ruptures: change point detection in Python

3 authors:

Charles Truong
11 PUBLICATIONS   391 CITATIONS

SEE PROFILE

Laurent Oudre
Ecole Normale Supérieure Paris-Saclay
79 PUBLICATIONS   782 CITATIONS

SEE PROFILE

Nicolas Vayatis
Ecole normale supérieure Paris-Saclay
191 PUBLICATIONS   2,730 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Graph signal processing View project

Change point detection View project

ruptures: change point detection in Python

Charles Truong   TRUONG@CMLA.ENS-CACHAN.FR
Nicolas Vayatis   VAYATIS@CMLA.ENS-CACHAN.FR
CMLA, ENS Cachan, CNRS, Université Paris-Saclay
94235, Cachan, France
COGNAC G, University Paris Descartes, CNRS
75006 Paris, France

Laurent Oudre   LAURENT.OUDRE@UNIV-PARIS13.FR
L2TI, University Paris 13
93430 Villetaneuse, France

Editor:

Abstract

`ruptures` is a Python library for offline change point detection. This package provides methods for the analysis and segmentation of non-stationary signals. Implemented algorithms include exact and approximate detection for various parametric and non-parametric models. `ruptures` focuses on ease of use by providing a well-documented and consistent interface. In addition, thanks to its modular structure, different algorithms and models can be connected and extended within this package.

Keywords: Change Point Detection, Signal Segmentation, Time Series, Python

1. Introduction

Change point detection is the task of finding changes in the underlying model of a signal. This subject has generated important activity in statistics and signal processing (Lavielle, 2005; Jandhyala et al., 2013; Haynes et al., 2017). Modern applications in bioinformatics, finance, monitoring of complex systems have also motivated recent developments from the machine learning community (Vert and Bleakley, 2010; Lajugie et al., 2014; Hocking et al., 2015).

We present `ruptures`, a Python scientific library for multiple change point detection in multivariate signals. It is meant to answer the growing need for fast exploration, by non-specialists, of non-stationary signals. In addition, we expect that removing the cost of reimplementation will facilitate composition of new algorithms. To that end, `ruptures` insists on an easy-to-use and consistent interface. Implementation is also modular to allow users to seamlessly plug their own code.

To the best of the authors' knowledge, `ruptures` is the first Python package dedicated to multiple change point detection. Most related softwares are implemented in R (Erdman and Emerson, 2007; James and Matteson, 2014; Killick and Eckley, 2014; Ross, 2015; Fryzlewicz, 2017; Haynes et al., 2017; Chakar et al., 2017). However, few provide more than one algorithm, and even fewer can be applied to detect changes other than mean shifts. On the other hand, `ruptures` contains several standard methods as well as recent contributions, most of which are not available elsewhere (in Python or R). Our work encompasses most packages and provides a unique framework to run and

evaluate all algorithms.

In the following, we quickly describe the change point detection framework. Then the main features of the library are detailed.

2. Change point detection framework

In the offline (or retrospective) change point detection framework, we consider a non-stationary random process $y = \{y_1, \ldots, y_T\}$ that takes value in $\mathbb{R}^d$ ($d \geq 1$). The signal y is assumed to be piecewise stationary, meaning that some characteristics of the process change abruptly at some unknown instants $t_1^\star < t_2^\star < \cdots < t_K^\star$. Change point detection consists in estimating those instants when a particular realization of y is observed. Note that the number of changes K is not necessarily known.

Most estimation methods adhere to or are an approximation of a general format where a suitable contrast function $V(\cdot)$ is minimized (Jandhyala et al., 2013; Lavielle, 2005). Usually, it is written as a sum of segment costs:

$$V(\mathbf{t}, y) := c(\{y_t\}_1^{t_1}) + c(\{y_t\}_{t_1+1}^{t_2}) + \cdots + c(\{y_t\}_{t_i+1}^{t_{i+1}}) + \ldots \tag{1}$$

where $\mathbf{t} = \{t_1, t_2, \ldots\}$ denotes a set of change point indexes and $c(\cdot)$ denotes a cost function that takes a process as input and measures its goodness-of-fit to a specified model. The contrast $V(\cdot)$ is the total cost associated with choosing a particular segmentation $\mathbf{t}$. Change point detection amounts to solving the following discrete optimization problem:

$$\min_{\mathbf{t}} V(\mathbf{t}, y) + \text{pen}(\mathbf{t}) \tag{2}$$

where $\text{pen}(\mathbf{t})$ is a regularizer on the value of the partition $\mathbf{t}$. Methods from the literature essentially differ by 1) the constraints they add to this optimization problem (fixed dimension of $\mathbf{t}$, penalty term, cost budget, etc.), 2) how they search for the solution (exact or approximate resolution, local or sequential, etc.) and 3) the cost function $c(\cdot)$ they use (which is related to the type of change).

3. Library overview

A basic flowchart is displayed on Figure 1. Each block of this diagram is described in the following brief overview of ruptures ' features. More information can be found in the related documentation (see link to source in Section 3.2).

3.1 Main features

- **Search methods** Our package includes the main algorithms from the literature, namely dynamic programming, detection with a l_0 constraint, binary segmentation, bottom-up segmentation and window-based segmentation. This choice is the result of a trade-off between exhaustiveness and adaptiveness. Rather than providing as many methods as possible, only algorithms which have been used in several different settings are included. In particular, numerous "mean-shift only" detection procedures were not considered. Implemented algorithms have sensible default parameters that can be changed easily through the functions' interface.
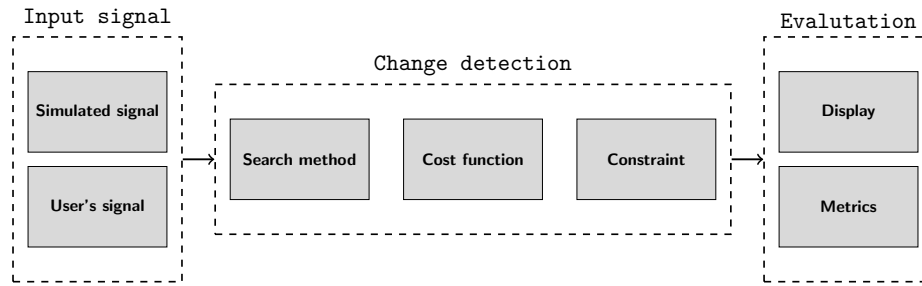
Figure 1: Schematic view of the `ruptures` package.

- **Cost functions** Cost functions are related to the type of change to detect. Within `ruptures`, one has access to parametric cost functions that can detect shifts in standard statistical quantities (mean, scale, linear relationship between dimensions, autoregressive coefficients, etc.) and non-parametric cost functions (kernel-based or Mahalanobis-type metric) that can, for instance, detect distribution changes (Harchaoui and Cappé, 2007; Lajugie et al., 2014).

- **Constraints** All methods can be used whether the number of change points is known or not. In particular, `ruptures` implements change point detection under a cost budget and with a linear penalty term (Killick et al., 2012; Maidstone et al., 2017).

- **Evaluation** Evaluation metrics are available to quantitatively compare segmentations, as well as a display module to visually inspect algorithms' performances.

- **Input** Change point detection can be performed on any univariate or multivariate signal that fits into a *Numpy* array. A few standard non-stationary signal generators are included.

- **Consistent interface and modularity** Discrete optimization methods and cost functions are the two main ingredients of change point detection. Practically, each is related to a specific object in the code, making the code highly modular: available optimization methods and cost functions can be connected and composed. An appreciable by-product of this approach is that a new contribution, provided its interface follows a few guidelines, can be integrated seamlessly into `ruptures`.

- **Scalability** Data exploration often requires to run several times the same methods with different sets of parameters. To that end, a cache is implemented to keep intermediate results in memory, so that the computational cost of running the same algorithm several times on the same signal is greatly reduced. We also add the possibility for a user with speed constraints to sub-sample their signals and set a minimum distance between change points.

3.2 Availability and requirements

The `ruptures` library is written in pure Python and available on Mac OS X, Linux and Windows platforms. Source code is available from reine.cmla.ens-cachan.fr[1] under the BSD license. We also provide a complete documentation that includes installation instructions, explanations with code snippets on advance use (ctruong.perso.math.cnrs.fr/ruptures).

Implementation relies on *Numpy* as the base data structure for signals and parameters and *Scipy*

---

1. https://reine.cmla.ens-cachan.fr/c.truong/ruptures/repository/latest/archive.zip

(a) Python code.

(b) Top and middle: simulated 2D signal; regimes are highlighted in alternating gray area. Below: scatter plots for each regime type.
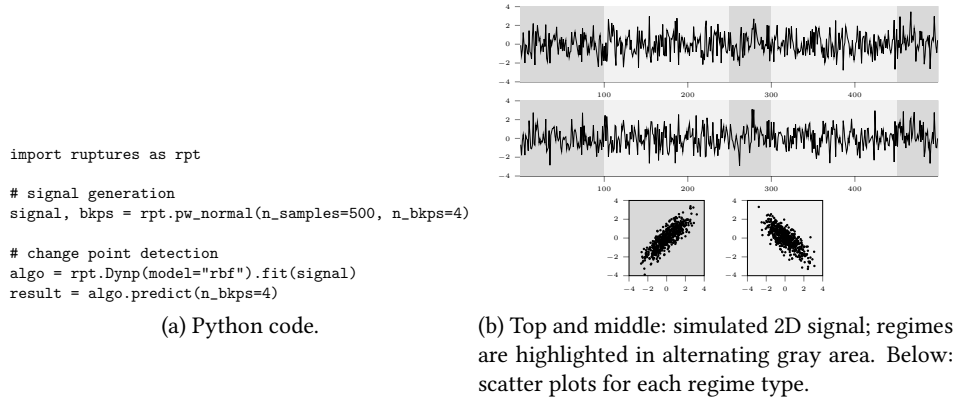
Figure 2: Illustrative example.

for efficient linear algebra and array operations. The *Matplotlib* library is recommended for visualization. Unit tests (through the *Pytest* library) are provided to facilitate the validation of new pieces of code.

3.3 Illustrative example

As an illustrative example, we perform a kernel change point detection on a simulated piecewise stationary process (Harchaoui and Cappé, 2007). In a nutshell, this method maps the input signal onto a high-dimensional Hilbert space $\mathcal{H}$ through a kernel function (here, we use the radial basis function) and searches for mean shifts.

First, random change point indexes are drawn and a 2D signal of i.i.d. centred normal variables with changing covariance matrix is simulated (Figure 2b). The algorithm's internal parameters are then fitted on the data. The discrete minimization of the contrast function is performed with dynamic programming and the associated estimates are returned. The related code lines are reported on Figure 2a.

It is worth mentioning that only a few instructions are needed to perform the segmentation. In addition, thanks to `ruptures` , variations of the kernel change point detection can be easily carried out by changing a few parameters in this code.

4. Conclusion

`ruptures` is the most comprehensive change point detection library. Its consistent interface and modularity allow painless comparison between methods and easy integration of new contributions. In addition, a thorough documentation is available for novice users. Thanks to the rich Python ecosystem, `ruptures` can be used in coordination with numerous other scientific libraries

Acknowledgments

References

S. Chakar, É. Lebarbier, C. Levy-Leduc, and S. Robin. A robust approach for estimating change-points in the mean of an AR(1) process. *Bernouilli Society for Mathematical Statistics and Probability*, 23(2):1408–1447, 2017.

C. Erdman and J. W. Emerson. bcp: an R package for performing a Bayesian analysis of change point problems. *Journal of Statistical Software*, 23(3):1–13, 2007.

P. Fryzlewicz. breakfast: multiple change-point detection and segmentation, 2017. URL https://cran.r-project.org/package=breakfast.

Z. Harchaoui and O. Cappé. Retrospective mutiple change-point estimation with kernels. In *Proceedings of the IEEE/SP Workshop on Statistical Signal Processing*, pages 768–772, Madison, Wisconsin, USA, 2007.

K. Haynes, I. A. Eckley, and P. Fearnhead. Computationally efficient changepoint detection for a range of penalties. *Journal of Computational and Graphical Statistics*, 26(1):134–143, 2017.

T. Hocking, G. Rigaill, and G. Bourque. PeakSeg: constrained optimal segmentation and supervised penalty learning for peak detection in count data. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 324–332, Lille, France, 2015.

N. A. James and D. S. Matteson. ecp: an R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software*, 62(7):1–25, 2014.

V. Jandhyala, S. Fotopoulos, I. Macneill, and P. Liu. Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*, 34(4):423–446, 2013.

R. Killick and I. A. Eckley. changepoint: an R package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19, 2014.

R. Killick, P. Fearnhead, and I. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

R. Lajugie, F. Bach, and S. Arlot. Large-margin metric learning for constrained partitioning problems. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 297–395, Beijing, China, 2014.

M. Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501–1510, 2005.

R. Maidstone, T. Hocking, G. Rigaill, and P. Fearnhead. On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27(2):519–533, 2017.

G. J. Ross. Parametric and nonparametric sequential change detection in R: the cpm package. *Journal of Statistical Software*, 66(3), 2015.

J.-P. Vert and K. Bleakley. Fast detection of multiple change-points shared by many signals using group LARS. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, volume 1, pages 2343–2351, Vancouver, Canada, 2010.