Discussion of *Automatic Change-Point Detection in Time Series via Deep Learning* by Jie Li, Paul Fearnhead, Piotr Fryzlewicz and Tengyao Wang

Yudong Chen and Yining Chen

London School of Economics and Political Science, London, U.K.

E-mail: y.chen**276**@lse.ac.uk, y.chen**101**@lse.ac.uk

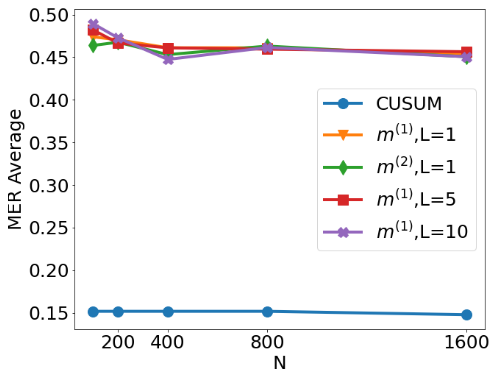
We congratulate the authors for providing this stimulating work which uses deep neural networks for detecting change-points. We shall comment on three aspects of the paper: (i) the standardisation, (ii) distributional assumptions of the training and testing samples, and (iii) potentially more efficient use of the training samples.

First, we emphasise that, by default, classification procedures based on neural networks will not automatically be invariant to shifting or scaling the data. To illustrate this point, we consider Scenario (S1) in the original manuscript with $\rho_t = 0.5$, but with no standardisation performed on the training or testing sets. Figure 1(a) shows the results where we additionally added an independent $U \sim U[-2, 2]$ mean-shift to each series from the testing sets. Here the proposed method (without standardisation) is no longer desirable even when we increase the size of the training samples, N, to 1600. The standardisation procedure proposed in the paper (i.e. scaling to [0,1]) will alleviate this issue. Using trimmed mean and standard deviation estimates would be a more robust option. From our experiments, other approaches such as adding a random baseline to all training samples would also work.

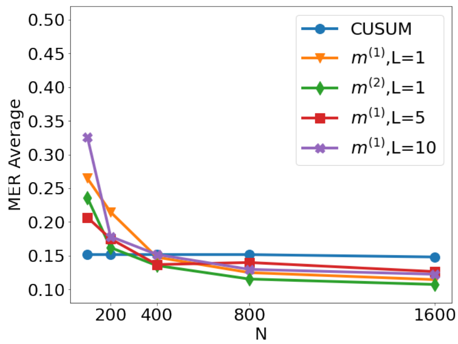
Nevertheless, if the distributions of the training and test samples, denoted by $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$, are indeed the same, then performing standardisation might lead to worse performance, as is demonstrated in Figure 1(b) and Figure 1(c). Although, as demonstrated before, without standardisation, the resulting classifier might not be transferable to even slightly different settings.

Note that the presented theory requires $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ to be the same. We anticipate similar results to hold if $\mathcal{D}_{train}$ is a finite mixture with one component being $\mathcal{D}_{test}$. More broadly speaking, consistency should hold if the support of $\mathcal{D}_{train}$ contains that of $\mathcal{D}_{test}$. In addition, one could use different labels for different types of change-points (e.g. change in mean/variance/etc) in the training set, and apply the existing approach to learn a multi-class classifier.

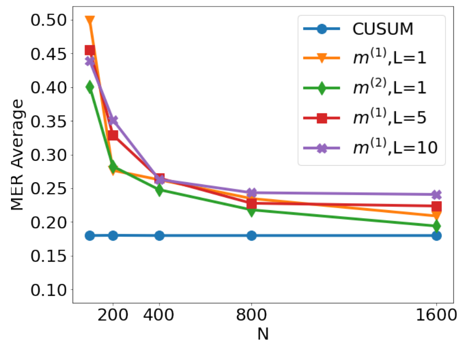
Finally, we believe that the training samples can be used in a more efficient manner with a minor twist. For $X_1, X_2, \ldots, X_n$ with label Y, its *reversed* sequence $X_n, X_{n-1}, \ldots, X_1$ should also have the same label. Consequently, we could double the size of N by adding all the reversed sequences into the training set. Sizeable improvement with this extra step can be seen, especially when N is small, by comparing Figure 1(d) with Figure 1(e).



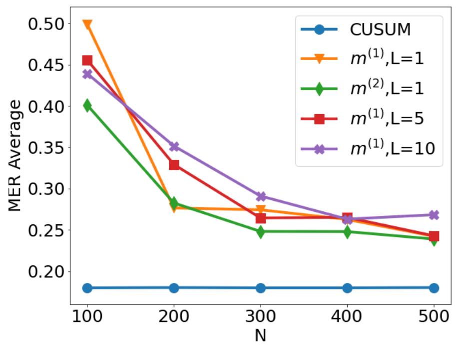
(a) $\mu_L = 0$ for the training samples, $\mu_L \sim U[-2, 2]$ for the testing samples; data not standardised



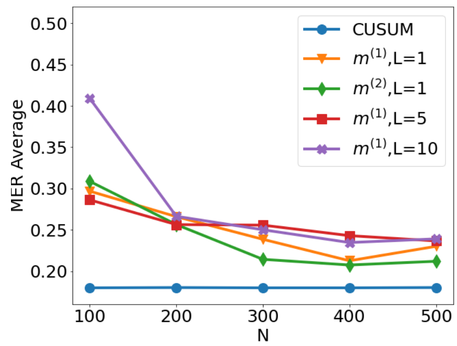
(b) $\mu_L = 0$ for both training and testing; data not standardised



(c) $\mu_L = 0$ for both training and testing; data standardised



(d) $\mu_L = 0$ for both training and testing; data standardised; no reversed samples added



(e) $\mu_L = 0$ for both training and testing; data standardised; reversed samples added

Fig. 1: Plot of test set misclassification error rate (MER), computed on a test set of size 150000, against training sample size N for detecting the existence of a change-point on data series of length $n = 100$ under Scenario S1 in Section 5 of the original paper, with the exception that $\rho_t = 0.5$. We compare the performance of the CUSUM test and neural networks from four function classes as specified in Section 5 of the paper. Here we use batch size of 32 for training with no regularisation. For standardisation, we use the approach suggested in the paper. For the testing samples in Figure (a) only, we added the independent mean-shift of $U[-2, 2]$ for each series.