Sparse Group Lasso for Change-points Detection

Xin Zhang and Ye Tian

***Abstract*—We propose a piecewise-constant model for detecting the multi-changepoints among a group of heterogeneous individuals. In this model, In this model, we think only small part of the individuals have changepoints and the number of changepoint locations is small. We want to identify the individual with changpoint and locate the shared changepoint loacation, estimate the jumps. Sparse Group Lasso and block coordinate descent are applied in solving our model.**

***Index Terms*—sparse group lasso.**

I. Introduction

Change-point detection, also known as analysis of structure break, origins in quality control. The aim of this issue is to find the position (place or time) where most or all of individuals jointly change in some specific way, including mean changing, variance changing and correlation changing [1]. So far, It has been broadly applied in many filed:

1) Audio and image processing [2][5]: Changepoint detection methods are applied for audio segmentation. Also it can be used to recognize the gaps between sentences and words, as well as detect the noise in the audio recording;
2) Intrusion detection in computer networks [6]: It is necessary to detect harmful and illegal intrusions into a computer network. By detecting the traffic change, we can locate the place where might have intrusions;
3) Financial and economics time series analysis [7]: Monitoring the price change in stock is a very important topic in finance. Changepoint method could be applied to detect whether the process of price change get into non-stationary status;
4) Biomarker analysis [8]: In order to identify the biomarker, biologists often recording a time series which reflects the physiological process. Comparing with the experiment conditions, we could find a "perfect" biomarker that indicates the status of organism.

So far, many methods have been developed for changepoint detection problem. In statistics, a famous one is hypothesis testing method, e.g. CUSUM test and ratio test. By giving a distribution to the data, we could form certain statistics which could reflect the difference between two time periods. Then the changepoint could be located by testing these statistics. Another method is from time series analysis. By fitting some time series model, we could find the changepoint from the estimated parameters. Also, some works focus on online changepoint detection. Common used is Bayesian method. First, combined with prior distribution, the known data could be used to compute the posterior distribution. Then using the posterior, we could check the upcoming data. Here we want to focus on a linear model with penalty term. According to the mechanism of data generating, the model and corresponding assumptions could be summarized. Then we could formulate the objective function by transforming assumptions into penalty in model solving.

In our project, we propose a piecewise-constant model for detecting the changepoints among a group of heterogeneous individuals. In this model, we think only small part of the individuals have changepoints and the number of changepoint locations is small. The goal of our method is first toidentify the individuals with changepoint, and then locate the changepoints and estimate the jumps.

II. Notation

For any two integers $u \leq v$, $[u, v]$ denotes the interval $\{u, u+1, \ldots, v\}$. For any $u \times v$ matrix M, $M_{i,j}$ denotes its (i,j)-th entry, and $\|M\|_F = \sqrt{\sum_{i=1}^{u} \sum_{j=1}^{v} M_{i,j}^2}$ is Frobenius norm (or Euclidean norm in the case of vectors).

For simplicity, we will use $\bullet$ to represent line or column instead of $[1, u]$ or $[1, v]$, i.e., $A_{i,\bullet}$ is the i-th row of A and $A_{\bullet,j}$ is the j-th column of A. We note $\mathbf{1}_{u,v}$ the $u \times v$ matrix of ones, and $\mathbf{I}_p$ the $p \times p$ identity matrix.

III. Formulation

We consider p real-valued profiles of length n, stored in an $n \times p$ matrix Y. The i-th profile $Y_{\bullet,i} = (Y_{1,i}, \ldots, Y_{n,i})$ is the i-th column of Y. And we have following assumptions: (1)

1) Profile sparsity: Most profiles have no changepoints, i.e. the number of individuals with changepoint $p_1 << p$;
2) Changepoint sparsity: The number of changepoints is very small, compared with n;
3) Shared location: Changepoints across profiles share the same location.

Then we propose piecewise constant model. We think the observation Y is true parameters U with error ϵ:

$$Y = U + \epsilon.$$

For those without changepoint, the corresponding column of U, let's say $U_{\bullet,i}$, is a constant vector $\gamma_i 1_{n\times 1}$; for those with changpoint, the corresponding column $U_{\bullet,i}$ is

$$(\gamma_i, ..., \gamma_i, \gamma_i + \beta_{i1}, ..., \gamma_i + \beta_{i1}, ..., \gamma_i + \beta_{ik}, \gamma_i + \beta_{ik}).$$

Thus, the objective function is following:

$$\min_{U \in \mathbb{R}^{n\times p}} ||Y - U||_F^2 \tag{1}$$

$$s.t \sum_{i=1}^{n-1} \delta(U_{i+1} - U_i) \leq k_1 \tag{2}$$

$$\sum_{i=1}^{n-1} \sum_{j=1}^{p} |U_{i+1,j} - U_{i,j}|_0 \leq k_2. \tag{3}$$

δ is the Dirac function, equals to 0 if its argument is 0, 1 otherwise. (1) is to minimize the error term $\epsilon = Y - U$; (2) is to control the total number of changepoints; and (3) is to control the number of individual with chang-points.

Then, we could get the lagrangian of above problem by relaxing the constrains:

$$\min_{U\in\mathbb{R}^{n\times p}} \frac{1}{2}||Y-U||_F^2 + \lambda_1 \sum_{i=1}^{n-1} ||U_{i+1,\bullet} - U_{i,\bullet}||_2^2 + \lambda_2 \sum_{i=1}^{n-1}\sum_{j=1}^{p} |U_{i+1,j} - U_{i,j}|. \qquad (4)$$

Here λ_1 and λ_2 are two tuning parameters. With piecewise constant model, we can get U as following:

$$\begin{aligned} &\beta_{i,\bullet} = U_{i+1,\bullet} - U_{i,\bullet}, \forall i = 1, \\ &U_{1,\bullet} = \gamma, \\ &U_{i,\bullet} = \gamma + \sum_{j=1}^{i-1} \beta_{j,\bullet}, \forall i. \end{aligned}$$

$$U = 1_{n,1}\gamma + X\beta,, \qquad (5)$$

where $X \in \mathbb{R}^{n\times(n-1)}$ and

$$X = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 1 & 1 \end{bmatrix}$$

. Plugging all the parameters, β and γ, into (4), we have

$$\min_{\beta\in\mathbb{R}^{(n-1)\times p}, \gamma\in\mathbb{R}^{1\times p}} \frac{1}{2}\| Y - X\beta - \mathbf{1}_{n,1}\gamma \|^2 + \lambda_1 \sum_{i=1}^{n-1} \| \beta_{i,\bullet} \| + \lambda_2 \sum_{i=1}^{n-1}\sum_{j=1}^{p} | \beta_{i,j} | , \qquad (6)$$

We could get γ by solving

$$\frac{\partial(6)}{\partial\gamma} = -1_{n,1}^T(Y - X\beta - 1_{n,1}\gamma) = 0 \qquad (7)$$

$$\Rightarrow \gamma = \frac{1_{1,n}(Y - X\beta)}{n} \qquad (8)$$

Plug $\gamma = \mathbf{1}_{1,n}(Y - X\beta)/n$ into (7), we have

$$\min_{\beta\in\mathbb{R}^{(n-1)\times p}} \frac{1}{2}\| \bar{Y} - \bar{X}\beta \|^2 + \lambda_1 \sum_{i=1}^{n-1} \| \beta_{i,\bullet} \| + \lambda_2 \sum_{i=1}^{n-1}\sum_{j=1}^{p} | \beta_{i,j} | . \qquad (9)$$

where $\bar{Y}$ and $\bar{X}$ are obtained from Y and X by centering each column.

IV. Implementation

A. Solution

For minimization problem (9), we consider the first order derivative:

$$\begin{aligned} \frac{\partial(9)}{\partial\beta_{i,\bullet}} &= -X_{\bullet,i}^{\top}(Y - X\beta) + \lambda_1 \frac{\beta_{i,\bullet}}{\|\beta_{i,\bullet}\|_2} + \lambda_2 sign(\beta_{i,\bullet}) \\ &= -X_{\bullet,i}^{\top}(Y - \sum_{j=1}^{n-1} X_{\bullet,j}\beta_{j,\bullet}) + \lambda_1 \frac{\beta_{i,\bullet}}{\|\beta_{i,\bullet}\|_2} + \lambda_2 sign(\beta_{i,\bullet}) \\ &= -X_{\bullet,i}^{\top}(Y - \sum_{j\neq i} X_{\bullet,j}\beta_{j,\bullet}) + X_{\bullet,i}^T X_{\bullet,i}\beta_{i,\bullet} \\ &\quad + \lambda_1 \frac{\beta_{i,\bullet}}{\|\beta_{i,\bullet}\|_2} + \lambda_2 sign(\beta_{i,\bullet}) \end{aligned} \qquad (10)$$

Set (10) as 0, we have

$$\begin{aligned} &(\lambda_1 \frac{1}{||\beta_{i,\bullet}||_2} + X_{\bullet,i}^T X_{\bullet,i})\beta_{i,\bullet} + \lambda_2 sign(\beta_{i,\bullet}) = \\ &\qquad X_{\bullet,i}^T(Y - \sum_{j\neq i} X_{\bullet,j}\beta_{j,\bullet}) \\ &(\lambda_1 \frac{1}{||\beta_{i,\bullet}||_2} + X_{\bullet,i}^T X_{\bullet,i})\beta_{i,k} + \lambda_2 sign(\beta_{i,k}) = \\ &\qquad X_{\bullet,i}^T(Y_{\bullet,k} - \sum_{j\neq i} X_{\bullet,j}\beta_{j,k}) \end{aligned} \qquad (11)$$

The second equation is considering each coordinate of the first equation. Also, for (11),

$$sign(\beta_{ij}) = \begin{cases} \frac{\beta_{ij}}{|\beta_{ij}|}, \text{if } \beta_{ij} \neq 0 \\ \in [-1, 1], o.w. \end{cases} \qquad (12)$$

The rest follows the algorithm in "a note" [9]. Block coordinate descent is used here for optimizing (9). Consider the residual $r^i = Y - \sum_{k\neq i} X_{\bullet,k}\beta_{k,\bullet}$. We have the subgradient equation

$$X_{\bullet,i}^T(r^i - \sum_{j=1}^{p} X_{\bullet,i}\beta_{i,\bullet}) + \lambda_1 s_i + \lambda_2 t_i = 0 \qquad (13)$$

where

$$\begin{cases} s_i = \frac{\beta_{i,\bullet}}{||\beta_{i,\bullet}||_2} & , \forall j \\ t_i = sign(\beta_{i,\bullet}) \end{cases} \qquad (14)$$

Define $a_i = X_{\bullet,i}r^i$. A necessary and sufficient condition for $\beta_{i,\bullet} = 0$ in (13) is following:

$$a_{ij} = \lambda_1 s_{ij} + \lambda_2 t_{ij}. \qquad (15)$$

Consider

$$J(t_i) = \frac{1}{\lambda_1^2}\sum_{j=1}^{p}(a_{ij} - \lambda_2 t_{ij})^2 = \sum_{j=1}^{p} s_{ij}^2. \qquad (16)$$

If $J(t_i) \leq 1$, then the minimizer for (16) is :

$$\hat{t}_{ij} = \begin{cases} \frac{a_{ij}}{\lambda_2}, \text{if } |\frac{a_{ij}}{\lambda_2}| \leq 1 \\ sign(\frac{a_{ij}}{\lambda_2}), o.w. \end{cases} \qquad (17)$$

So in this case, $\beta_{i,\bullet} = 0$. However if $J(t_i) > 1$, then it means that there is no feasible solution for (11) and not all the β_{ij} are zero. Then we need to solve the objective problem

$$\min_{\beta_{i,\bullet}} \frac{1}{2}||r^i - X_{\bullet,i}\beta_{i,\bullet}||_F^2 + \lambda_1||\beta_{i,\bullet}||_2 + \lambda_2 \sum_{j=1}^{p} |\beta_{ij}|. \quad (18)$$

Then we use coordinate descent method to solve problem (18). The subgradient equation is

$$-X_{\bullet,i}^T(r_{\bullet,j}^i - X_{\bullet,i}\beta_{ij}) + 2\lambda_1\beta_{ij} + \lambda_2 sign(\beta_{ij}) = 0, \quad (19)$$

$,\forall j = 1, 2, ..., p$. According to the subgradient equation of (18), if $|X_{\bullet,i}r_{\bullet,j}^i| \leq \lambda_2$, we have $\hat{\beta}_{ij} = 0$; otherwise, we minimize (18) with one-dimensional research over β_{ij}.

B. Pseudo-code

Algorithm 1 Sparse Group Lasso for Change Point detection

```
Require: $Y \in \mathbb{R}^{n \times p}$, $\lambda_1$, $\lambda_2$
Ensure: $\hat{\beta} \in \mathbb{R}^{(n-1) \times p}$
Start with $\hat{\beta} = \beta^0$;
Generate $X$;
for $k = 1$ to $K$ do
    for $i = 1$to $n - 1$ do
        Compute $J(t_i)$ in (16) ;
        if $J(t_i) \leq 1$ then
            $\hat{\beta}_{i,\bullet}^k = 0$
        else
            for $j = 1$ to $p$ do
                if $|X_{\bullet,i}^T r_{\bullet,j}^i| \leq \lambda_2$ then
                    $\hat{\beta}_{ij}^k = 0$
                else
                    solve the following objective function to get $\hat{\beta}_{ij}^k$:
                    $\min_{\beta_{ij}} \frac{1}{2}||r_{\bullet,j}^i - X_{\bullet,i}\beta_{ij}||_2^2 + \lambda_1||\beta_{ij}||_2 + \lambda_2|\beta_{ij}|;$
                end if
            end for
        end if
    end for
end for
```

V. Experiment

In this section we test the Sparse Group Lasso on simulated data sets. All experiments were written in Python.

A. Simulation

Each changepoint of each simulated profile were drawn from a Gaussian with mean 0 and variance 1 with n = 1000 (i.e. length is 1000), then we randomly choose some positions of changepoints. To consider the case of many shared changepoints, we simulated profiles as many as our machine can efficiently compute. Since we are considering sparsity of changepoints, we then added Gaussian noise with variance of 6 to those chosen position in a few profiles. Do 100 trials, check if the results are almost true changepoints, this can be considered as accuracy. Figure 1 is one of the simulated results. For profile 2 and profile 4, almost all the changepoints are marked.

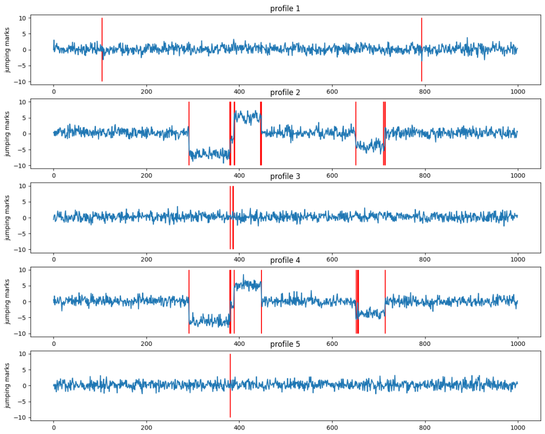

Figure 1. Simulated profiles with n = 1000, p = 300. We selected 5 profiles to observe. Blue lines indicate profiles, vertical red lines indicate changepoints

VI. Conclusion

We have used a piecewise constant model to describe the high-dimensional time series with sparse changepoints problem. In this model, we assume that both the changepoints and the individuals with changepoints are sparse. Also, the locations of changepoints are shared. We formulate the objective function by relaxing the constrains as l_2 norm and l_1 norm penalty terms into the basic minimizing error term $||Y - U||_F^2$. Then with piecewise constant model, we introduce parameter β_i which is the different between time $i+1$ and time i. Thus the original objective function is transformed into a new on for β. Then, we apply block coordinate gradient descent method to solve this sparse group lasso. The location and jump scale can be directly obtained from β.

References

[1] Bleakley, Kevin, and Jean-Philippe Vert. "The group fused lasso for multiple changepoint detection." arXiv preprint arXiv:1106.4199 (2011).

[2] Desobry, Frédéric, Manuel Davy, and Christian Doncarli. "An online kernel change detection algorithm." IEEE Transactions on Signal Processing 53.8 (2005): 2961-2974.

[3] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. "A note on the group lasso and a sparse group lasso." arXiv preprint arXiv:1001.0736 (2010).

[4] Harchaoui, Zad, and Céline Lévy-Leduc. "Multiple changepoint estimation with a total variation penalty." Journal of the American Statistical Association 105.492 (2010): 1480-1493.

[5] Harchaoui, Zaid, et al. "A regularized kernel-based approach to unsupervised audio segmentation." Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009.

[6] Tartakovsky, Alexander G., et al. "A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential changepoint detection methods." IEEE Transactions on Signal Processing 54.9 (2006): 3372-3382.

[7] Talih, Makram, and Nicolas Hengartner. "Structural learning with time-varying components: tracking the cross-section of financial time series." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67.3 (2005): 321-341.

[8] Picard, Franck, et al. "A statistical approach for array CGH data analysis." BMC bioinformatics 6.1 (2005): 27.

[9] Friedman, J., Hastie, T., Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:1001.0736.