Change point detection

Document: **functional specification**
Date: **September 21, 2014**
Author: **Buzun Nazar, postrealist@gmail.com**
Version: **0.1**
Organization: **http://premolab.ru/**
Project repository:
/Users/<MacUser>/Dropbox/aamo/eklmn/projects/brain/origin/change_point.git

Contents

1 Data format

Input data file consists of header and dataset parts. Header contains description for each column in the dataset (column name, type, participation). Type could be integer, real, boolean or categorical. Participation indicates active parameters of projection where change point is considered. Each dataset row allows one type of separators (",", " ", "tab").

Example:

```
DATASET_NAME: name;
column1 integer 0;
column2 categorical cat1, cat2 1;
column3 real 1;

DATASET
1,cat1,0.32
3,cat1,23.23
3,cat2,21.2
...
```

2 Probabilistic model types [PMT]

1. Standart: $X_i \sim P_\theta$;
2. Partially standart: $Y_i \sim P_{f(X_i,\theta)}$, where P is standart and f is user defined, Y and X are columns from dataset;
3. User specified $X_i \sim P_\theta$, P is defined manually;
4. Diagrams: data without model.

3 Usage scenarios

1. **Data generation:**
   - user chooses PMT (1,2 or 3);
   - dataset length (n);
   - defines CP (change point) locations $0 = t_0 < t_1, \ldots, t_k \leq n$ in file;

- for each change point PMT parameter is defined in the same file.
- program returns file with dataset;

2. **PMT parameters generation:** using Markov chains user could obtain CP locations and corresponded parameters.

   - user chooses PMT (1,2 or 3);
   - provides array of parameters for the selected PMT;
   - sets parameter change probability;
   - sets probability that next value is equal to previous;
   - sets whether parameter changes stepwise or continuously;
   - defines length of continuous parameter change interval.

3. **Change points search:**

   Given time interval (t_1, t_2) (range in Dataset) system finds the only one CP (s), where achieved minimum of

$$d_h(s,b) = \|f_\triangle(b) - \sqrt{2\triangle L}[s - nh/2, s + nh/2]\| \to \min_{b,s} \quad t_1 \le s < t_2,$$

$$f_\triangle(b) = b^2\left(1 - \frac{2|t|}{nh}\right),\ t = [-nh/2, nh/2].$$

   (a) Offline version:

   - user gives the program file with dataset and describe PMT;
   - sets through command line the other parameters (output file path, window size of array of window sizes, mistake probability).
   - in the output file each row has following format:
     CP number (sequent CPs have the same number),
     CP location in the input dataset (s),
     confidence interval of the location,
     probability of CP absence,
     max $\sqrt{2\triangle L}$ value in CP,
     $\mathbb{E}_{boot}\sqrt{2\triangle L}$ in CP,
     $\sqrt{\mathbb{D}_{boot}\sqrt{2\triangle L}}$ in CP,
     $\|\triangle\theta^*_{12}\|$ in CP;
   - program also creates file with $\sqrt{2\triangle L}(t)$, $\mathbb{E}_{boot}\sqrt{2\triangle L}(t)$ with sd intervals in each point, $\|\triangle\xi^*_{12}\|(t)$, $\|\triangle\theta^*_{12}\|(t)$.

   (b) Online version:

   - user gives the program file with dataset and describe PMT;
   - user receives answer whether there is a CP a cote to the end of the dataset;
   - if change point is detected its details should be written to output file in format described in the offline version.

4. **CP removing:** For bootstrap application Dataset ranges without change point are required. CP removing carried out by steps:

   - system creates $\mathbb{E}_{boot}(2\triangle L)(t) = E(t)$,

$$2\mathbb{E}\triangle L \approx \triangle(\theta^*_{12})^2 + \mathbb{E}\|\triangle\xi_{12}\|^2.$$

   - Dataset is divided into parts, where $E(t)$ close to $\min_t E(t)$.

- [Optional] if model has different effective dimensions, Dataset is divided into parts, where $E(t)$ close to $\text{const}(t_a, t_b), t_b - t_a > \delta t_{\min}$.

5. **CP intervals detection** User provides array of window lengths $[h_1, \ldots, h_m]$. For each h_i and each window position t system fits $f_\triangle(b)$. If $b(t) > b_{boot}$, then t is a candidate for CP.

   System unites CP from different h_i and returns array of time intervals of change points.

6. **CP intervals single point verification**

   Given time interval (t_1, t_2), where each internal point is a candidate for CP, system checks whether the interval contains multiple CP. If there are more than one CP system splits (t_1, t_2) and returns subintervals with single CP.

   Split points could be detected by decreasing h and checking hypothesis $\theta_1^* = \theta_2^*$.

7. **Visualization:** System is able to create following graphics:

   - $\sqrt{2\triangle L}(t)$, $t \in [a, b]$;
   - $\mathbb{E}_{boot}\sqrt{2\triangle L}(t)$ with sd interval in each point;
   - $\sqrt{2\triangle L}(t)$ with marked CP intervals and fitted $f_\triangle(b)$ in each CP interval;
   - Regression plot $y_t \sim P_{(t,\theta_t)}$, $y_t \in \mathbb{R}$, and its mean $\overline{y}(t) = \mathbb{E}y_t$;
   - Animated regression plot sync with $\sqrt{2\triangle L}(t)$.

8. **Web demo:**

   - user loads web page from address like *demo.premolab.ru/cpd*;
   - selects model $y_t \sim P_{(t,\theta_t)}$ from the proposed list;
   - selects PMT parameters (θ_t) generator;
   - the page displays dynamic plot of y_t and $\sqrt{2\triangle L}(t)$;
   - system executes online CP search, displays triangles on $\sqrt{2\triangle L}(t)$ in each CP and draws forecast line $(\overline{y}(t))$ on y_t plot, that has bend in each CP.