Algorithms for detecting changepoints in data

Cody Buntain, Christopher Natoli, Miroslav Živković

OSDC PIRE 2014, hosted at University of Amsterdam

24 July 2014

Building the test statistic

$$\text{test statistic} = \frac{\sum_{t=1}^{h} \boldsymbol{e}_t}{}$$

Galeano, Pedro, and Daniel Peña. "Covariance changes detection in multivariate time series."

Journal of Statistical Planning and Inference 137.1 (2007): 194-211.

Building the test statistic

$$\text{test statistic} = \frac{\sum_{t=1}^{h} \boldsymbol{e}_t{}^{\top} \qquad \boldsymbol{e}_t}{}$$

Galeano, Pedro, and Daniel Peña. "Covariance changes detection in multivariate time series." *Journal of Statistical Planning and Inference* 137.1 (2007): 194-211.

Building the test statistic

$$\text{test statistic} = \frac{\sum_{t=1}^{h} \boldsymbol{e}_t{}^\top (\hat{\Sigma})^{-1} \boldsymbol{e}_t}{h}$$

$$\hat{\Sigma} := \frac{1}{n-1} \sum_{t=1}^{n} \boldsymbol{e}_t \boldsymbol{e}_t^\top$$

Galeano, Pedro, and Daniel Peña. “Covariance changes detection in multivariate time series.” *Journal of Statistical Planning and Inference* 137.1 (2007): 194-211.

Building the test statistic

$$\text{test statistic} = \frac{\sum_{t=1}^{h} \boldsymbol{e}_t{}^{\top}(\hat{\Sigma})^{-1}\boldsymbol{e}_t}{h} - \frac{\sum_{t=1}^{n} \boldsymbol{e}_t^{\top}(\hat{\Sigma})^{-1}\boldsymbol{e}_t}{n}$$

Galeano, Pedro, and Daniel Peña. "Covariance changes detection in multivariate time series."

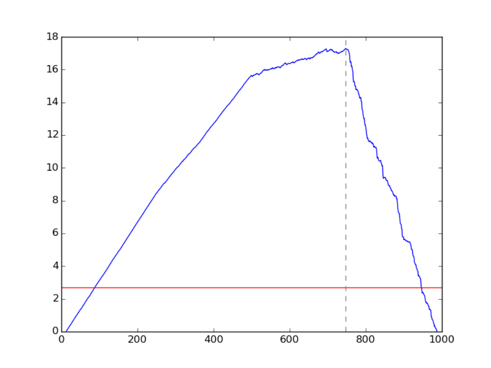
Journal of Statistical Planning and Inference 137.1 (2007): 194-211.

Building the test statistic

$$\text{test statistic} = \frac{h}{\sqrt{2kn}}\left(\frac{\sum_{t=1}^{h} \boldsymbol{e}_t{}^\top (\hat{\Sigma})^{-1}\boldsymbol{e}_t}{h} - \frac{\sum_{t=1}^{n} \boldsymbol{e}_t^\top (\hat{\Sigma})^{-1}\boldsymbol{e}_t}{n}\right)$$

Galeano, Pedro, and Daniel Peña. "Covariance changes detection in multivariate time series." *Journal of Statistical Planning and Inference* 137.1 (2007): 194-211.

Finding a single changepoint



Cusum statistic for a time series with three changepoints.

Galeano, Pedro, and Daniel Peña. "Covariance changes detection in multivariate time series."
Journal of Statistical Planning and Inference 137.1 (2007): 194-211.

Finding a single changepoint

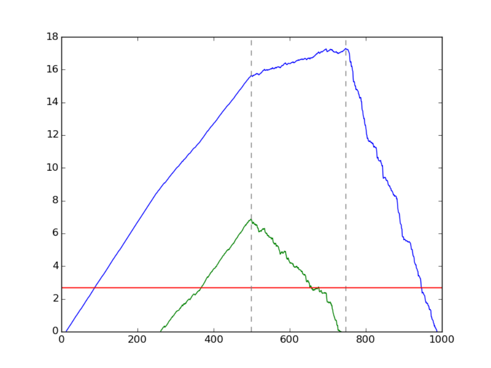
$$\text{test statistic} = \frac{h}{\sqrt{2kn}}\left(\frac{\sum_{t=1}^{h} \boldsymbol{e}_t^\top (\hat{\Sigma})^{-1}\boldsymbol{e}_t}{h} - \frac{\sum_{t=1}^{n} \boldsymbol{e}_t^\top (\hat{\Sigma})^{-1}\boldsymbol{e}_t}{n}\right)$$

$$\max(\text{test statistic}) \xrightarrow{D} \sup\left\{\text{Brownian bridge}\right\},$$

which is a known distribution, so we can compute critical values.

Galeano, Pedro, and Daniel Peña. “Covariance changes detection in multivariate time series.” *Journal of Statistical Planning and Inference* 137.1 (2007): 194-211.

Finding more changepoints



Cusum statistic for a time series with three changepoints.

Galeano, Pedro, and Daniel Peña. "Covariance changes detection in multivariate time series."

Journal of Statistical Planning and Inference 137.1 (2007): 194-211.

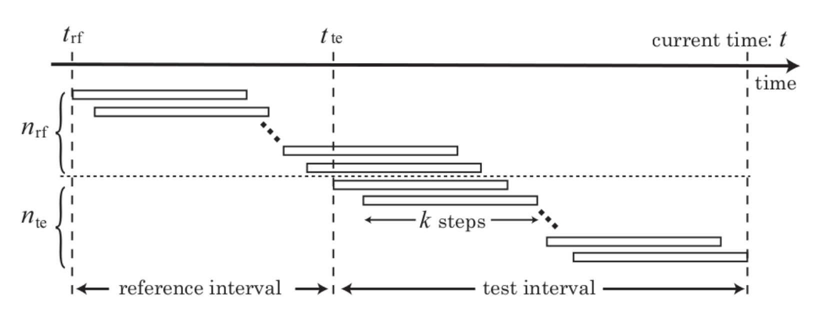
Moving from datapoints to subsequences

Suppose time series $\boldsymbol{y}(t) \in \mathbb{R}^3$ and denote its ith component by $y_i(t) \in \mathbb{R}$. Define the subsequence

$$\boldsymbol{Y}(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \\ y_1(t+1) \\ y_2(t+1) \\ y_3(t+1) \\ \vdots \\ y_1(t+m-1) \\ y_2(t+m-1) \\ y_3(t+m-1) \end{pmatrix}.$$

Kawahara, Yoshinobu and Masashi Sugiyama. "Sequential changepoint detection based on direct densityratio estimation." *Statistical Analysis and Data Mining* 5.2 (2012): 114-127.

Moving from datapoints to subsequences



Sliding window of subsequences.
(Reference interval has pdf p_{rf} and test interval has pdf p_{te}.)

Kawahara, Yoshinobu and Masashi Sugiyama. "Sequential changepoint detection based on direct densityratio estimation." *Statistical Analysis and Data Mining* 5.2 (2012): 114-127.

Hypothesis test

$$
\begin{aligned}
& H_0\colon p(\boldsymbol{Y}(i)) = p_{\mathrm{rf}}(Y(i)) && \text{for } i = t_{\mathrm{rf}}, \ldots, t-1 \\
\text{vs} \quad & H_1\colon p(\boldsymbol{Y}(i)) = p_{\mathrm{rf}}(Y(i)) && \text{for } i = t_{\mathrm{rf}}, \ldots, t_{\mathrm{te}}-1 \\
& \qquad p(\boldsymbol{Y}(i)) = p_{\mathrm{te}}(Y(i)) && \text{for } i = t_{\mathrm{te}}, \ldots, t-1.
\end{aligned}
$$

Kawahara, Yoshinobu and Masashi Sugiyama. "Sequential changepoint detection based on direct densityratio estimation." *Statistical Analysis and Data Mining* 5.2 (2012): 114-127.

Hypothesis test

$$
\begin{aligned}
& H_0\colon p(\boldsymbol{Y}(i)) = p_{\mathrm{rf}}(Y(i)) && \text{for } i = t_{\mathrm{rf}}, \ldots, t-1 \\
\text{vs} \quad & H_1\colon p(\boldsymbol{Y}(i)) = p_{\mathrm{rf}}(Y(i)) && \text{for } i = t_{\mathrm{rf}}, \ldots, t_{\mathrm{te}}-1 \\
& \quad\;\; p(\boldsymbol{Y}(i)) = p_{\mathrm{te}}(Y(i)) && \text{for } i = t_{\mathrm{te}}, \ldots, t-1.
\end{aligned}
$$

$$
\text{likelihood ratio} = \prod_{i=1}^{n_{\mathrm{te}}} \frac{p_{\mathrm{te}}(\boldsymbol{Y}_{\mathrm{te}}(i))}{p_{\mathrm{rf}}(\boldsymbol{Y}_{\mathrm{te}}(i))}
$$

Kawahara, Yoshinobu and Masashi Sugiyama. “Sequential changepoint detection based on direct densityratio estimation.” *Statistical Analysis and Data Mining* 5.2 (2012): 114-127.

Directly estimating the density ratio

$$w(\boldsymbol{Y}) = \frac{p_{\mathrm{te}}(\boldsymbol{Y})}{p_{\mathrm{rf}}(\boldsymbol{Y})}$$

Kawahara, Yoshinobu and Masashi Sugiyama. "Sequential changepoint detection based on direct densityratio estimation." *Statistical Analysis and Data Mining* 5.2 (2012): 114-127.

Directly estimating the density ratio

$$w(\boldsymbol{Y}) = \frac{p_{\mathrm{te}}(\boldsymbol{Y})}{p_{\mathrm{rf}}(\boldsymbol{Y})}$$

$$\hat{w}(\cdot) = \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\cdot)$$

Kawahara, Yoshinobu and Masashi Sugiyama. “Sequential changepoint detection based on direct densityratio estimation.” *Statistical Analysis and Data Mining* 5.2 (2012): 114-127.

Directly estimating the density ratio

$$w(\boldsymbol{Y}) = \frac{p_{\mathrm{te}}(\boldsymbol{Y})}{p_{\mathrm{rf}}(\boldsymbol{Y})}$$

$$\hat{w}(\cdot) = \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\cdot) = \sum_{\ell=1} \alpha_\ell K_\sigma(\cdot, \qquad)$$

Gaussian kernel $K_\sigma(\boldsymbol{Y}, \boldsymbol{Y}') = \exp\left(-\frac{\|\boldsymbol{Y}-\boldsymbol{Y}'\|^2}{2\sigma^2}\right)$

Kawahara, Yoshinobu and Masashi Sugiyama. "Sequential changepoint detection based on direct densityratio estimation." *Statistical Analysis and Data Mining* 5.2 (2012): 114-127.

Directly estimating the density ratio

$$w(\boldsymbol{Y}) = \frac{p_{\mathrm{te}}(\boldsymbol{Y})}{p_{\mathrm{rf}}(\boldsymbol{Y})}$$

$$\hat{w}(\cdot) = \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\cdot) = \sum_{\ell=1}^{n_{\mathrm{te}}} \alpha_\ell K_\sigma(\cdot, \boldsymbol{Y}_{\mathrm{te}}(\ell))$$

Gaussian kernel $K_\sigma(\boldsymbol{Y}, \boldsymbol{Y}') = \exp\left(-\frac{\|\boldsymbol{Y}-\boldsymbol{Y}'\|^2}{2\sigma^2}\right)$

Kawahara, Yoshinobu and Masashi Sugiyama. "Sequential changepoint detection based on direct densityratio estimation." *Statistical Analysis and Data Mining* 5.2 (2012): 114-127.

Directly estimating the density ratio

$$\hat{p}_{\mathrm{te}}(\cdot) := p_{\mathrm{rf}}(\cdot)\hat{w}(\cdot) = p_{\mathrm{rf}}(\cdot)\sum_{\ell=1}^{n_{\mathrm{te}}}\alpha_\ell K_\sigma(\cdot, \boldsymbol{Y}_{\mathrm{te}}(\ell))$$

Kawahara, Yoshinobu and Masashi Sugiyama. "Sequential changepoint detection based on direct densityratio estimation." *Statistical Analysis and Data Mining* 5.2 (2012): 114-127.

Directly estimating the density ratio

$$\hat{p}_{\mathrm{te}}(\cdot) := p_{\mathrm{rf}}(\cdot)\hat{w}(\cdot) = p_{\mathrm{rf}}(\cdot)\sum_{\ell=1}^{n_{\mathrm{te}}} \alpha_\ell K_\sigma(\cdot, \boldsymbol{Y}_{\mathrm{te}}(\ell))$$

$$\begin{array}{rl} \min\limits_{\{\alpha_\ell\}} & \text{Kullback-Leibler divergence}(p_{\mathrm{te}} \| \hat{p}_{\mathrm{te}}) \\ \text{subject to} & \text{some uninteresting constraints} \end{array}$$

Kawahara, Yoshinobu and Masashi Sugiyama. “Sequential changepoint detection based on direct densityratio estimation.” *Statistical Analysis and Data Mining* 5.2 (2012): 114-127.

Directly estimating the density ratio

$$\hat{p}_{\mathrm{te}}(\cdot) := p_{\mathrm{rf}}(\cdot)\hat{w}(\cdot) = p_{\mathrm{rf}}(\cdot)\sum_{\ell=1}^{n_{\mathrm{te}}}\alpha_\ell K_\sigma(\cdot, \boldsymbol{Y}_{\mathrm{te}}(\ell))$$

$$\begin{aligned} \min_{\{\alpha_\ell\}} \quad & \text{Kullback-Leibler divergence}(p_{\mathrm{te}} \| \hat{p}_{\mathrm{te}}) \\ \text{subject to} \quad & \text{some uninteresting constraints} \end{aligned}$$

Kawahara, Yoshinobu and Masashi Sugiyama. "Sequential changepoint detection based on direct densityratio estimation." *Statistical Analysis and Data Mining* 5.2 (2012): 114-127.