9&0.9\\0.9&1&0.9\\0.9&0.9&1\end{bmatrix}\right)$

case for multiple consecutive windows, this signals that the observations that are added to the window might come from a different distribution, and, hence, that a change point occurred. Specifically, DeCon consists of the following four steps.

1. *Apply Robust PCA in each time window and determine "outlyingness" of the last time point.*

Per time window, DeCon computes a robust multivariate center, $\boldsymbol{\mu}_w$, and a covariance matrix, Σ_w, to determine the distribution of the regular observations (standardized per variable since we are interested in correlations rather than covariances), and generates an outlyingness measure for the last time point of the window. To this end, the robust principal components approach (ROBPCA) of Hubert et al. is used (for details, see Hubert, Rousseeuw & Vanden Branden, 2005). In this paper, we retained all principal components to avoid the issue of how to choose the optimal number of components.[1] Given that we used all components, the outlyingness measure is the so-called score distance, which equals the Mahalanobis distance between the last time point $\boldsymbol{X}_{last}$ and the robust window-specific center $\boldsymbol{\mu}_w$:

$$SD_{last} = \sqrt{(\boldsymbol{X}_{last}-\boldsymbol{\mu}_w)^T \sum\nolimits_w^{-1} (\boldsymbol{X}_{last}-\boldsymbol{\mu}_w)} \qquad (1)$$

Note that the Mahalanobis distance differs from the Euclidean one, in that the covariance matrix of the variables under consideration is taken into account. If the data are normally distributed,[2] the squared Mahalanobis distances follow a χ^2 distribution with degrees of freedom equal to the number of variables. Thus, the last time point is classified as an outlier if the Mahalanobis distance exceeds the square root of the 97.5th quantile of this χ^2 distribution.

For analyzing our hypothetical data, the window size was set to $W = 20$. In general, this parameter should be chosen considering the minimum time period within which no change is expected to occur (for more considerations and detailed simulation results, see Bulteel et al., 2014). ROBPCA was applied to the first window, $\boldsymbol{X}_{1:20}$, then to the second window, $\boldsymbol{X}_{2:21}$, and so on, until the last window, $\boldsymbol{X}_{31:50}$. Since the change point occurs at $T = 26$, chances are high that the last time point of the first time window that includes a new phase observation as its last time point, $\boldsymbol{X}_{7:26,}$ has a large score distance. In general, this probability depends on how different the means and correlations are in the subsequent phases. For the time window, $\boldsymbol{X}_{7:26,}$ where the robust center equals $\mu_{7:26} = \begin{bmatrix}1.41\\1.89\\3.18\end{bmatrix}$ and the robust covariance matrix is given by $\Sigma_{7:26} = \begin{bmatrix}0.29&-0.15&0.11\\-0.15&0.62&0.13\\0.11&0.13&0.57\end{bmatrix}$, the last observation, $x_{26} = \begin{bmatrix}1.08\\4.37\\7.36\end{bmatrix}$, has a score distance equal to

$$SD_{26} = \sqrt{\left(\begin{bmatrix}1.08\\4.37\\7.36\end{bmatrix}-\begin{bmatrix}1.41\\1.89\\3.18\end{bmatrix}\right)^T \begin{bmatrix}0.29&-0.15&0.11\\-0.15&0.62&0.13\\0.11&0.13&0.57\end{bmatrix}^{-1}\left(\begin{bmatrix}1.08\\4.37\\7.36\end{bmatrix}-\begin{bmatrix}1.41\\1.89\\3.18\end{bmatrix}\right)}$$
$$= 6.06.$$

For this example, the cut-off for the score distance is 3.06. Figure 2 (left panel) shows that the score distance of the last observation, X_{26}, indeed clearly exceeds the cut-off indicated by the red line.

[1] Note that Bulteel et al. (2014) implemented an automatic procedure to determine the number of components and classified a time point as outlying if either the orthogonal distance or the score distance exceeds the respective cut-off. However, for the simulations settings reported in this paper, retaining all components worked equally well.

[2] If the normality assumption may not hold for a specific data set, a variant of ROBPCA for skewed data (Hubert, Rousseeuw & Verdonck 2009) can be plugged in into DeCon.

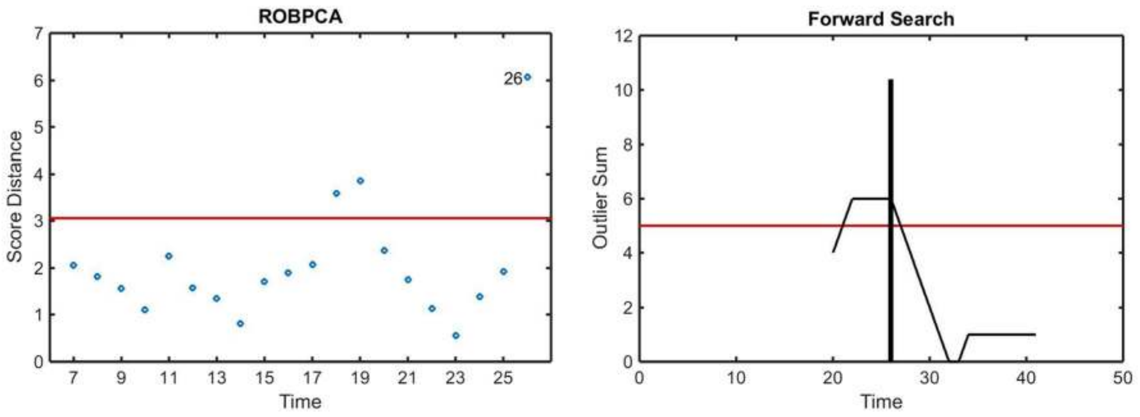

Fig. 2 ROBPCA outlier plot for window, $X_{7:26}$, and plot of the moving outlier sum generated by the forward procedure for the hypothetical data. The left panel shows that X_{26} exceeds the score distance cut-off (indicated by the horizontal line), hence this observation is flagged as an outlier. The right panel reveals that the moving outlier sum reaches the moving sum cut-off at $T = 21$ when the sum is computed across observations, X_{21}, X_{22}, X_{23}, …, X_{30}. The vertical line indicates that the location of the change point is set at $T = 26$, since X_{26} is the first outlier within these ten observations

2. *Track the moving sum of the outlyingness of ten subsequent last time points and declare a change point when this sum equals five at least.*

The outlyingness of the last time point is a binary variable and thus a binary time series is created (with 1 indicating an outlier and 0 a regular observation). Although the outlyingness of the last time point of a window may correctly signal the presence of a change point, false negatives or false positives can of course occur. To mitigate their impact, the results of multiple windows are combined. Specifically, a moving sum of the outlyingness of the final time point of ten subsequent windows is computed. As soon as this sum equals at least five, the first outlying time point in the corresponding set of time points is declared as the change point. Note that the 5 out of 10 rule is recommended to balance Type 1 and Type 2 errors (see Bulteel et al., 2014). Going back to our example, Fig. 2 (right panel) shows the moving outlier sum for the whole sequence. The moving sum cut-off was reached for the first time when the moving sum included observations, X_{21}, $X_{22}, X_{23}, \ldots, X_{30}$. Out of these observations, the first outlying one was X_{26}. Thus, the change point is detected at time point $T = 26$.

When a change point occurs, it is quite likely that the moving outlier sum stays relatively high for a while, because the change in correlation structure and means will only start to influence the ROBPCA estimates if at least .25W time points within a window pertain to the next phase. This is because by default, the ROBPCA estimates are based on the 75 % least outlying cases only. Therefore, the minimum distance between subsequent change points equals .25W. In our hypothetical time series no further change points were detected.

3. *Repeat steps 1 and 2 in the backward direction.*

The original DeCon procedure executed steps 1 and 2 only. The simulations in the present paper revealed that this “forward procedure” (forward, because we slide the time window from the first to the last time point), works well for settings where the correlation decreases. Indeed, in these cases, the first time points in the moving window come from a more compact joint distribution (higher correlation), while later observations come from a more scattered joint distribution (low correlation). These later observations, thus, will generate larger score distances, and will be correctly flagged as outliers. However, when there is an increase in correlation across time, the time window moves from a phase with a lower correlation (Phase 1) to a phase with a higher one (Phase 2). If the means and variances remain the same, observations in Phase 2 will most likely have a small score distance, when compared to the distribution in Phase 1, because the distribution of Phase 2 will mostly overlap with that of Phase 1

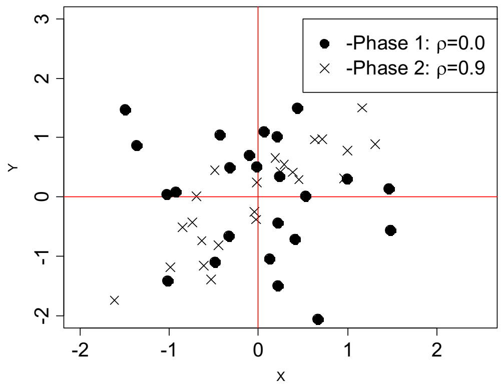

Fig. 3 Overlapping phase distributions. Observations are bivariate normal with all means equal to zero and all variances equal to 1. Phase 1 observations are uncorrelated (ρ = 0.0) and Phase 2 observations are highly correlated (ρ = 0.9)

(see Fig. 3). Since no outliers are detected, neither is the change point.

This limitation of the forward DeCon procedure can be resolved by also performing a "backward search," which boils down to reversing the time order (last time point of the sequence becomes the first one, etc.) and conducting steps 1 and 2 on this reversed sequence. Indeed, in the backward search, the increase in correlation becomes a drop in correlation, since the time points in Phase 2 constitute the standard against which observations in Phase 1 will be compared. Therefore, new observations are identified as outliers, and the change point can be detected. For the illustrative example, the change point was detected at $T = 25$. However, one should be aware that we are now working in the backward sense, implying that the detected change point actually is the last observation of a phase, rather than the first one of a new phase. Thus, to transform the backward estimate of the change point locations to the correct time order, we should add one to it. Thus, we obtain $T = 26$, which is indeed the correct change point.

4. *Combine change points detected in the forward and the backward procedure.*

Finally, the change points detected in the forward and backward procedure are pooled together. Of course, it will often happen that the forward and backward search will detect the same phase change, but yield slightly different estimates of the change point. In the simulations, change point estimates that are within a 10-time point distance will be pooled by computing their means. However, this maximum between distance for pooled change points may be adjusted by the user to a higher number when one deals with a much longer time series. Moreover, when the mean does not correspond to an exact time point in the series, we round the estimate to the next time point.

E-divisive

E-divisive detects change points by quantifying how different the characteristic functions of the distributions of subsequent segments of the time series are (Matteson & James, 2014). Indeed, given that characteristic functions uniquely describe a probability distribution (Gnedenko, 2005), changes in the characteristic function signal distributional change (Matteson & James, 2014). E-divisive performs the following segmentation steps.

1. *Segment the time series into two phases for which the characteristic functions maximally differ.*

To segment the time series into two phases for which the characteristic functions maximally differ, based on derivations from Szekely and Rizzo (2005), the following divergence measure of phases, $\boldsymbol{X}_{1:\tau}$ and $\boldsymbol{X}_{\tau+1:n}$, is computed for different τ-values:

$$\hat{Q}(\tau) = \frac{\tau(n-\tau)}{n}\left[\frac{2}{\tau(n-\tau)}\sum_{i=1}^{\tau}\sum_{j=\tau+1}^{n}\|X_i-X_j\| - \binom{\tau}{2}^{-1}\sum_{i=1}^{\tau-1}\sum_{k=i+1}^{\tau}\|X_i-X_k\| - \binom{n-\tau}{2}^{-1}\sum_{j=\tau+1}^{n-1}\sum_{k=j+1}^{n}\|X_j-X_k\|\right], \tag{2}$$

where ‖ · ‖ denotes the Euclidean distance.[3] The left-most term within the square brackets expresses the average Euclidean distance of time points belonging to different phases, whereas the two right-most terms quantify the average within-phase distances, separately for Phase 1 and Phase 2. For instance, if we divide our illustrative time series into two candidate phases $\boldsymbol{X}_{1:25}$ and $\boldsymbol{X}_{26:50}$ by setting τ equal to 25, the corresponding divergence measure equals:

$$\hat{Q}(\tau) = \frac{25(25)}{50}$$

$$[\text{Dist. Between } \boldsymbol{X}_{1:25}, \boldsymbol{X}_{26:50} - \text{Dist. Within } \boldsymbol{X}_{1:25} - \text{Dist. Within } \boldsymbol{X}_{26:50}] = 136.42$$

Indeed,

$$\begin{aligned}
\text{Dist . Between } \boldsymbol{X}_{1:25}, \boldsymbol{X}_{26:50} &= \frac{2}{25(25)}\sum_{i=1}^{25}\sum_{j=26}^{50}\|X_i-X_j\| \\
&= \frac{2}{25(25)}\left\{\sqrt{(2.35-1.08)^2+(1.98-4.37)^2+(3.77-7.36)^2}\right. \\
&\quad + \sqrt{(2.35-4.50)^2+(1.98-7.61)^2+(3.77-10.44)^2} + \ldots \\
&\quad \left. + \sqrt{(2.27-3.77)^2+(2.10-6.72)^2+(3.36-9.67)^2}\right\} \\
&= 15.28
\end{aligned}$$

$$\begin{aligned}
\text{Dist . Within } \boldsymbol{X}_{1:25} &= \binom{25}{2}^{-1}\sum_{i=1}^{24}\sum_{k=i+1}^{25}\|\boldsymbol{X}_i-\boldsymbol{X}_k\| \\
&= \binom{25}{2}^{-1}\left\{\sqrt{(2.35-2.04)^2+(1.98-2.82)^2+(3.77-3.62)^2}\right. \\
&\quad + \sqrt{(2.35-2.52)^2+(1.98-2.07)^2+(3.77-4.29)^2} + \ldots \\
&\quad \left. + \sqrt{(1.19-2.27)^2+(1.72-2.10)^2+(3.86-3.36)^2}\right\} \\
&= 1.90
\end{aligned}$$

[3] Matteson and James (2014) mention the option of raising the Euclidean distances to a power α, with 0< α <2. In this paper, we used the default α-value of 1, since the same authors claim that similar results are obtained when other α -values are used.

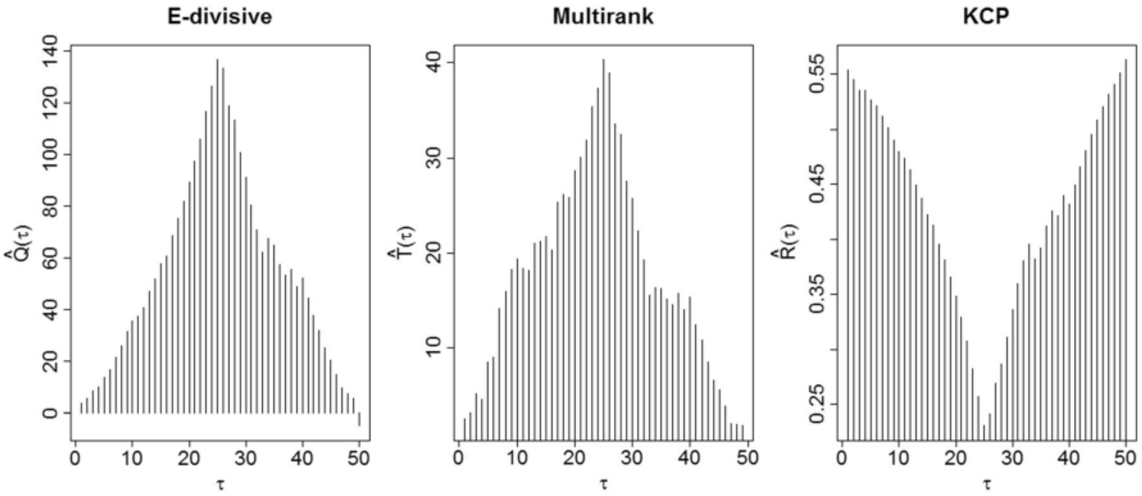

Fig. 4 Optimization of the E-divisive, Multirank, and KCP segmentation statistics over all possible change point locations, $\tau \in 1:n$, for the hypothetical data. The first panel shows the maximization of the divergence measure, $\hat{Q}$, for E-divisive. The second panel displays the maximization of the homogeneity statistic, $\hat{T}$, for Multirank. The last panel exhibits the minimization of the variance-like criterion, $\hat{R}$, for KCP. For all three methods, the statistics were optimal when $\tau = 25$, implying that a change point occurred at $T = 26$

$$\begin{aligned}
\text{Dist. Within } \boldsymbol{X}_{26:50} &= \binom{25}{2}^{-1} \sum_{j=26}^{49} \sum_{k=j+1}^{50} \|\boldsymbol{X}_j - \boldsymbol{X}_k\| \\
&= \binom{25}{2}^{-1} \left\{ \sqrt{(1.08-4.50)^2 + (4.37-7.61)^2 + (7.36-10.44)^2} \right. \\
&+ \quad \sqrt{(1.08-2.01)^2 + (4.37-5.26)^2 + (7.36-7.88)^2} + \ldots \\
&\left. + \sqrt{(2.37-3.77)^2 + (6.19-6.72)^2 + (9.41-9.67)^2} \right\} \\
&= 2.46
\end{aligned}$$

The optimal estimate of the change point location can be derived by inspecting which τ-value maximizes $\hat{Q}$. For our illustrative time series, Fig. 4 (first panel) shows the $\hat{Q}$ -values that are obtained when τ is varied from 1 to 50. As expected, $\hat{Q}$ is maximal for τ = 25, implying that the distribution changes after $T = 25$, generating a change point estimate at $T = 26$.

2. *Determine if the change point is significant through a permutation test.*

After estimating the change point location, its significance is tested by means of a permutation test on the maximal $\hat{Q}$ -value. This test is conducted by generating R permuted time series that are obtained by randomly changing the time order of the sequence. Step 1 is applied to each of these permuted sequences, yielding R new maximal $\hat{Q}$ -values. The p-value of the permutation test equals the percentage of permuted sequences that generated a larger maximal $\hat{Q}$ -value than the one obtained for the original sequence. For the illustrative data, the p-value for the maximal $\hat{Q}$ is 0.002, implying that the change point, $T = 26$, is considered significant (at a pre-specified significance level of 0.05).

3. *Divide the sequence into separate phases according to the detected change point and look for further change points in each of them.*

If the change point corresponding to the maximal $\hat{Q}$ -value obtained in Step 1 is found significant in Step 2, the sequence is divided into the corresponding phases. Steps 1 to 3 are then applied within each phase to detect additional change points, yielding a change point detection process that is hierarchically structured. Applying the procedure to the hypothetical data, the time series is split into two phases after the first change point was found significant in Step 2. The first phase, $\boldsymbol{X}_{1:25}$, was further bisected, and the optimal change point estimate was $T = 11$. For the second phase, $\boldsymbol{X}_{26:50}$, the optimal change point location was $T = 41$. The p-values were 0.668 and 0.268, respectively, for the first and second phases, implying non-significance of these additional change points. The phases were not further bisected and it was concluded that the time series contains one change point only.

Multirank

Multirank makes use of a homogeneity statistic, which is a multivariate extension of the Kruskal-Wallis test statistic (Lung-Yut-Fong, Lévy-Leduc, & Cappé, 2012). Hence, Multirank only takes the rank order of the scores per variable into account. The method consists of two steps.

1. *Check whether the time series contains at least one significant change point.*

Considering all possible τ-values, the sequence is divided into two phases $\boldsymbol{X}_{1:\tau}$ and $\boldsymbol{X}_{\tau+1:n}$. For each τ-value, the dissimilarity of these phases is determined by computing the following homogeneity statistic

$$\hat{T}(\tau) = \frac{4}{n^2}\left[(\tau)\overline{\boldsymbol{R}}_1'\hat{\boldsymbol{\Sigma}}^{-1}\overline{\boldsymbol{R}}_1 + (n-\tau)\overline{\boldsymbol{R}}_2'\hat{\boldsymbol{\Sigma}}^{-1}\overline{\boldsymbol{R}}_2\right] \qquad (3)$$

where $\hat{\boldsymbol{\Sigma}}$ is the empirical covariance matrix of the rank orders of the scores, and $\overline{\boldsymbol{R}}_k$ is a phase specific vector containing deviations of the observed mean phase ranks from the expected mean phase rank if the whole sequence is homogeneous. In case of homogeneity, the rank order of a score is completely random and, thus, the expected mean rank within a phase equals $\frac{n+1}{2}$. However, if a change point segments the sequence into phases with different distributions, the rank orders would not be random anymore but dependent on the distributions. Consequently, the deviations of the mean phase ranks from the expected rank under homogeneity, and thus also $\hat{T}$, would be large. Hence, to decide whether the time series contains at least one change point, the significance of the highest $\hat{T}$ –value is tested by computing the associated asymptotic *p*-value under the assumption of homogeneity. Details on this computation, which is based on Bessel functions of the first kind and the gamma function, can be found in Lung-Yut-Fong et al. (2012).

Figure 4 (second panel) displays the $\hat{T}$ -values that were obtained for our illustrative example using different τ-values, and indicates that τ = 25 yields the highest $\hat{T}$ -value. This implies a possible change point at the 26th observation. Specifically, the maximal homogeneity statistic equals 40.27, since

$$\overline{\boldsymbol{R}}_1 = \begin{bmatrix} \frac{R_1^{(1)}+R_2^{(1)}+R_3^{(1)}+\ldots+R_{25}^{(1)}}{25}-\frac{n+1}{2} \\ \frac{R_1^{(2)}+R_2^{(2)}+R_3^{(2)}+\ldots+R_{25}^{(2)}}{25}-\frac{n+1}{2} \\ \frac{R_1^{(3)}+R_2^{(3)}+R_3^{(3)}+\ldots+R_{25}^{(3)}}{25}-\frac{n+1}{2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{31+26+33+\ldots+30}{25}-25.5 \\ \frac{14+22+15+\ldots+16}{25}-25.5 \\ \frac{19+16+24+\ldots+15}{25}-25.5 \end{bmatrix} = \begin{bmatrix} -9.22 \\ -12.30 \\ -12.50 \end{bmatrix},$$

$$\overline{\boldsymbol{R}}_2 = \begin{bmatrix} \frac{R_{26}^{(1)}+R_{27}^{(1)}+R_{28}^{(1)}+\ldots+R_{50}^{(1)}}{25}-\frac{n+1}{2} \\ \frac{R_{26}^{(2)}+R_{27}^{(2)}+R_{28}^{(2)}+\ldots+R_{50}^{(2)}}{25}-\frac{n+1}{2} \\ \frac{R_{26}^{(3)}+R_{27}^{(3)}+R_{28}^{(3)}+\ldots+R_{50}^{(3)}}{25}-\frac{n+1}{2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{12+49+25+\ldots+50}{25}-25.5 \\ \frac{29+49+31+\ldots+40}{25}-25.5 \\ \frac{29+48+31+\ldots+41}{25}-25.5 \end{bmatrix} = \begin{bmatrix} 9.22 \\ 12.30 \\ 12.50 \end{bmatrix}$$

and

$$\hat{\Sigma}^{-1} = \begin{bmatrix} 8.48 & -0.83 & -6.09 \\ -0.83 & 11.75 & -9.45 \\ -6.09 & -9.45 & 16.03 \end{bmatrix}.$$

In Step 1, $\overline{\boldsymbol{R}}_2$ is always equal to $-\overline{\boldsymbol{R}}_1$, since we are looking for one change point. When considering multiple change points, this property will of course not hold. The associated *p*-value for the maximal $\hat{T}$ is 1.38×10^{-7}, confirming that the change point, $T = 26$, is highly significant. Henceforward, we will denote the maximal $\hat{T}$ as $\hat{T}_{max}$.

2. *Decide on the number of change points and on their location.*

If the change point obtained in Step 1 is found to be significant, multiple change point detection is conducted by computing the generalized form of the homogeneity statistic in Eq. 3, where K denotes the number of change points, $\tau_0 = 0$ and $\tau_{K+1} = n$:

$$\hat{T}(\tau_1, \tau_2, \ldots \tau_K) = \frac{4}{n^2}\sum_{k=0}^{K}(\tau_{k+1}-\tau_k)\overline{\boldsymbol{R}}_k'\hat{\boldsymbol{\Sigma}}^{-1}\overline{\boldsymbol{R}}_k \qquad (4)$$

To determine the number of change points and their location, K is varied from 0 to $K_{\max}$. For each K-value, the phase boundaries, $\tau_1, \tau_2, \ldots, \tau_K$, in Eq. 4 are varied and the homogeneity statistic, $\hat{T}$, for the resulting phases is computed. The change point locations that generate the maximal homogeneity statistic, $\hat{T}_{max}$, are stored (see Table 1). Next, the $\hat{T}_{max}$ values are plotted against the number of change points K (see Fig. 5, left panel). To choose the optimal K, two linear regressions are performed for each K -one starting from 0 up to K and one on the points from K onwards. The total residual sum of squares of both regressions[4] is then computed. The K-value associated with the lowest sum is retained as the optimal estimate of the number of change points. Based on Table 1, the

[4] The before and after regression is only done for $K = 1, \ldots K_{max}$, since the test for K being equal to zero or not is already conducted in Step 1.

Table 1 Maximal Multirank homogeneity statistic, $\hat{T}_{\max}$, total residual sum of squares of the before and after K regressions and estimated change point locations for the hypothetical data

K	$\hat{T}_{max}$	Total residual sum of squares	Change points
0	0.00	-	-
1	40.27	10.22	26
2	47.55	189.54	9, 26
3	57.67	291.47	19, 21, 26
4	62.99	427.62	19, 21, 23, 26
5	67.71	565.21	4, 9, 19, 21, 26

K-value with the lowest total residual sum of squares is $K = 1$, revealing that our hypothetical data contain only one change point, located at $T = 26$. This is easily spotted as well in Fig. 5 (left panel), where $K = 1$ generated the best before and after regression fit as shown by the black lines.

KCP

The Kernel Change Point (KCP) method proposed by Arlot et al. (2012) detects change points by evaluating how similar or dissimilar the scores at the observed time points are to each other. To this end, the observations are transformed to similarities by means of a kernel function (Shawe-Taylor & Christianini, 2004). In this paper, like Arlot et al. (2012), we used the Gaussian kernel, which is the most widely applied kernel in the literature (Sriperumbudur, Gretton, Fukumizu, Lanckriet, & Scholkopf, 2010).

1. *Compute pairwise similarities using a Gaussian kernel function.*

For each pair of observations, $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, the pairwise similarity is computed using a Gaussian kernel function,

$$k\left(\boldsymbol{X}_i,\, \boldsymbol{X}_j\right) = \exp\left(\frac{-\left\|\boldsymbol{X}_i - \boldsymbol{X}_j\right\|^2}{2h^2}\right)$$

The similarities take on values close to 0 when $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are distant and values close to 1 when $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are similar. The bandwidth, h, is a smoothing parameter that indicates how strict one is when deciding if two observations are similar (see examples of usage in Hastie, Tibshirani, & Friedman, 2009). In this paper, we determined the bandwidth using the procedure of Arlot et al. (2012) which sampled 250 observations $\boldsymbol{X}_i$ from the whole time series and set h to the median Euclidean distance among those 250 observations.

2. *For different numbers of change points K, minimize the total intra-phase scatter to detect their location.*

For varying numbers of change points, $K = 0, \ldots, K_{max}$, KCP minimizes the following criterion across all possible change point locations $(\tau_1, \tau_2, \ldots, \tau_K)$:

$$\hat{R}(\tau_1, \tau_2, \ldots, \tau_K) = \frac{1}{n}\sum_{k=0}^{K} \hat{V}_k$$

where $\hat{V}_k$ is the intra-phase scatter. $\hat{V}_k$ measures how homogeneous the corresponding phase is,

$$\hat{V}_k = (\tau_k - \tau_{k-1}) - \frac{1}{\tau_k - \tau_{k-1}} \sum_{i=\tau_{k-1}+1}^{\tau_k} \sum_{j=\tau_{k-1}+1}^{\tau_k} k\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)$$

Indeed, the more similar the observations in a segment, $\boldsymbol{X}_{\tau_{k-1}+1}$: $\boldsymbol{X}_{\tau_k}$, are, the larger the sum that is subtracted by the

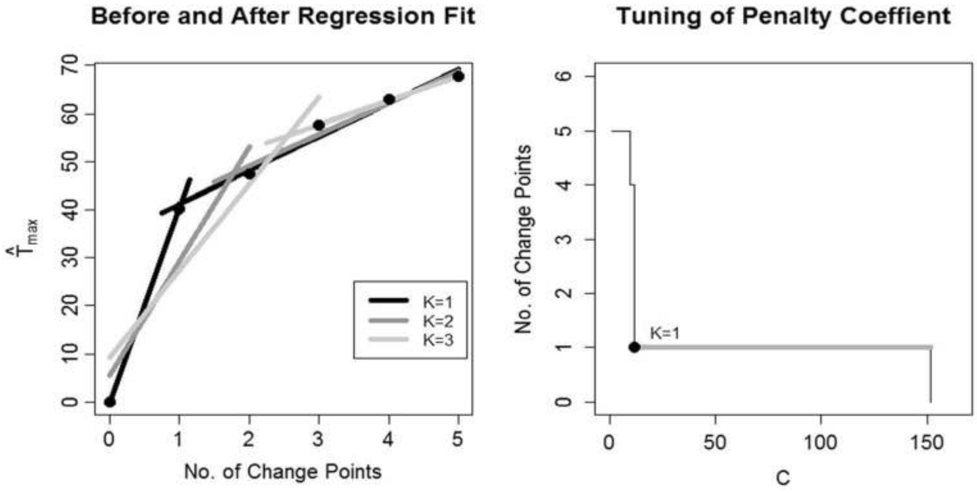

Fig. 5 MultiRank and KCP heuristic procedures for choosing the number of change points, K, for the hypothetical data. The left panel shows the $\hat{T}_{max}$ vs. K plot for Multirank, where the best before and after regression fit is generated when K is set to 1. The right panel displays the tuning of the penalty coefficient, C, for KCP, which generated $K = 1$ as the most stable K

Table 2 Minimal KCP criterion, $\hat{R}_{\min}$, and change point locations for different values of K for the hypothetical data

K	$\hat{R}_{min}$	Change points
0	0.56	-
1	0.23	26
2	0.21	26, 41
3	0.19	26, 34, 41
4	0.17	26, 27, 34, 41
5	0.16	26, 27, 34, 35, 41

rightmost term of $\hat{V}_k$ and thus the smaller the intra-phase scatter. Moreover, the smaller are the $\hat{V}_k$'s for all k, the smaller the criterion, $\hat{R}$, will be. For example, Fig. 4 (third panel) which shows the $\hat{R}$ -values obtained when looking for the optimal location of a single change point ($\tau = \tau_1$) in our illustrative time series, reveals that $\hat{R}$ is minimal for $\tau = 25$, creating segments, $\boldsymbol{X}_{1:25}$ and $\boldsymbol{X}_{26:50}$. Specifically, the intra-phase scatters for these two phases are

$$\hat{V}_1 = 25 - \frac{1}{25}\sum_{i=1}^{25}\sum_{j=1}^{25} k\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) = 4.55$$

$$\hat{V}_2 = 25 - \frac{1}{25}\sum_{i=26}^{50}\sum_{j=26}^{50} k\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) = 7.00,$$

generating the minimal criterion value, $\hat{R}(\tau = 25) = \frac{\hat{V}_1 + \hat{V}_2}{n} = \frac{11.55}{50} = 0.23$. Henceforward, this minimal KCP criterion generated from the optimal change point locations will be denoted as $\hat{R}_{min}$. Applying the method for $K = 0$ to $K_{max} = 5$, yields the change points listed in Table 2.

3. *Decide on the optimal number of change points.*

After Step 2, what remains to be determined is the most appropriate number of change points, K. Since the $\hat{R}_{min}$ -values decrease with increasing K, Arlot et al. (2012) proposed to penalize the $\hat{R}_{min}$ -values for the additional complexity that is introduced by allowing for extra change points. Specifically, they select the number of change points which minimizes

$$crit_K = \hat{R}_{min} + pen_K, \tag{5}$$

where $pen_K = C\frac{V_{max}\,(K+1)}{n}\left[1 + \log\left(\frac{n}{K+1}\right)\right]$. The constant, C, is a tuning parameter that controls the influence of the penalty term (see below). The remaining constant, v_{max}, is determined by computing the trace of the estimated covariance matrix for the first 5 % time points as well as for the last 5 % time points, and choosing whichever is larger.

As can be expected, the value chosen for C, greatly influences the performance of the method, where a smaller C favors numerous change points, while a larger C causes undersegmentation. Whereas in previous simulations (Matteson & James, 2014) this tuning issue was dealt with by just setting a particular C value, without further motivation, we propose selecting C by plugging linearly increasing values starting from C = 1 into Eq. 5. When $C = 1$, the generated estimate for K is K_{max}. If C is increased, the effect of the penalty term is strengthened and the generated estimate for K becomes smaller. Thus, the procedure terminates when C becomes so high that the associated estimate for K equals 0. Based on the theoretical motivations in Lavielle (2005), the K-value that is selected most often in this grid search is retained as the optimal number of change points. Figure 5 (right panel) shows the K-values that are selected across different C-values between 1 and 151.9 for our illustrative example. Since the mode of the selected K-values equals 1, as could be expected, we decided that the time series contains one change point. Note that in case only K_{max} and 0 are selected in the grid search, it should be concluded that the time series contains no change points (see details in Lebarbier, 2005).

Software

For DeCon and for the tuning steps of the other methods, Matlab codes are available upon request from the first author of this paper. E-divisive can be applied using the *ecp* package in R. Multirank was programmed in Python, and the codes can be requested from the second author of the corresponding paper (Lung-Yut-Fong et al., 2012). For KCP, R codes are included in the supplementary files provided by Matteson and James (2014). The hypothetical data used for illustrating the methods was simulated with the "mvtnorm" R-package and can be obtained from the first author as well.

Simulation studies

Two simulation studies will be performed to compare the four methods, i.e., DeCon, E-divisive, Multirank and KCP, using the settings of Bulteel et al. (2014) and Matteson and James (2014). Neither of those earlier studies compared all four methods examined in this paper.

Simulation settings of Bulteel et al. (2014): Mean changes, correlation changes or both

The first simulation study is conducted to compare the performance of DeCon, E-divisive, MultiRank and KCP in detecting changes in mean, changes in correlation, or both. In particular, we used the simulation settings of Bulteel et al. (2014) to generate time series of 300 time points with 5 variables. Each time series consisted of two phases containing 150 time points each. The time series varied with respect to the following three factors, which were fully crossed with 1,000 replicates per cell of the design:

1. *Change in mean between the two phases (three levels)*: There could be no change, an increase of 1 standard deviation for 3 variables, or an increase of 2 standard deviations for three variables (in the latter two levels, the mean of the other two variables remains the same).
2. *Change in correlation structure between the two phases (two levels):* The correlation structure of the variables is manipulated by generating true scores according to a principal component model. For settings with no correlation change, a 300×5 matrix was generated according to a model with 3 components. For settings with correlation change, two 150×5 matrices were generated, where the first one is based on three components and the second one on two components. The loadings on these components were sampled from a uniform distribution on the interval (-1,1). The component scores, on the other hand, were drawn from a standard normal distribution. Note that the loadings and the error values were rescaled to obtain data that contain 25 % noise.
3. *Strength of autocorrelation within the phases (three levels):* 0, .3 and .7.

   Each simulated time series was constructed by adding true scores and noise. The noise was sampled from a multivariate normal distribution having a zero vector as its mean vector and the identity matrix as its covariance matrix. Next, we imposed a lag-one autocorrelation on these noise scores by means of a recursive filter (Hamilton, 1994).

All four methods under study were then applied to these simulated data sets. The tuning parameters for each method are tabulated in Table 3. For E-divisive, default settings by Matteson and James (2014) were maintained imposing a maximum of ten phases and a minimum phase size of 30. Equivalently, for Multirank and KCP, K_{max}[5] was set to 9. For DeCon, on the other hand, a window size of 75 was chosen to impose a minimum phase size[6] of $.25WS \approx 19$.

To quantify how well the four methods revealed the underlying phases, we computed the Rand Index (RI) between the recovered phases and the true phases. An RI value of 1 implies perfect recovery of the underlying phases, while 0 implies that recovered and underlying phases do not resemble one another (Rand, 1971). We also recorded the detected number of change points.

Results show that KCP outperforms the three other methods, exhibiting the highest RIs in almost all settings (Table 4). It also proved to be the most robust to the presence of autocorrelation, which leads to false detections for the other methods. All methods succeeded in detecting changes in mean, though DeCon performed worse for settings with a small mean change (1 standard deviation). Furthermore, change in correlation (without change in mean) proved to be harder to detect. Though KCP (RI≥ 0.94) and DeCon (RI≥0.87) still showed acceptable detection performance in these settings, performance of Multirank (RI≤0.61) and E-divisive (RI≤ 0.80) was inadequate. The RI-values for Multirank were close to 0.50, because the method either did not detect the change point and concludes that the time series contains only a single phase (no or weak autocorrelation) or yields too many change points (strong autocorrelation), rather than the correct two phases. Finally, for settings where both changes in mean and in correlation were introduced, all methods retrieved the change point in most cases. Note that a repeated measures ANOVA with method as within subjects factor and size of mean change, size of correlation change, and size of autocorrelation as between subjects factors, revealed that two effects had a generalized effect size (η_G^2) larger than .13, indicating a medium effect size (Bakeman, 2005): size of mean change (η_G^2=.27) and its interaction with the method used (η_G^2=.18).

Simulation settings of Matteson and James (2014): Correlation change and presence of noise variables

The first simulation study already showed that changes in correlation structure are more difficult to detect. To get further insight into the performance of the four methods in revealing correlation change, we ran a second study[7] based on the simulation settings of Matteson and James (2014). An interesting feature of these settings is that they allow to investigate how performance is affected by the presence of noise variables. Noise variables are variables that do not change in means and correlations. Specifically, Matteson and James (2014) generated normally distributed time series that consist of three phases of equal length. In the first phase, the variables are uncorrelated. In the middle phase, they become strongly correlated, such that their pairwise correlations amount to 0.9. And in the final phase, the variables are uncorrelated again. Three factors were manipulated, with 1,000 replicates per possible combination:

[5] $Max\,Number\,of\,Phases = \frac{300}{\min phase\ size} = K_{max} + 1 \approx 10.$

[6] The window size for DeCon cannot be tuned to a minimum phase size of 30 since that would imply choosing a window size equal to 120, which is larger than the real phase size of 100.

[7] Tuning parameters in the first simulation study were maintained in the second simulation study, with the exception of K_{max} being set to 19 and 29 for settings with N = 600 and N = 900, respectively.

Table 3 Tuning parameters for the four change point detection methods: first simulation study based on simulation settings of Bulteel et al. (2014)

Method	Initial parameters
Decon	Window Size = 75, Moving Sum Cut-off = 5/10
E-divisive	Min. Cluster Size = 30, R = 499, Significance level = 0.05
Multirank	$K_{max} = 9$
KCP	$K_{max} = 9$

1. *Number of variables D (three levels):* 2, 5, and 9.
2. *Number of time points n (three levels):* 300, 600, and 900.
3. *Number of noise variables (two levels)*: 0 (i.e., all variables correlate in the middle phase) and number of variables minus two (only two variables become correlated in the middle phase).

When no noise variables are present, KCP is the best method in all conditions but one (two variables, 300 time points), with RI values larger than 0.97 (see Table 5). Multirank, on the other hand, consistently failed, being the worst method. Its RI values were close to 0.33, because no change points are detected, thus generating 1 phase only, instead of the three underlying phases. A repeated measures ANOVA, with method as within subjects factor and number of variables and number of time points as between subjects factors, revealed that RI was indeed clearly influenced by method (η_G^2=.87), as well as by its two-way interaction with number of variables (η_G^2=.34) and three-way interaction with number of variables and number of time points (η_G^2=.15); the main effect of number of variables (η_G^2=.29) was strong as well.

When noise variables were present, DeCon was the clear winner, with RIs being consistently larger than 0.81. Moreover, its RI performance was not extremely affected by the number of noise variables or the number of time points, although both factors have an impact on the number of detected change points. All the other methods yielded inadequate RI values in all with noise settings and thus are severely affected by the presence of noise variables. In almost all settings (except for KCP on five variables and 900 time points), their RI values were close to 0.33, because no change point was detected for most data sets. Not surprisingly, the repeated measures ANOVA revealed that the main effect of method (η_G^2=.72) explained the bulk of the differences in the RIs for settings with noise.

In summary, the following conclusions can be drawn. First, KCP and Multirank seem to be reliable methods for detecting mean changes, whereas E-divisive and DeCon often yield false change points. Second, KCP and DeCon are the best methods for detecting correlation change, although KCP often fails if noise variables are present. DeCon is too sensitive however and frequently yields false positives. Thus, change points that are only found by DeCon should be approached cautiously: They can signal real correlation changes as well as false positives.

Illustrative application

Change point detection

We further assessed the performance of the methods by applying them to multivariate time series data obtained from a study on cardiorespiratory assessment of mental load in the field of aviation (Grassmann, Vlemincx, von Leupoldt, & Van den Bergh, in press). Male pilot applicants were subjected to four experimental periods: a resting baseline, a "vanilla" baseline, a highly demanding multiple task and a recovery period. During the resting baseline, participants were instructed to fix their eyes on a cross that was presented on the screen. In the vanilla baseline, they were asked to complete a minimally demanding vigilance task which was intended to reduce anticipatory arousal, hence improving the validity of baseline measures (see Jennings, Kamarck, Stewart, Eddy, & Johnson, 1992). In the multiple task period, participants had to perform three tasks simultaneously, tapping perceptual speed, spatial orientation and working memory (for a detailed description see Grassmann et al., in press). Finally, during the recovery period, participants watched a relaxing underwater movie. Each period lasted for 6 min, however the first and last 30 s were cut before data processing to procure stationary data, and to exclude artifacts that were occasionally caused by speech and movement during the periods of transition. Heart rate, respiration rate and partial pressure of end-tidal CO2 (petCO2) were monitored throughout the experiment.

Based on previous findings, the means of all three physiological variables were expected to change across the phases (e.g., Backs & Seljos, 1994; Brookings, Wilson, & Swain, 1996; Veltman & Gaillard, 1998; Wientjes, Grossman, & Gaillard, 1998). Heart rate, for instance, was hypothesized to decrease during the vanilla baseline and to increase during the multiple task while readjustments were expected for the recovery period (Jennings et al., 1992). Regarding correlation changes, we expect an increase in the correlation of cardiorespiratory variables in the vanilla baseline, as it requires focused attention and low cognitive activity (Wu & Lo, 2010). During the multiple task, where tasks are more highly demanding, a decrease in correlation could occur (Zhang, Yu, & Xie, 2010).

The study included 115 pilot applicants; however, for this paper, we analyzed data from a single randomly chosen pilot. The variables were initially measured in different frequencies. Cardiac data (sampled at 1,000 Hz) were processed beat-by-beat whereas respiratory data (sampled at 20 Hz) were processed breath-by-breath. For the present analyses, common time points were re-aligned by up-sampling the respiratory

Table 4 Mean Rand indices and number of detected change points: first simulation study based on simulation settings of Bulteel et al. (2014) with one real change point

ΔCorr	ΔMean	AR Coeff	E-divisive		Multirank		KCP		DeCon	
			RI	No. of CPs	RI	No. of CPs	RI	No. of CPs	RI	No. of CPs
0	0	0.0	0.97	0.06	1.00	0.03	0.91	0.49	0.97	0.08
		0.3	0.94	0.14	0.96	0.30	0.90	0.53	0.94	0.14
		0.7	0.68	1.24	0.39	5.32	0.81	0.94	0.75	0.72
	1	0.0	0.97	1.05	1.00	1.00	1.00	1.00	0.81	0.71
		0.3	0.97	1.12	0.99	1.01	1.00	1.00	0.82	0.76
		0.7	0.90	1.89	0.97	1.34	0.99	1.00	0.85	1.40
	2	0.0	0.99	1.05	1.00	1.00	1.00	1.00	0.98	1.09
		0.3	0.98	1.14	1.00	1.00	1.00	1.00	0.97	1.10
		0.7	0.91	1.91	1.00	1.01	1.00	1.00	0.94	1.62
1	0	0.0	0.77	0.68	0.50	0.07	0.97	1.07	0.87	1.01
		0.3	0.78	0.81	0.51	0.35	0.96	1.11	0.88	1.11
		0.7	0.80	1.79	0.61	4.40	0.94	1.21	0.88	1.79
	1	0.0	0.98	1.07	1.00	1.00	1.00	1.00	0.96	1.33
		0.3	0.98	1.13	0.99	1.01	1.00	1.00	0.96	1.40
		0.7	0.91	1.78	0.97	1.30	0.99	1.00	0.91	1.98
	2	0.0	0.99	1.06	1.00	1.00	1.00	1.00	0.97	1.33
		0.3	0.98	1.13	1.00	1.00	1.00	1.00	0.97	1.37
		0.7	0.92	1.80	1.00	1.00	1.00	1.00	0.92	1.93

data (i.e., respiration rate and petCO2 values of one breath were assigned to each heart rate value that was initiated within the corresponding respiratory cycle). It is also important to note that variables were all scaled to have a variance of 1 as three methods, E-divisive, DeCon, and KCP, calculate distance measures which are influenced by the scale of the data.

Table 5 Mean Rand indices and number of detected change points: second simulation study based on simulation settings of Matteson and James (2014) with two real change points

Noise	N	D	E-divisive		Multirank		KCP		DeCon	
			RI	No. of CPs	RI	No. of CPs	RI	No. of CPs	RI	No. of CPs
None	300	2	0.40	0.23	0.33	0.03	0.76	2.15	0.86	1.79
		5	0.91	1.93	0.34	0.09	0.98	2.24	0.96	2.65
		9	0.98	2.14	0.34	0.24	0.98	2.21	0.94	2.76
	600	2	0.53	0.68	0.34	0.06	0.97	2.21	0.88	1.88
		5	0.99	2.07	0.34	0.14	1.00	2.03	0.96	3.15
		9	0.99	2.07	0.35	0.30	1.00	2.01	0.90	4.59
	900	2	0.91	1.80	0.33	0.04	0.99	2.03	0.89	1.92
		5	1.00	2.07	0.34	0.20	1.00	2.00	0.96	3.35
		9	1.00	2.08	0.35	0.42	1.00	2.00	0.87	5.99
With	300	5	0.36	0.11	0.33	0.03	0.39	0.39	0.81	1.53
		9	0.36	0.12	0.33	0.01	0.35	0.12	0.87	2.21
	600	5	0.37	0.17	0.34	0.03	0.47	0.83	0.84	1.77
		9	0.36	0.13	0.34	0.04	0.36	0.19	0.88	3.75
	900	5	0.38	0.21	0.33	0.03	0.61	1.56	0.84	1.88
		9	0.37	0.15	0.33	0.03	0.37	0.22	0.85	5.13

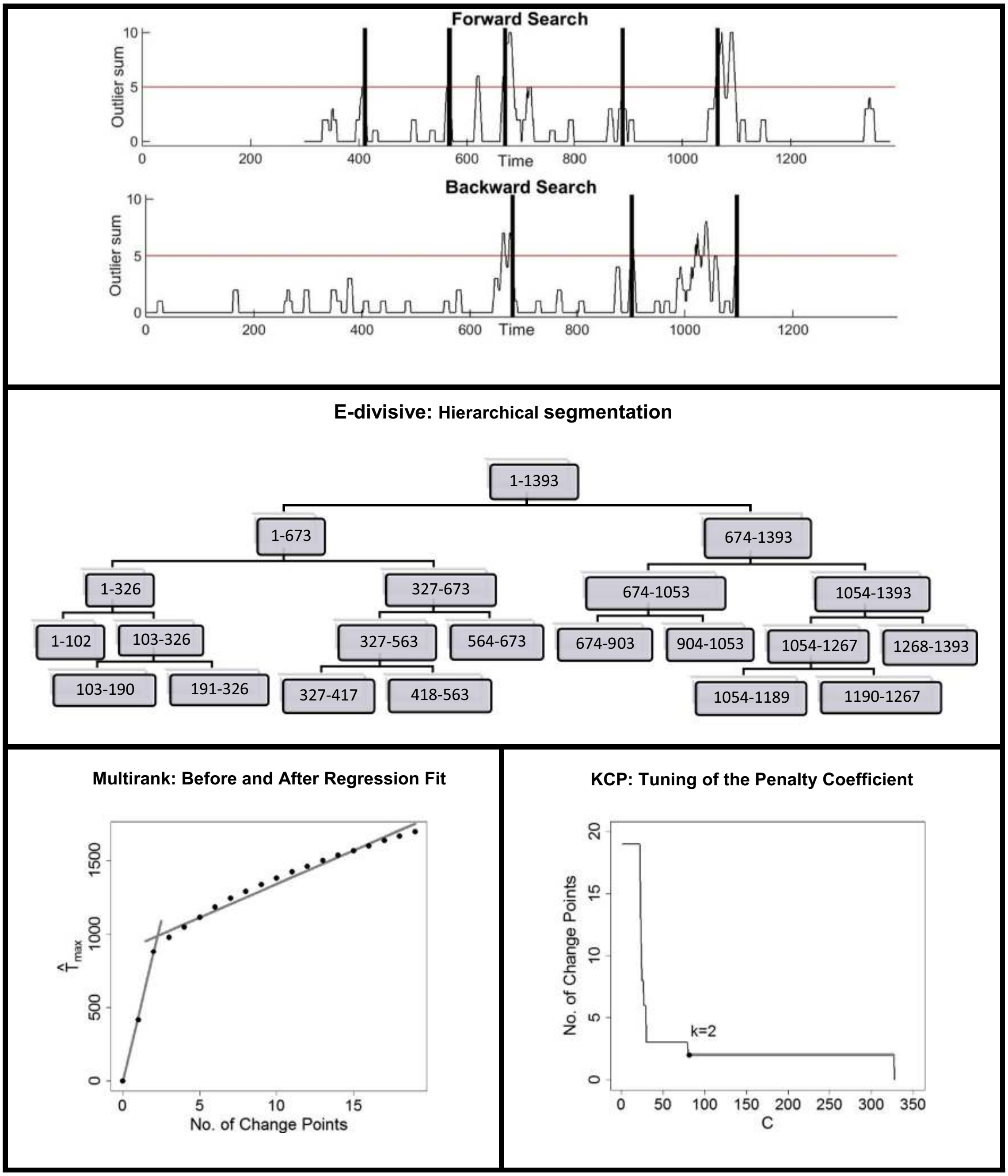

Fig. 6 Change point selection output of the four methods for the cardio-respiratory data. The topmost panel displays the DeCon moving outlier sum from the forward and the backward procedure, implying five change points. The next panel exhibits the hierarchical change point detection process by E-divisive, generating ten change points. The lowest left panel shows the $\hat{T}_{max}$ vs K plot for Multirank, indicating two change points. The lowest right panel demonstrates the linear tuning of the penalty coefficient for KCP, suggesting using two change points

Where possible, methods were initialized in such a way that at maximum 20 phases[8] would be discerned.

The change point detection results are displayed in Fig. 6. Employing DeCon, five change points were detected by the forward search (411, 568, 672, 889, 1,064), and three by the backward search (681, 901, 1,055). Given that the change point estimates from the backward search were considerably close to the last three change point estimates from the forward search, and given that the time series is quite long, we pooled

[8] The following tuning parameters were employed. DeCon: Window Size = 300, Moving Sum Cut-off = 5/10. E-divisive: Min. Cluster Size = 75, R = 499, significance level = 0.05. Multirank: K_{max} = 19. KCP: K_{max} = 19, C = 79.3

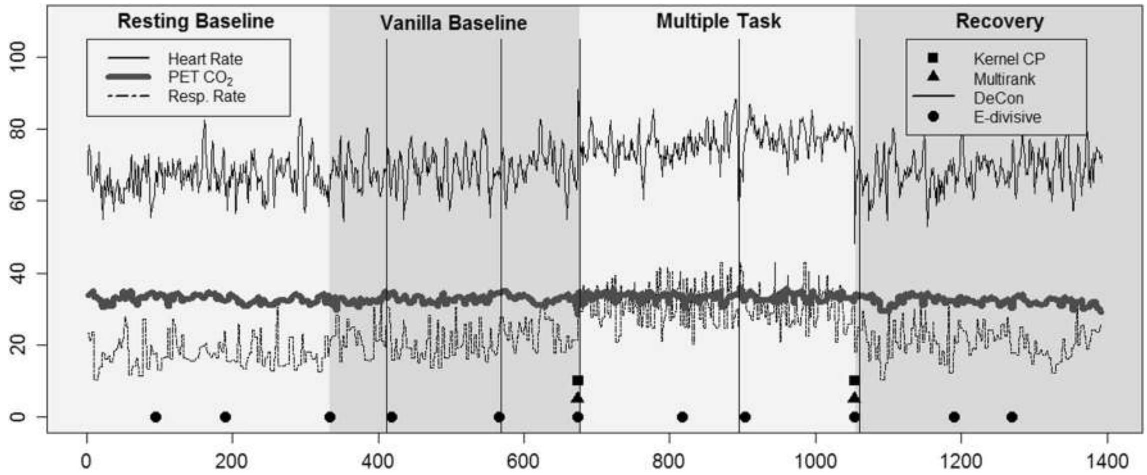

Fig. 7 Cardio-respiratory data and change points detected by the four methods. The experimental phases: resting baseline, vanilla baseline, multiple task, and recovery are indicated by the varying background shading

these change points by computing their means. Thus, the final set of change points generated by DeCon is (411, 568, 676, 895, 1,060).

E-divisive yields ten change points: 103, 191, 327, 418, 564, 674, 904, 1,054, 1,190, and 1,268, five of which are very close to the ones detected by DeCon. We initially attributed the five additional change points to Type 1 errors as E-divisive does not correct for multiple testing. However, changing the significance level did not dramatically change the results (nine change points for significance level = .01). Multirank and KCP both suggest that two change points might be present. Their change point estimates, 674 and 1,054 are identical, and were also obtained with DeCon and E-divisive. Examining Fig. 7, these two time points correspond to the boundaries of the Multiple task, confirming that the cardio respiratory measures from this specific pilot exhibited changes at the moment the highly demanding task was introduced as well as when the recovery period started. Given that in the simulations KCP and Multirank were reliable in detecting change points signaling changes in mean, whereas KCP and DeCon succeed rather well in revealing correlation change, we may say that the two common change points probably indicate changes in mean as well as changes in correlation.

Auxiliary analyses

To verify that both mean and correlation changed during the Multiple task (as hypothesized above on the basis of the simulation results), and to determine which variables specifically exhibited these changes, we conducted some auxiliary analyses. Focusing on mean changes, Mann-Whitney U tests revealed that the mean of all variables increased during the multiple task, and decreased again in the recovery period (see Fig. 8).

In order to check for correlation changes during the multiple task, we utilized the test for the difference of two correlations based on the Fisher's z-transformation of the sample correlation coefficients (Cohen, Cohen, West, & Aiken, 2003). On one hand, we concluded that heart rate and respiration rate became more negatively correlated during the multiple task, and correlated less during the recovery period. For petCO$_2$, no significant correlation changes were found during the transition to the multiple task. However, during the transition to recovery, petCO$_2$ significantly changed correlation with heart rate (negatively) and with respiration rate (positively).

Discussion and Conclusion

Change point detection in multivariate time series data presents a major data-analytical challenge because the variables involved can exhibit changes in means, in correlation, or in both (Terien et al., 2009). Detecting changes in correlation is crucial when one wants to understand the behavior of the system that is comprised of these variables. In this study, we compared the performance of four recently proposed non-parametric multivariate change point detection methods, focusing on changes in correlation.

In the first simulation study, Multirank and KCP, and to a somewhat lesser extent E-divisive, succeeded well in detecting mean changes. These results confirmed previous findings of Matteson and James (2014) regarding mean changes. DeCon, on the other hand, could only compete with these methods for large mean changes ($\Delta\, mean \geq 2sd$). Change in correlation without mean changes, proved to be harder to capture. For this specific setting, KCP and DeCon clearly outperformed E-divisive and Multirank. E-divisive missed to detect the change points in a considerable number of replicates, while Multirank failed in almost all cases. The second

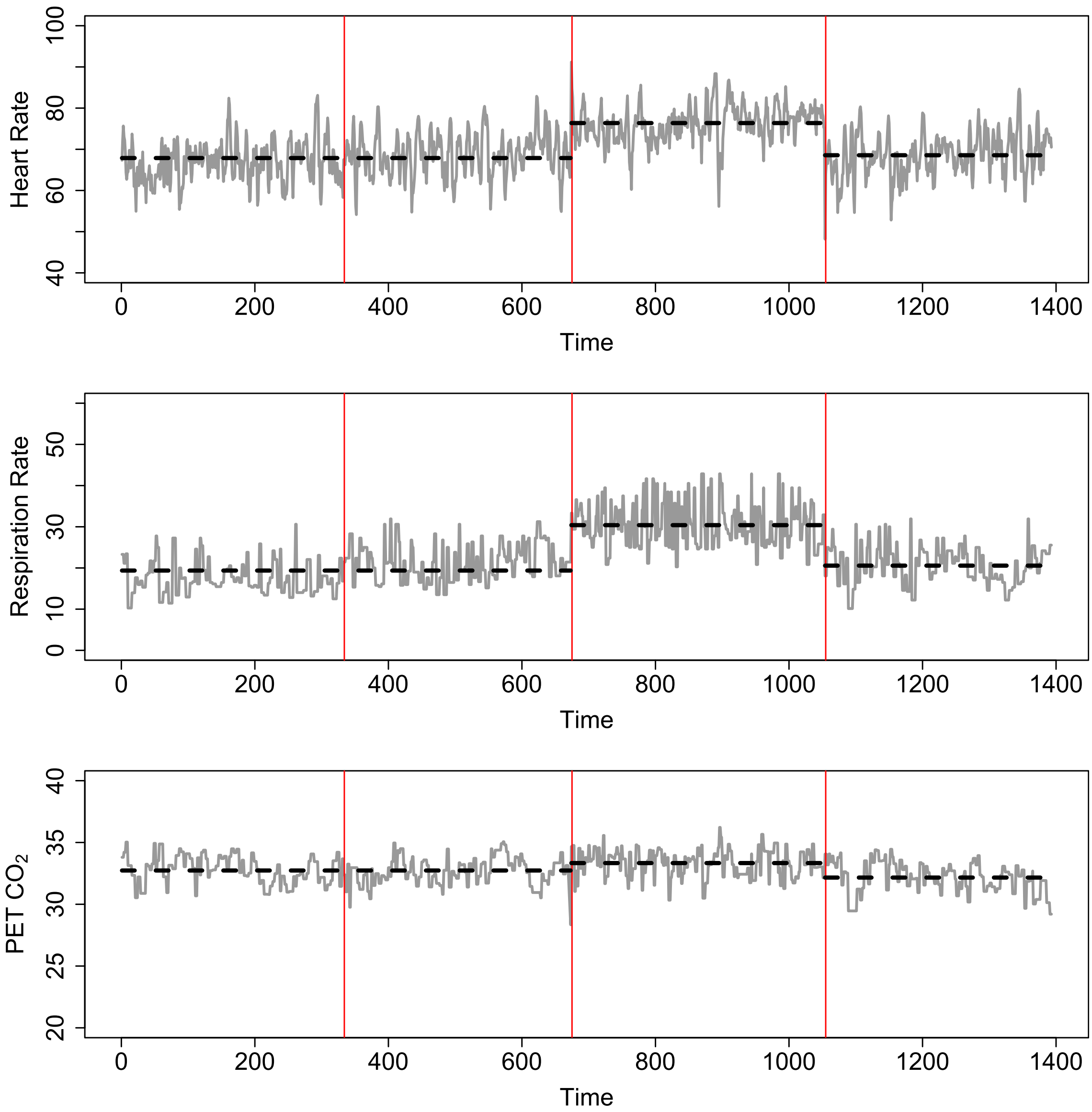

Fig. 8 Mean changes for the two change points, indicated by the vertical lines, that were detected by all four methods. Levels of cardio-respiratory variables increased in the second phase, then decreased again

simulation study revealed that when the correlation change to be detected is sizeable (no noise settings), KCP performs the best. These results are somewhat different from those reported by Matteson and James (2014) which suggested that KCP performs poorly compared to E-divisive and Multirank for a relatively small sample size (n = 300). These differences are attributed to the additional tuning step of the penalty coefficient, C, that we implemented for KCP. In contrast, our Multirank results are worse than those of Matteson and James (2014), because we included the significance test for a single change point, which was proposed by the original authors but not implemented by Matteson and James (2014), leading to false positives. It is important to note that DeCon was clearly the best method in detecting changes in correlation for settings with only two variables, as well as settings in which the majority of the variables were noise variables (with noise settings). All other methods performed badly in these settings, failing to detect all change points.

Overall, we thus conclude that which methods perform well strongly depends on the specific data setting. Therefore, we recommend using multiple methods in order to be more sensitive to different types of changes. However, we see a major issue that needs to be tackled when applying multiple methods. For the simulation study, it was straightforward to know which methods produced correct detections and which ones generated false positives because we introduced the changes ourselves. When applying the methods to real data, such as in the application section, this is almost always not the case. The task of deciphering which change points are important then lies in the hands of the user. Based on our simulation results, we provide the following advice: For detecting mean changes, one should inspect the changes that are detected by both KCP and Multirank. For detecting correlation changes, change points yielded by both KCP and DeCon are probably trustworthy. Lastly, when numerous variables are monitored, without prior knowledge whether some of them are noise variables, change points unique to DeCon should be scrutinized as well as they may signal correlation change.

Aside from the simulation results, an overview of the similarities and differences between the four change point detection methods under study with respect to statistical method, segmentation strategy and number of change points heuristic used could help an applied researcher in deciding which method or set of methods is appropriate for the data at hand, and could yield interesting directions for future methodological research. Regarding the statistical method used, Multirank is based on a multivariate version of the Kruskal-Wallis test statistic, which looks at deviations from the overall median, thus it is mainly sensitive to changes in levels. DeCon looks at score distances computed using a robust center and covariances, thus it is expected to pick up not just changes in levels but also in correlations. E-divisive and KCP, on the other hand, are both based on Euclidean distances which can be influenced by changes in any moments of the distribution. This explains why these methods can capture both mean and correlation changes. An extra feature of the similarity measure in KCP, though, is that it uses a non-linear transformation of this distance through a Gaussian kernel, magnifying the differences. Therefore, it is not unexpected that KCP performs better than E-divisive when the change introduced in the simulations was purely correlational. Regarding segmentation strategy, KCP and Multirank optimize an overall homogeneity statistic to locate multiple change points simultaneously. E-divisive employs binary segmentation such that only one change point is estimated at a time, leaving previously found change points untouched. DeCon on the other hand does not look for the optimal location for a change point, but indicates for every time point whether or not it is likely that a change occurred (because the time point is outlying with respect to the previous ones). When deciding on the optimal number of change points, the number of change points obtained with DeCon can hardly be controlled; the method for instance cannot be used to retrieve the three most likely change points. E-divisive employs a permutation test which is embedded in the hierarchical segmentation, but this test disrupts the natural ordering of time points and is not corrected for family wise error rate. Both KCP and Multirank use a heuristic procedure which weighs both the minimization (maximization) of the distance measure and the number of change points. This weighting proved to be effective in avoiding false detections for KCP and Multirank. One could postulate that generalizing the E-divisive divergence measure to more than two groups might decrease false detections. A pruning step, in which all change points are re-examined and only the most evident ones are retained, could possibly improve the performance of DeCon.

Finally, a common limitation of all considered change point detection methods is that they neither indicate which type of change (mean/correlation/both) occurred nor which variables are involved. Regarding the type of change, the methods under study might even indicate changes in higher moments. This is the price to pay, of course, when applying non-parametric methods as the distance measures used can be caused by numerous types of changes in the joint distribution. In contrast, most parametric methods monitor specific parameters (Chen & Gupta, 2012). Thus, when a change is detected, one immediately knows which type of changes was exhibited. Regarding the variables involved, the four non-parametric methods are not able to pinpoint which channels demonstrate the changes and which ones did not. To address this limitation, one could implement auxiliary analyses as we did for the illustrative application. However, this is cumbersome, especially when there are numerous recovered phases. Future research may therefore aim to determine the type of change and the variables involved during change point detection.

In conclusion, KCP was generally the best method in detecting changes in mean, changes in correlation, or both, and can therefore be recommended. When the goal is capturing changes in mean, results can be confirmed by Multirank, as it detects this type of change with comparable reliability. When the focus is capturing changes in correlation, we recommend inspecting DeCon change points as well. Although in general, DeCon performed less reliably than KCP, the method is quite sensitive to correlation changes, especially when the multivariate time series contains multiple noise variables.

Author Note The research leading to the results reported in this paper was sponsored in part by a research grant from the Fund for Scientific Research-Flanders (FWO, Project No. G.0582.14 awarded to Eva Ceulemans, Peter Kuppens and Francis Tuerlinckx), by the Belgian Federal Science Policy within the framework of the Interuniversity Attraction Poles program (IAP/P7/06), and by the Research Council of KU Leuven (GOA/15/003).

References

Amsel, E., & Renninger, K. A. (1997). *Change and development: Issues of theory, method and application*. Lawrence Erlbaum Associates.

Arlot, S., Celisse, A., & Harchaoui, Z. (2012). Kernel change-point detection. Retrieved from http://arxiv.org/abs/1202.3878

Aue, A., H rmann, S., Horváth, L., & Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics, 37*(6B), 4046–4087. doi:10.1214/09-AOS707

Backs, R. W., & Seljos, K. A. (1994). Metabolic and cardiorespiratory measures of mental effort: The effects of level of difficulty in a working memory task. *International Journal of Psychophysiology, 16,* 57–68. doi:10.1016/0167-8760(94)90042-6

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*(3), 379–384. doi:10.3758/BF03192707

Basseville, M., & Nikiforov, I. (1993). *Detection of abrupt changes: Theory and application*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Bhattacharya, G., & Johnson, R. (1968). Nonparametric tests for shift at an unknown time point. *The Annals of Mathematical Statistics, 39*(5), 1731–1743. doi:10.1214/aoms/1177698156

Brookings, J., Wilson, G., & Swain, C. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology, 42,* 361–377. doi:10.1016/0301-0511(95)05167-8

Bulteel, K., Ceulemans, E., Thompson, R., Waugh, C., Gotlib, I., Tuerlinckx, F., & Kuppens, P. (2014). DeCon: A tool to detect emotional concordance in multivariate time series data of emotional responding. *Biological Psychology, 98*(1), 29–42. doi:10.1016/j.biopsycho.2013.10.011

Chen, J., & Gupta, A. (2012). *Parametric statistical change point analysis with applications to genetics, medicine and finance* (2nd ed.). New York, New York: Springer.

Christie, I., & Friedman, B. (2004). Autonomic specificity of discrete emotion and dimensions of affective space: A multivariate approach. *International Journal of Psychophysiology, 51*(2), 43–153. doi:10.1016/j.ijpsycho.2003.08.002

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Crowley, P., & Schultz, A. (2011). Measuring the intermittent synchronicity of macroeconomic growth in Europe. *International Journal of Bifurcation and Chaos, 21*(04), 1215–1231. doi:10.1142/S0218127411028957

Gnedenko, B. V. (2005). *The theory of probability*. Rhode Island: American Mathematical Society.

Grassmann, M., Vlemincx, E., von Leupoldt, A., & Van den Bergh, O. (in press). The role of respiratory measures to assess mental load in pilot selection. *Ergonomics*.

Hamilton, J. D. (1994). *Time series analysis*. Princeton, N.J.: Princeton University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Retrieved from http://statweb.stanford.edu/~tibs/ElemStatLearn/

Hoover, A., Singh, A., Fishel-Brown, S., & Muth, E. (2011). Real-time detection of workload changes using heart rate variability. *Biomedical Signal Processing and Control, 7*(4), 333–341. doi:10.1016/j.bspc.2011.07.004

Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics, 47,* 64–79. doi:10.1198/004017004000000563

Hubert, M., Rousseeuw, P. J., & Verdonck, T. (2009). Robust PCA for skewed data and its outlier map. *Computational Statistics and Data Analysis, 53,* 2264–2274. doi:10.1016/j.csda.2008.05.027

Jarusikova, D. (1997). Some problems with application of change point detection methods to environmental data. *Environmetrics, 8,* 469–483. doi:10.1002/(SICI)1099-095X(199709/10)8:5<469::AID-ENV265>3.0.CO;2-J

Jennings, J. R., Kamarck, T., Stewart, C., Eddy, M., & Johnson, P. (1992). Alternate cardiovascular baseline techniques: Vanilla or resting baseline. *Psychophysiology, 29,* 742–750. doi:10.1111/j.1469-8986.1992.tb02052.x

Kander, Z., & Zacks, S. (1966). Test procedures for possible changes in parameters of statistical distributions occurring at unknown time points. *The Annals of Mathematical Statistics, 37,* 1196–1210. doi:10.1214/aoms/1177699265

Klausmeier, H., & Sipple, T. (1982). Factor structure of the Piagetian stage of concrete operations. *Contemporary Educational Psychology, 7,* 161–180. doi:10.1016/0361-476X(82)90041-8

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing, 85*(4), 1501–1510. doi:10.1016/j.sigpro.2005.01.012

Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing, 85*(4), 717–736. doi:10.1016/j.sigpro.2004.11.012

Lindquist, M., Waugh, C., & Wager, T. (2007). Modeling state-related fMRI activity using change-point theory. *NeuroImage, 35,* 1125–1141. doi:10.1016/j.neuroimage.2007.01.004

Lung-Yut-Fong, A., Lévy-Leduc, C., & Cappé, O. (2012). Homogeneity and change-point detection tests for multivariate data using rank statistics. Retrieved from http://arxiv.org/abs/1107.1971

Matteson, D., & James, N. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association, 109*(505), 334–345. doi:10.1080/01621459.2013.849605

Mauss, I., Levenson, R., McCarter, L., Wilhelm, F., & Gross, J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion, 5,* 175–190. doi:10.1037/1528-3542.5.2.175

Müller, M., Baier, G., Galka, A., Stephani, U., & Muhle, H. (2005). Detection and characterization of changes of the correlation structure in multivariate time series. *Physical Review E, 71,* 046116. doi:10.1103/PhysRevE.71.046116

Page, E. (1954). Continuous inspection schemes. *Biometrika, 41*(1/2), 100–115. doi:10.1093/biomet/41.1-2.100

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human Development, 15,* 1–12. doi:10.1159/000271225

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846–850. doi:10.2307/2284239

Rosenfield, D., Zhou, E., Wilhelm, F., Conrad, A., Roth, W., & Meuret, A. (2010). Change point analysis for longitudinal physiological data: Detection of cardio-respiratory changes preceding panic attacks. *Biological Psychology, 84,* 112–120. doi:10.1016/j.biopsycho.2010.01.020

Shawe-Taylor, J., & Christianini, N. (2004). *Kernel methods for pattern analysis*. New York: Cambridge University Press.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Lanckriet, G., & Scholkopf, B., (2010). Hilbert Space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, *11*, 1517-1561. Retrieved from http://www.jmlr.org/papers/volume11/sriperumbudur10a/sriperumbudur10a.pdf

Szekely, G. J., & Rizzo, M. L. (2005). Hierarchical clustering via coint between-within distances: Extending Ward's minimum variance method. *Journal of Classification, 22,* 151–183. doi:10.1007/s00357-005-0012-9

Terien, J., Germain, G., Marque, C., & Karlsson, B. (2013). Bivariate piecewise stationary segmentation; improved pre-treatment for synchronization measures used on non-stationary biological signals. *Medical Engineering & Physics, 35*(8), 1188–1196. doi:10.1016/j.medengphy.2012.12.010

Terien, J., Marque, C., Germain, G., & Karlsson, B. (2009). Sources of bias in synchronization measures and how to minimize their effects on the estimation of synchronicity: Application to the uterine electromyogram. In G. R. Naik (Ed.), *Recent Advances in Biomedical Engineering* (pp. 73-99). InTech. doi:10.5772/7486

Van der Maas, H., & Molenaar, P. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review, 99*(3), 395–417. doi:10.1037/0033-295X.99.3.395

Veltman, J. A., & Gaillard, A. W. K. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics, 41,* 656–669. doi:10.1080/001401398186829

Wientjes, C. J. E., Grossman, P., & Gaillard, A. W. K. (1998). Influence of drive and timing mechanisms on breathing pattern and ventilation during mental task performance. *Biological Psychology, 49,* 53–70. doi:10.1016/S0301-0511(98)00026-X

Wright, P., & Wallace, J. (1988). Correlation structure of the El Niño/southern oscillation phenomenon. *American Meteorological Society, 1,* 609–625. doi:10.1175/1520-0442(1988)001<0609:CSOTEN>2.0.CO;2

Wu, S. D., & Lo, P. C. (2010). Cardiorespiratory phase synchronization during normal rest and inward-attention meditation. *International Journal of Cardiology, 141*(3), 325–328. doi:10.1016/j.ijcard.2008.11.137

Zhang, J., Yu, X., & Xie, D. (2010). Effects of mental tasks on the cardiorespiratory synchronization. *Respiratory Physiology & Neurobiology, 170*(1), 91–95. doi:10.1016/j.resp.2009.11.003