Consistent and powerful non-Euclidean graph-based change-point test with applications to segmenting random interfered video data

Xiaoping Shi[a,1], Yuehua Wu[b,1], and Calyampudi Radhakrishna Rao[c,d,1]

[a]Department of Mathematics and Statistics, Thompson Rivers University, Kamloops, BC, Canada V2C0C8; [b]Department of Mathematics and Statistics, York University, Toronto, ON, Canada M3J1P3; [c]Department of Biostatistics, University at Buffalo, The State University of New York, Buffalo, NY 14221-3000; and [d]CR RAO Advanced Institute of Mathematics, Statistics and Computer Science, Hyderabad 500046, India

Contributed by Calyampudi Radhakrishna Rao, April 13, 2018 (sent for review March 19, 2018; reviewed by Miklós Csörgő and Runze Li)

The change-point detection has been carried out in terms of the Euclidean minimum spanning tree (MST) and shortest Hamiltonian path (SHP), with successful applications in the determination of authorship of a classic novel, the detection of change in a network over time, the detection of cell divisions, etc. However, these Euclidean graph-based tests may fail if a dataset contains random interferences. To solve this problem, we present a powerful non-Euclidean SHP-based test, which is consistent and distribution-free. The simulation shows that the test is more powerful than both Euclidean MST- and SHP-based tests and the non-Euclidean MST-based test. Its applicability in detecting both landing and departure times in video data of bees' flower visits is illustrated.

non-Euclidean distance | shortest Hamilton path | minimum spanning tree | change-point | distribution-free

Webcams are used in both daily life and research projects. Such examples include visitors' motion detection via a home security system, video fire detection via a fire detection system, study of the behavior of a rare and big but dangerous animal, magnetic resonance imaging analysis of some functions (e.g., learning function) of a part of a brain, and investigation of bees' flower visitation. A webcam collects tons of short video clips. However, many of video clips do not contain any information of interest, and hence, they can be and should be removed for efficiency, effectiveness, and economy of the variety of purposes. For example, to investigate the movement pattern of bees, one may use webcams to record a bee's flower visitation to find the duration of the bee on the flower (1). Huge video data need to be analyzed to accurately investigate bees' flower visitation. However, in all of the above examples, there may exist unexpected (i.e., random) interferences. For instance, in the bees' flower visitation example, relatively smaller insects, such as ants, may also visit the flower unexpectedly; these are the random interferences and are not avoidable. Fig. 1 displays four selected frames from those extracted every second from a recorded video. It can be seen that the flower was visited by both bees and ants, and it can also be seen that they were landed in different places of the flower. Thus, we are interested in keeping the video data only containing bees' flower visitation.

The removal of the informationless video data in the examples given above can actually be converted into a change-point detection problem for high-dimensional data that contain random interferences. Consider the bees' flower visitation example for demonstration. Here, the sequence of data consists of a sequence of vectorized pixel values from frames, and both landing and leaving times of bees make large changes in vectorized pixel values and can thus be considered as change points in the data sequence. A well-performed change-point detection method would allow users to remove large but informationless video data in such examples, which can be carried out daily or other regular basis.

Many change-point detection methods can be found in literature. Refs. 2 and 3, among others, gave change-point analyses for high-dimensional time series, where the data structure changes in a fixed subset of components. Graph-based change-point tests have been developed recently for their advantage in describing high-dimensional data, which date back to Friedman and Rafsky (4) for a two-sample test via applying the minimum spanning tree (MST). In terms of Friedman and Rafsky (4), Chen and Zhang (5) proposed a change-point test. However, the test based on MST is not distribution-free and is not consistent when there is a shift in variance in high-dimensional settings as shown in ref. 6. By applying the shortest Hamiltonian path (SHP) introduced in ref. 7 for a two-sample test, ref. 6 proposed a distribution-free and consistent change-point test. We remark that both of the above change-point tests are constructed in terms of Euclidean graph, which may not perform for the problem considered in this paper. Consider the bees' flower visitation example. The issue arises because (*i*) the random interferences by small insects may lead to the changes of vectorized pixel values, which could not be ignored, and (*ii*) the pixel values change at a relatively large number of unknown image points for a bee but a small number of unknown image points for a small insect, such as an ant. Both locations and amounts of changes in pixel values are unknown, which make the Euclidean graph-based change-point tests not to perform. The issue may be resolved by replacing the Euclidean distance by a non-Euclidean distance.

Significance

A webcam collects tons of short video clips for home security, fire inspection, and animal's behavior study, etc. For efficiency, effectiveness, and economy of the variety of purposes of the use of the video data, we need to remove informationless data. To solve this problem, we propose two non-Euclidean graph-based change-point tests: SHP* and MST*. The proposed test SHP* is not only distribution-free but also, consistent. As shown in the real data example of a bee's flower visitation, it has successfully detected both landing and departure times of a bee from the video data with random interference.

Author contributions: X.S., Y.W., and C.R.R. designed research; X.S., Y.W., and C.R.R. performed research; X.S. analyzed data; and X.S., Y.W., and C.R.R. wrote the paper.

Reviewers: M.C., Carleton University; and R.L., Pennsylvania State University.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. Email: xshi@tru.ca, wuyh@mathstat.yorku.ca, or crr1@psu.edu.

Published online May 21, 2018.

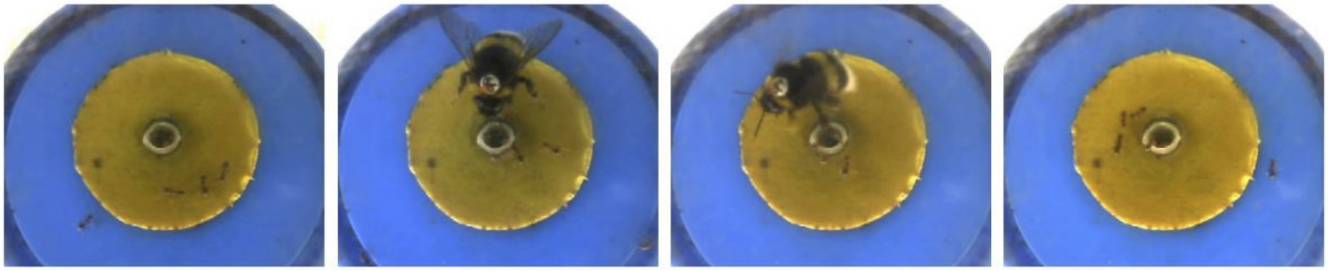

Fig. 1. Extracted frames with dimension 288 × 352 located at 1, 5, 40, and 49 from the original video 09-10-2010_15h_49_27.17.mpg (faculty.tru.ca/xshi/09-10-2010_15h_49_27.17.mpg). The landing and departure times of the bee are 5 and 41, respectively.

Non-Euclidean Graph-Based Change-Point Test

We first vectorized the matrix of pixel values of the tth image into a d-dimensional vector, with d being the number of pixels in this image. To show how to remove large but information-less video data in any example described above, we use the bees' flower visitation example for demonstration. We assume that, at most, a single bee has visited the flower for simple presentation. The possible scenarios include (*i*) there is no insect on the flower; (*ii*) there are only a few small insects, such as ants, on the flower; and (*iii*) there is a bee on the flower. It is noted that both small insects and a bee may land on any place of the flower. The following model can be used to model such video data, where the three parts of this model correspond to the above three scenarios, respectively:

$$y_{t,j}=\begin{cases}\mu_j+\sigma e_{t,j}, & \text{if } j\notin\mathcal{J}_t \text{ and } j\notin\mathcal{K}_t,\\ a_1+b_1\varepsilon_{t,j}, & \text{if } j\in\mathcal{J}_t \text{ and } j\notin\mathcal{K}_t,\\ a_2+b_2\eta_{t,j}, & \text{if } j\in\mathcal{K}_t,\end{cases} \quad [1]$$

where $y_{t,j}$ is observation at time $t\in\{1,\dots,N\}$ and location of component $j\in\{1,\dots,d\}$; μ_j, σ, and a_ℓ b_ℓ for $\ell=1,2$ are unknown parameters; and $e_{t,j}$, $\varepsilon_{t,j}$, and $\eta_{t,j}$ are independently and identically distributed errors with mean zero and variance one. The first part is to model the distribution of pixel values when there are no insects or bees, the second part is for those pixel values of the insects, and the third part is for those pixel values of the bee. We let $\mathcal{J}_t$ for $1\le t\le N$ and $\mathcal{K}_t$ for $t^*<t\le N$ ($\mathcal{K}_t=\emptyset$ for $1\le t\le t^*$) be independently and identically distributed samples with sizes $|\mathcal{J}_t|$ and $|\mathcal{K}_t|$, respectively, from $\{1,\dots,d\}$ without replacement to reflect the random movement of insects and bee. If there exists a $t^*<N$, such that $\mathcal{K}_{t^*}\neq\emptyset$, the time t^* is called a change point. For simple presentation, we let $|\mathcal{J}_t|\equiv|\mathcal{J}|$ for $1\le t\le N$ and $|\mathcal{K}_t|\equiv|\mathcal{K}|$ for $t^*<t\le N$.

To detect whether there is a change point t^*, it is equivalent to test the following hypotheses:

$$H_0:\mathcal{K}_t\equiv\emptyset \text{ for } 1\le t\le N \text{ vs. } H_a:\begin{cases}\mathcal{K}_t\equiv\emptyset \text{ for } 1\le t\le t^*,\\ \mathcal{K}_t\neq\emptyset \text{ for } t^*<t\le N.\end{cases}$$

Let G be a connected, edge-weighted undirected graph made up of a set of nodes $V(G)=\{1,\dots,N\}$ together with a set of edges $\mathcal{E}(G)$. The Euclidean edge weight between two nodes t_1 and t_2 is given by

$$w_{t_1,t_2}=\sqrt{\sum_{j=1}^{d}(y_{t_1,j}-y_{t_2,j})^2}. \quad [2]$$

However, the Euclidean graph-based tests have low power in detecting a change point by simulation studies. Thus, we replace the Euclidean edge weight by the following non-Euclidean edge weight:

$$w^*_{t_1,t_2}=\left|\sum_{j=1}^{d}y_{t_1,j}-\sum_{j=1}^{d}y_{t_2,j}\right|. \quad [3]$$

For example, a graph G can be an MST or an SHP based on the Euclidean edge weight Eq. **2** or the non-Euclidean edge weight Eq. **3** denoted MST(w), MST(w^*) or SHP(w), SHP(w^*), where the sum of edge weights attains the minimum among all of the spanning trees or paths, respectively. To test the null hypothesis, we cut the whole set of nodes $\{1,\dots,N\}$ at an arbitrary point t into two sets $\{1,\dots,t\}$ (until t) and $\{t+1,\dots,N\}$ (after t). As in ref. 5, for the Euclidean edge weight Eq. **2**, we define

$$C_t^{\mathrm{G}(w)}=\sum_{(t_1,t_2)\in\mathcal{E}(G(w))}I\{I(t_1>t)\neq I(t_2>t)\}. \quad [4]$$

Ref. 5 proposed an MST-based test, denoted **MST** in this paper, with the test statistic

$$\widetilde{C}_N^{\mathrm{MST}(w)}=\max_{n_0\le t\le n_1}Z_t^{\mathrm{MST}(w)}, \quad [5]$$

where $Z_t^{\mathrm{MST}(w)}=-\frac{C_t^{\mathrm{MST}(w)}-E_0\left(C_t^{\mathrm{MST}(w)}\right)}{\sqrt{\mathrm{var}_0\left(C_t^{\mathrm{MST}(w)}\right)}}$, with G($w$) being replaced by MST($w$) in Eq. **4**; n_0 and n_1 are prespecified constraints, and $E_0(C_t^{\mathrm{MST}(w)})$ and $\mathrm{var}_0(C_t^{\mathrm{MST}(w)})$ are expectation and variance, respectively, of $C_t^{\mathrm{MST}(w)}$ under the permutation null. Note that this test is nonparametric but not distribution-free. If the null hypothesis is rejected, the change-point estimate is given by

$$\arg\max_{n_0\le t\le n_1}Z_t^{\mathrm{MST}(w)}. \quad [6]$$

For the same Euclidean edge weight Eq. **2**, ref. 6 proposed an SHP-based test, denoted **SHP** in this paper, with the test statistic

$$S_N^{\mathrm{SHP}(w)}=\frac{1}{N-1}\sum_{t=1}^{N-1}\left(Z_t^{\mathrm{SHP}(w)}\right)^2, \quad [7]$$

where $Z_t^{\mathrm{SHP}(w)}=\frac{C_t^{\mathrm{SHP}(w)}-E_0\left(C_t^{\mathrm{SHP}(w)}\right)}{\sqrt{\mathrm{var}_0\left(C_t^{\mathrm{SHP}(w)}\right)}}$, with G($w$) being replaced by SHP($w$) in Eq. **4**, $E_0(C_t^{\mathrm{SHP}(w)})=2t(N-t)/N$, and $\mathrm{var}_0(C_t^{\mathrm{SHP}(w)})=2t(N-t)\{2t(N-t)-N\}/(N^3-N^2)$ (8). If the null hypothesis is rejected, the change-point estimate is given by

$$\arg\min_{1\le t<N}C_t^{\mathrm{SHP}(w)}/\{t(N-t)\}. \quad [8]$$

We now define both non-Euclidean SHP- and MST-based test statistics by replacing the Euclidean edge weights Eq. **2** with the non-Euclidean edge weights Eq. **3** in both test statistics Eq. **7** and Eq. **5**. We denote so-defined test statistics $S_N^{\mathrm{SHP}(w^*)}$ and $\widetilde{C}_N^{\mathrm{MST}(w^*)}$, respectively. These lead to the non-Euclidean SHP- and MST-based change-point tests, which are denoted **SHP*** and **MST***, respectively, in this paper. If the null hypothesis is rejected by **SHP*** or **MST***, the change-point estimate is given by the ratio cut as in Eq. **8** or by Eq. **6**, respectively. **SHP*** and **MST*** will also denote the non-Euclidean SHP-based change-point detection and the non-Euclidean MST-based change-point detection, respectively, for convenience.

To illustrate the different edge weights, we consider a mean shift model by setting $N=10$, $d=3$, $u_1=u_2=0$, $u_3=-1$, $a_1=0.1$, $a_2=2$, and $\sigma=b_1=b_2=1$; $e_{t,j}$, $\varepsilon_{t,j}$, and $\eta_{t,j}$ are standard normal errors. $|\mathcal{J}|=1$ and $|\mathcal{K}|=2$ in Eq. **1**, with a change point at five. Note that the Euclidean MST-based change-point detection was implemented in the R package gSeg (9) with default settings $n_0=0.05N$ and $n_1=0.95N$, while the Euclidean SHP-based change-point detection can be carried out by using the msTreeKruskal function in the R package optrees (10).

An illustration is given in Fig. 2, from which it can be seen that the change-point estimate by **SHP*** is the same as the true one, but the other three tests produce biased change-point estimates that are eight, six, and eight, respectively.

It seems that the model Eq. **1** would only be suitable for studying a bee's landing. In fact, it is also the right model for studying a bee's departure, because we may set $\mathcal{K}_t\neq\emptyset$ for $1\le t\le t^*$ and $\mathcal{K}_t\equiv\emptyset$ otherwise. Thus, **SHP*** and **MST*** remain unchanged.

Consistency of the Non-Euclidean SHP-Based Test

Consider **SHP*** given above. We will show that it is consistent with fixed N and $d\to\infty$.

Assumption 1. *Suppose that* $a_1\neq a_2$, $|\mathcal{J}|\ll|\mathcal{K}|$, $\sqrt{d}\ll|\mathcal{K}|$, *and* $\max(|\mathcal{L}_1|,|\mathcal{L}_2|)\ll|\mathcal{K}|$ *as* $d\to\infty$, *where* $\mathcal{L}_1=\{j:\mu_j-a_2=0\}$ *and* $\mathcal{L}_2=\{(i,j):\mu_i\neq\mu_j \text{ for } i\neq j\}$.

Assumption 2. *There exists an* N_α, *such that*

$$\min\Big\{\sum_{|t-t^*|\le N_\alpha}\frac{1}{N-1}\sum_{t=1}^{N-1}\frac{\{\kappa_t-E_0(C_t^{\mathrm{SHP}})\}^2}{\mathrm{var}_0(C_t^{\mathrm{SHP}})}:\kappa_{t^*}\le 2,\ |\kappa_t-\kappa_{t\pm1}|\le 2\Big\}>c_\alpha.$$

We remark that *Assumption 1* stems from the bee's flower visitation example, in which the following are met: (*i*) a bee is quite large compared with other small insects, such as ants ($|\mathcal{J}|\ll|\mathcal{K}|$ and $a_1\neq a_2$); (*ii*) a bee is not much smaller than the flower ($\sqrt{d}\ll|\mathcal{K}|$); (*iii*) the color of a bee is different from that of the flower ($|\mathcal{L}_1|\ll|\mathcal{K}|$); and (*iv*) the whole flower has almost the same color ($|\mathcal{L}_2|\ll|\mathcal{K}|$). *Assumption 2* comes from ref. 6.

The following theorem shows that **SHP*** is consistent.

Theorem. *Under* Assumptions 1 *and* 2, *for a predefined positive number* α, *the power of the non-Euclidean SHP-based test of significance level* α *converges to one as* $d\to\infty$.

Proof: By Eq. **3**,

$$w^*_{t_1,t_2}=\Big|\sum_{j=1}^{d}\{y_{t_1,j}-E(y_{t_1,j})\}-\sum_{j=1}^{d}\{y_{t_2,j}-E(y_{t_2,j})\}+\sum_{j=1}^{d}\{E(y_{t_1,j})-E(y_{t_2,j})\}\Big|.$$

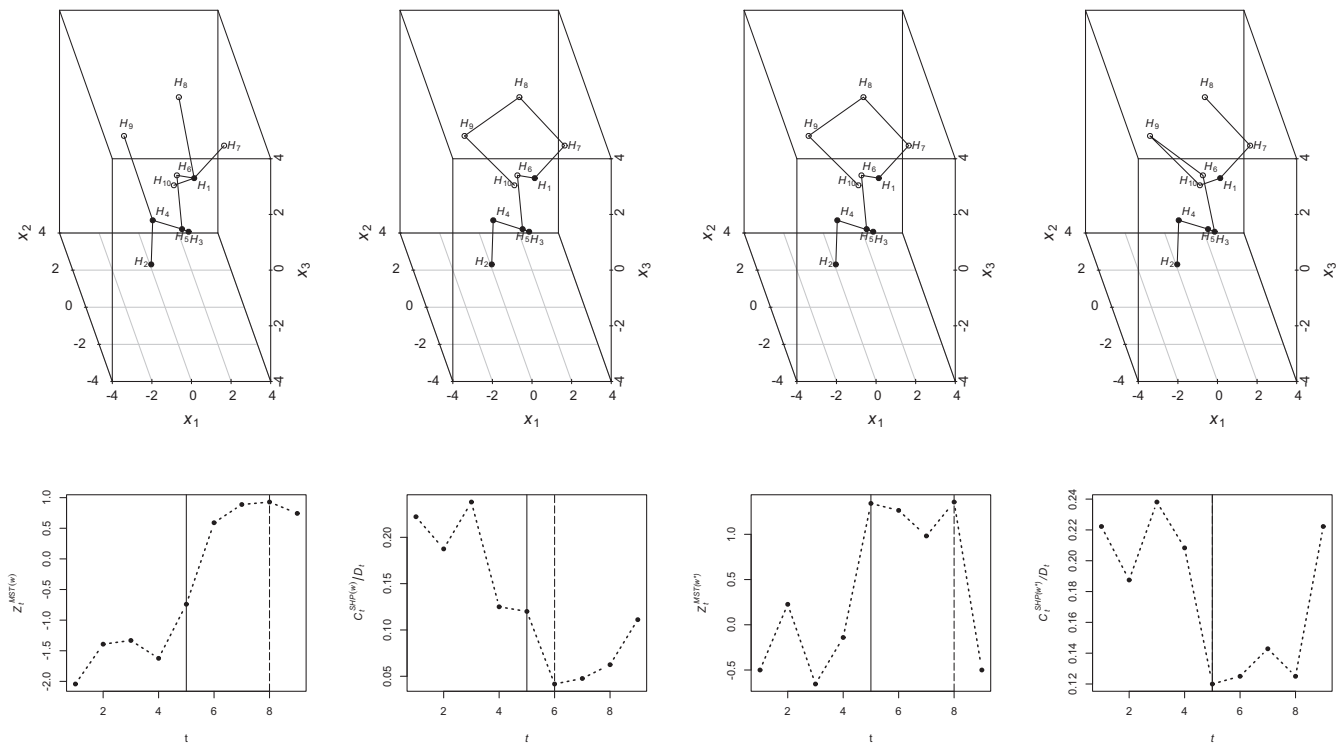

Fig. 2. An illustration of SHP and MST based on Euclidean or non-Euclidean edge weight and the corresponding change-point estimates for $N=10$ and $d=3$. *Upper* depicts the MST in a complete Euclidean graph (column 1), the SHP in a complete Euclidean graph (column 2), the MST in a complete non-Euclidean graph (column 3), and the SHP in a complete non-Euclidean graph (column 4). *Lower* displays the corresponding change-point estimates for the four graphs given in *Upper*.

Table 1. Simulated type I errors for SHP*

d	10	50	100	500	1,000	5,000
$N=20$	0.052	0.042	0.059	0.060	0.057	0.047
$N=40$	0.039	0.052	0.039	0.042	0.054	0.055
$N=60$	0.045	0.045	0.044	0.061	0.051	0.044
$N=80$	0.048	0.047	0.032	0.038	0.047	0.048
$N=100$	0.045	0.046	0.055	0.046	0.039	0.037
$N=200$	0.059	0.056	0.054	0.051	0.058	0.053
$N=300$	0.058	0.047	0.050	0.048	0.050	0.054

Note that $\sum_{j=1}^{d}\{y_{t,j}-E(y_{t,j})\}=O_p(\sqrt{d})$ for $1\le t\le N$. By Eq. **1**, $\left|\sum_{j=1}^{d}\{E(y_{t_1,j})-E(y_{t_2,j})\}\right|/|\mathcal{K}|\to 0$ if $t_1,t_2\le t^*$ or $t_1,t_2>t^*$; otherwise, $\left|\sum_{j=1}^{d}\{E(y_{t_1,j})-E(y_{t_2,j})\right|/|\mathcal{K}|\nrightarrow 0$ by *Assumption 1*. Thus, we have

$$\lim_{d\to\infty} w^*_{t_1,t_2}/|\mathcal{K}|\begin{cases}=0, & \text{if } t_1,t_2\le t^* \text{ or } t_1,t_2>t^*\\ \neq 0, & \text{otherwise}\end{cases} \quad [9]$$

as $d\to\infty$. The rest follows the proof of theorem 1 in ref. 6.

We remark that the conditions in Eq. **1** may be relaxed by letting $a_1=a_{1,j}$ and $a_2=a_{2,j}$ (i.e., they may depend on j). With some additional conditions to *Assumption 1*, the test can still be shown to be consistent.

It is noted that the change-point estimate given by using the ratio cut as in ref. 6 has the same error bound as given in theorem 2 in ref. 6.

Data Examples

Simulations. For simple presentation, we only carry out the simulation studies for the model Eq. **1**, with $e_{t,j}$, $\varepsilon_{t,j}$, and $\eta_{t,j}$ being standard normal errors.

In Table 1, the simulated type I errors for **SHP*** are compared based on 1,000 simulations, with $N=40,60,80,100,200,300$; $\alpha=0.05$; $u_j\equiv 0$; $a_1=0.1$; $\sigma=b_1=1$; $|\mathcal{J}|=\lfloor\log d\rfloor$; $\mathcal{K}_t\equiv\emptyset$; and $d=10,50,100,500,1{,}000,5{,}000$ in Eq. **1**, where the critical values are taken from table 1 of ref. 6, and $\lfloor c\rfloor$ denotes the greatest integer less than or equal to a real number c. It can be seen from Table 1 that the test performs well.

To examine the powers of **MST**, **SHP**, **MST***, and **SHP***, 200 simulations are carried out for $a_2=0.3$, $b_2=1$, and $\mathcal{K}_t=\emptyset$ for $t\le t^*$ and $|\mathcal{K}|=\lfloor d^{0.7}\rfloor$. To investigate the effect of a change-point location for different N, the locations $t^*=N/2$, $N/4$ are considered with $N=40,100$. We set $\alpha=0.05$. Fig. 3 displays the percentages of the rejections of the null hypothesis at 0.05 significance level by each of **MST**, **SHP**, **MST***, and **SHP*** in the simulation study.

It can be seen from Fig. 3*A* that the powers of **MST*** and **SHP*** monotonically increase as d increases, which suggests that both of them may be consistent. Compared with the powers of both **MST** and **SHP**, both MST* and SHP* have much better performances, especially when the change point is located at $N/2$. Fig. 3*A* also reveals that both **MST** and **SHP** may not be consistent.

Further comparisons are carried out for each of the four change-point estimates based on $\arg\max_{n_0\le t\le n_1} Z_t^{\mathrm{MST}(w)}$ for **MST**, $\arg\max_{n_0\le t\le n_1} Z_t^{\mathrm{MST}(w^*)}$ for **MST***, $\arg\min_{1\le t<N} C_t^{\mathrm{SHP}(w)}/\{t(N-t)\}$ for **SHP**, and $\arg\min_{1\le t<N} C_t^{\mathrm{SHP}(w^*)}/\{t(N-t)\}$ for **SHP***. Fig. 3*B* displays the boxplots of these estimates for each of the dimensions $d_1=10$, $d_2=50$, $d_3=100$, $d_4=500$, $d_5=1{,}000$, and $d_6=5{,}000$.

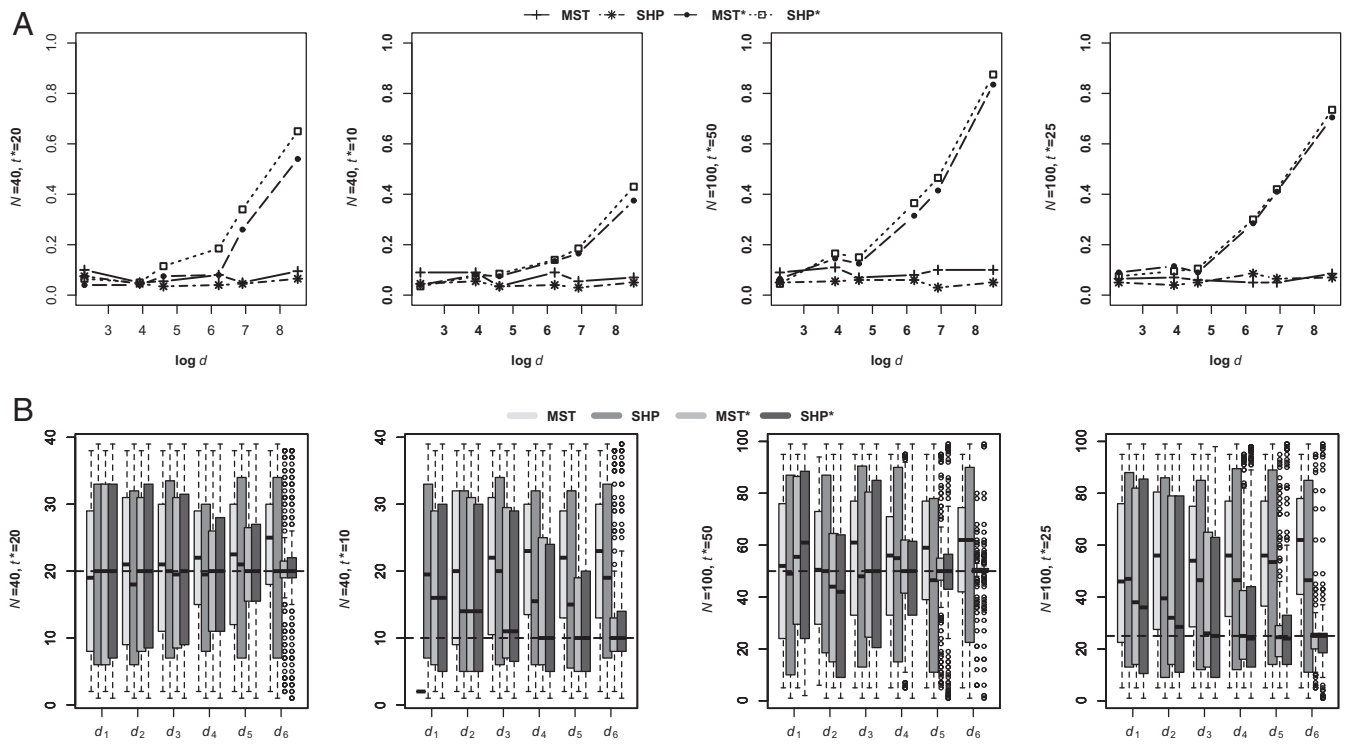

Fig. 3. (*A*) Powers of **MST**, **SHP**, **MST***, and **SHP*** in 200 simulations with $|\mathcal{J}|=\lfloor\log d\rfloor$, $|\mathcal{K}|=\lfloor d^{0.7}\rfloor$, sample size of $N=40$ or $N=100$, and the change point located at $t^*=N/2$ or $t^*=N/4$. (*B*) Boxplots of the change-point estimates by $\arg\max_{n_0\le t\le n_1} Z_t^{\mathrm{MST}(w)}$ for **MST**, $\arg\min_{1\le t<N} C_t^{\mathrm{SHP}(w)}/\{t(N-t)\}$ for **SHP**, $\arg\max_{n_0\le t\le n_1} Z_t^{\mathrm{MST}(w^*)}$ for **MST***, and $\arg\min_{1\le t<N} C_t^{\mathrm{SHP}(w^*)}/\{t(N-t)\}$ for **SHP*** for the dimensions of $d_1=10$, $d_2=50$, $d_3=100$, $d_4=5{,}00$, $d_5=1{,}000$, and $d_6=5{,}000$; $|\mathcal{J}|=\lfloor\log d\rfloor$; $|\mathcal{K}|=\lfloor d^{0.7}\rfloor$; the respective change point located at $t^*=N/2$ or $N/4$; and two sample sizes of $N=40$ or 100.

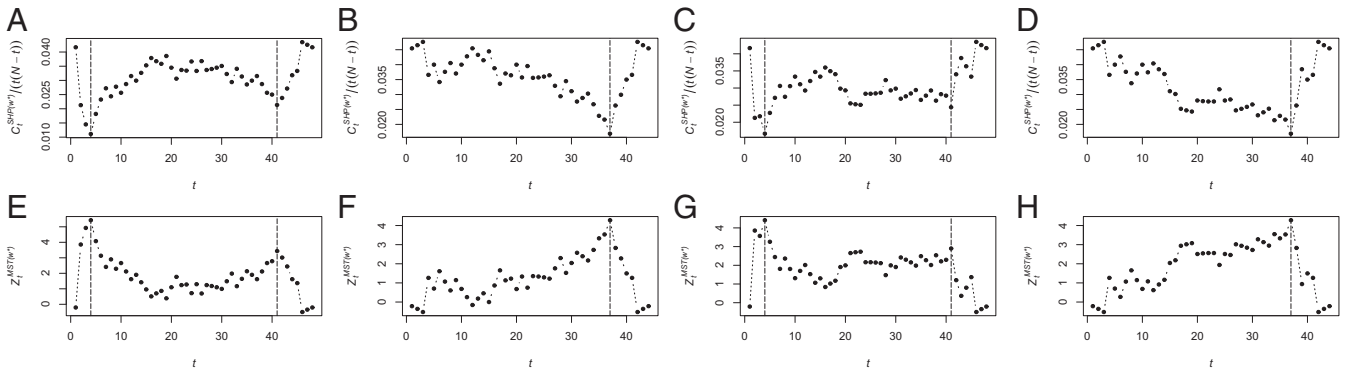

Fig. 4. $C_t^{\mathrm{SHP}(w^*)}/(t(N-t))$ based on $x_{t,j}$ for $t=1,\ldots,49$ (*A*) and $t=5,\ldots,49$ (*B*). $C_t^{\mathrm{SHP}(w^*)}/(t(N-t))$ based on $\tilde{x}_{t,j}$ for $t=1,\ldots,49$ (*C*) and $t=5,\ldots,49$ (*D*). $C_t^{\mathrm{SHP}(w^*)}/(t(N-t))$ based on $\tilde{x}_{t,j}$ for $t=1,\ldots,49$ (*E*) and $t=5,\ldots,49$ (*F*). $C_t^{\mathrm{SHP}(w^*)}/(t(N-t))$ based on $\tilde{x}_{t,j}$ for $t=1,\ldots,49$ (*G*) and $t=5,\ldots,49$ (*H*).

It can be seen from Fig. 3*B* that both change-point estimates given by non-Euclidean graph-based methods, which are comparable, outperform both change-point estimates by Euclidean graph-based methods when d increases. It is noted that the power and change-point estimate based on **MST** and **MST*** are dependent on the constraints of $n_0 \geq 1$ and $n_1 \leq N-1$. Their performances with $n_0 = 1$ and $n_1 = N-1$ (i.e., no constraints) are not as good as those with $n_0 = 0.05N$ and $n_1 = 0.95N$ in the default settings, which are used to produce Fig. 3. The detailed simulation studies are omitted.

Video Data. We still use the bees' flower visitation example for demonstration. As we described before, relatively smaller insects, such as ants, may also visit the flowers unexpectedly. We analyze the original video 09-10-2010_15h_49_27.17.mpg (faculty.tru.ca/xshi/09-10-2010_15h_49_27.17.mpg); some of 49 frames extracted from the video data are shown in Fig. 1 with one frame per second. The details of the experiment that produced the data are given in the work by Lihoreau et al. (1). As shown in Fig. 1, there were only some ants in the first and fourth frames. Our aim is to locate the frame where a bee appears.

To proceed, we first convert the three intensities of the R, G, and B components (three-dimensional array) to one intensity of the grayscale (one-dimensional array). As reviewed by Kanan and Cottrell (11), there are two popular methods for the conversion. The simpler one is to convert three intensities by their average as follows (12):

$$x_{t,j} = (x_{1,t,j} + x_{2,t,j} + x_{3,t,j})/3, \qquad \textbf{[10]}$$

where $x_{1,t,j}$, $x_{2,t,j}$, and $x_{3,t,j}$ represent the three-dimensional pixel values of the R, G, and B components, respectively. Another one proposed by Pratt (13) is to match human brightness perception by using a weighted average:

$$\tilde{x}_{t,j} = 0.3x_{1,t,j} + 0.59x_{2,t,j} + 0.11x_{3,t,j}. \qquad \textbf{[11]}$$

The quality of the videos was variable depending on climatic conditions, such as light. To have the same contrast, we make the same-scale transformations on the pixel values $x_{t,j}$ from Eq. **10** or $\tilde{x}_{t,j}$ from Eq. **11** of 49 frames:

$$y_{t,j} = (x_{t,j} - \min_j x_{t,j})/(\max_j x_{t,j} - \min_j x_{t,j}),$$
$$\tilde{y}_{t,j} = (\tilde{x}_{t,j} - \min_j \tilde{x}_{t,j})/(\max_j \tilde{x}_{t,j} - \min_j \tilde{x}_{t,j}),$$

and thus, $d = 288 \times 352 = 101,376$. We construct the non-Euclidean SHP with weight defined in Eq. **3** using $y_{t,j}$. The test statistic $S_{49}^{\mathrm{SHP}(w^*)} = 4.7$, which suggests that there is a change point at 4 for significance level 0.05. Fig. 4*A* displays $C_t^{\mathrm{SHP}(w^*)}/(t(N-t))$, where the change-point estimate is four. Let us consider the segment for $\{y_{t,j}, t=5,\ldots,49\}$. In view of the fact that the test statistic $S_{45}^{\mathrm{SHP}(w^*)} = 2.8$, there exists a change-point estimate at 37 for significance level 0.05, which is displayed in Fig. 4*B*. Thus, there are two change-point estimates at 4 and 41 (=37 + 4), which in fact, are located at the local minimums in Fig. 4*A*. As a matter of fact, 4 and 41 are the true change points corresponding to the landing and departure times, respectively, of a bee.

If we replace $y_{t,j}$ with $\tilde{y}_{t,j}$ above, the same change points are detected by applying **SHP***, which are displayed in Fig. 4 *C* and *D*. This suggests that the impact of different weighted algorithms for converting RGB to grayscale may be negligible.

The application of **MST*** yields the same change-point estimates that are detected by **SHP***, which are displayed in Fig. 4 *E*–*H*. However, if we apply **MST** or **SHP**, their performances are not satisfactory, as shown in Fig. 5 *A* and *B* for $y_{t,j}$ and Fig. 5 *C* and *D* for $\tilde{y}_{t,j}$.

Discussion and Conclusions

Using the bees' flower visitation example as a demonstration, two non-Euclidean graph-based change-point tests are developed for high-dimensional data with random interferences. The proposed non-Euclidean SHP-based change-point test **SHP*** is not only distribution-free but also, consistent. For analyzing the video data, we first extract it to a sequence of frames and make the

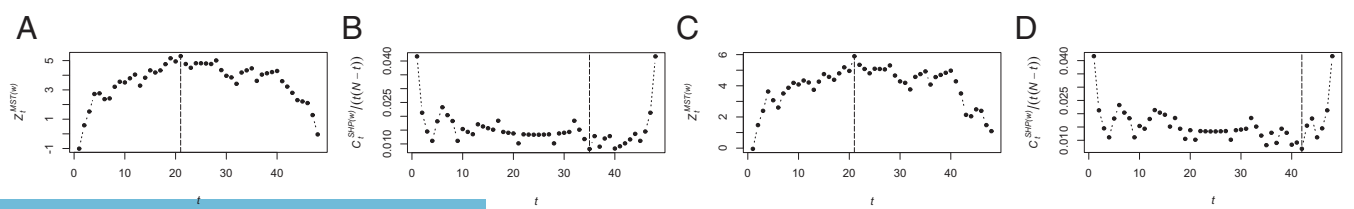

Fig. 5. (*A*) $Z_t^{\mathrm{MST}(w)}$ based on $\{x_{t,j}\}$. (*B*) $C_t^{\mathrm{SHP}(w)}/(t(N-t))$ based on $\{x_{t,j}\}$. (*C*) $Z_t^{\mathrm{MST}(w)}$ based on $\{\tilde{x}_{t,j}\}$. (*D*) $C_t^{\mathrm{SHP}(w)}/(t(N-t))$ based on $\{\tilde{x}_{t,j}\}$.

same-scale transformation of the data; then, we apply **SHP*** to detect the bee's landing and departure times. It is noted that the performance of the resulting change-point estimate depends on the number of frames per second. We have limited the discussion to the case that there is, at most, one bee on the flower. If we remove this assumption and allow multiple bees to visit a flower, the following are possible cases: (*i*) they show up at almost same time, or (*ii*) they visit the flower at different times. For both cases *i* and *ii*, **SHP*** can be used to find out the landing of the first bee and departure of the last bee. We remark that we may also use **MST***; however, its performance is heavily dependent on the constraints n_0 and n_1.

The model Eq. **1** can be modified to adapt to other video data examples containing random interferences, which can be used to remove the informationless data. The change-point detection method can be given similarly as above. The details are omitted.

ACKNOWLEDGMENTS. We thank Dr. Mathieu Lihoreau for data sharing. The research is partially supported by the Natural Sciences and Engineering Research Council of Canada.

1. Lihoreau M, Chittka L, Raine NE (2016) Monitoring flower visitation networks and interactions between pairs of bumble bees in a large outdoor flight cage. *PLoS One* 11:e0150844.
2. Cho H, Fryzlewicz P (2015) Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J R Stat Soc Ser B Stat Methodol* 77:475–507.
3. Wang T, Samworth RJ (2017) High dimensional change point estimation via sparse projection. *J R Stat Soc Ser B Stat Methodol* 80:57–83.
4. Friedman JH, Rafsky LC (1979) Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. *Ann Statist* 7:697–717.
5. Chen H, Zhang N (2015) Graph-based change-point detection. *Ann Statist* 43:139–176.
6. Shi X, Wu Y, Rao CR (2017) Consistent and powerful graph-based change-point test for high-dimensional data. *Proc Natl Acad Sci USA* 114:3873–3878.
7. Biswas M, Mukhopadhyay M, Ghosh AK (2014) A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika* 101:913–926.
8. Wald A, Wolfowitz J (1940) On a test whether two samples are from the same distribution. *Ann Math Statist* 11:147–162.
9. Chen H, Zhang N (2014) gSeg: Graph-based change-point detection (g-Segmentation). R package, version 0.1. Available at https://www.rdocumentation.org/packages/gSeg/versions/0.1. Accessed October 27, 2015.
10. Fontenla M (2014) optrees: Optimal trees in weighted graphs, version 1.0. Available at https://cran.r-project.org/web/packages/optrees/index.html. Accessed October 12, 2016.
11. Kanan C, Cottrell GW (2012) Color-to-grayscale: Does the method matter in image recognition? *PLoS ONE* 7:e29740.
12. Jack K (2007) *Video Demystified: A Handbook for the Digital Engineer* (HighText, Solana Beach, CA), 5th Ed.
13. Pratt W (2007) *Digital Image Processing* (John Wiley & Sons, Hoboken, NJ).