On change points analysis based on resampling methods

Asanao Shimokawa*
Tokyo University of Science, Tokyo, Japan - shimokawa@rs.tus.ac.jp

Etsuo Miyaoka
Tokyo University of Science, Tokyo, Japan - miyaoka@rs.kagu.tus.ac.jp

Abstract

In this study, we focus on the construction method of the prediction model, and estimation methods of the change points for the generalized linear model with piecewise different coefficients. When there are multiple change points in data, the application of the hierarchical splitting (HS) algorithm can be considered to detect the location of change points. Although the algorithm is easy to execute and the computational efficiency is good, there is a risk that the estimated change points do not become the maximum likelihood estimators, and as a result its variance increases and it has no consistency and asymptotic normality. Moreover, if the locations of change points are estimated incorrectly, the prediction accuracy of the finally obtained model will rapidly decrease. To deal with this problem, we focused on the application of the bootstrap method based on the HS algorithm. In our approach, the prediction model is constructed by the bagging of the models obtained from resampling data and the HS algorithm. From the high diversity of the models obtained by the HS algorithm, it is expected that the prediction accuracy of the obtained model would be high. In addition to this, we study the two estimators of change points which obtained by bootstrap-based method. These approaches are compared to the ordinal HS algorithm through simulation studies. From the result, we confirmed the utility of the bootstrap-based methods for change point analysis. Especially, the prediction accuracy of the model obtained by the bagging algorithm was obviously higher than the model obtained by the HS algorithm. Moreover, the standard error of the estimator obtained by the proposed approach is smaller than that of the HS algorithm.

Keywords, Bagging; Ensemble method; Generalized linear model; Hierarchical splitting.

1. Introduction

Generalized linear models (GLM) are widely used to modeling an interesting response variable based on explanatory variables. In ordinal analysis based on a generalized linear model, the model is assumed to be hold for the whole data. However, it is widely understood that the assumption is not hold for several situations. For example in epidemiological studies in occupational medicine, there is often a threshold concentration of a specific agent which have an adverse health effect (Ulm (1991)). As an another example, in medical research, there is a possibility that the mortality rate for a certain disease changes suddenly with a certain threshold of age. To deal with these data, we can consider linear models, where the structure is changed at some points of an explanatory variable. These points are called as change points or break points. In this study, we focus on the construction method of the prediction model, and estimation methods of the change points for the GLM with piecewise different coefficients.

The change point analysis have been studied for a number of years. For example, Hawkins (1977), Worsley (1979), Inclán (1993), and Chen and Gupta (1997) studied the detection of change point locations in a sequence of random variables which follows normal distribution. Hawkins (1977) and Worsley (1979) described a method based on the likelihood procedure test. On the other hand, Inclán (1993) proposed a Bayesian based approach, and Chen and Gupta (1997) studied a Bayesian information criterion (BIC) based approach. If there are multiple change points, the grid search could be used on each method. However if the number of search points is large, this method is not practical from the viewpoint of computational complexity. To deal with this problem, the hierarchic splitting (HS) algorithm which dichotomize data recursively like classification and regression tree (Breiman et al. (1984)) is widely used (Chen and Gupta (2012)). The studies in change point analysis for a sequence of random variables are summarized in Csörgő and Horváth (1997), and Chen and Gupta (2012).

As a disadvantage of HS algorithm, the estimated location of change points are fixed until the end of algorithm. From this, an optimal combination of change points may not be found in some cases. As a result, there is a high risk that the variance of the estimator becomes large. Moreover, if the location of change points are estimated incorrectly, it is easily expected that the prediction accuracy of finally obtained model will be less. To deal with this problem, we consider to apply the HS algorithm with bootstrap method in this study. It is expected to decrease the variance of estimators of change points by aggregating the estimators obtained from each models which are given by resampling data and HS algorithm. Moreover, it is expected to increase the prediction accuracy by bagging the models obtained from each resampling data and HS algorithm.

2. Model

Let $\mathcal{L} = \{(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i); i = 1, 2, \cdots, n\}$ denote a set of observed learning samples, where y_i is the response, and $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \cdots, x_{iq-1})'$ and $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \cdots, z_{ip})'$ denote the vectors of independent explanatory variables. We assume that y_i comes from a distribution in the exponential family with a probability density function f where the dispersion parameter is ϕ. In addition to this, we assume that the n pairs $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ of observations are arranged in ascending order based on the continuous variable x_{i1} which considering change points.

We can consider piecewise different mean and variance structure by thinking about the GLM with piecewise different coefficients. Let $\mu_i = E(y_i)$ represents the mean of y_i, we consider the linear predictor model with $d-1$ change points:

$$\eta_i \equiv g(\mu_i) = \begin{cases} \boldsymbol{x}_i'\boldsymbol{\beta}_1 + \boldsymbol{z}_i'\boldsymbol{\alpha}, & \tau_0 < x_{i1} \leq \tau_1 \\ \boldsymbol{x}_i'\boldsymbol{\beta}_2 + \boldsymbol{z}_i'\boldsymbol{\alpha}, & \tau_1 < x_{i1} \leq \tau_2 \\ \quad \vdots & \\ \boldsymbol{x}_i'\boldsymbol{\beta}_d + \boldsymbol{z}_i'\boldsymbol{\alpha}, & \tau_{d-1} < x_{i1} \leq \tau_d \end{cases},$$

where $g(\cdot)$ represents the link function, and $\tau_0 = -\infty$ and $\tau_d = +\infty$ are assumed. $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \cdots, \beta_{q-1k})'$ represents the piecewise different coefficients vector $(k = 1, 2, \cdots, d)$, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_p)'$ represent the common coefficient vector in all segments. As a restriction for guarantee the estimable of coefficients, it is assumed that the number of learning samples include in each segment is larger than q. Then, our purpose is to estimate the location of change points $\tau_1, \tau_2, \cdots, \tau_{d-1}$, coefficients vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_d, \boldsymbol{\alpha}$, dispersion parameter ϕ if present, and the number of segments d if it is unknown case.

To estimate these parameters with the maximum likelihood method, the log likelihood of this model can be represented by the sum of the d log likelihoods using samples included in each segment.

$$l(\boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \phi | \boldsymbol{y}) = \sum_{k=1}^{d} \sum_{\substack{i \\ x_{i1} \in (\tau_{k-1}, \tau_k]}} \log f(y_i | \boldsymbol{\beta}_k, \boldsymbol{\alpha}, \phi),$$

where $\boldsymbol{\tau} = (\tau_1, \tau_2, \cdots, \tau_{d-1})$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_d)$. The maximum likelihood estimators (MLE) of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and ϕ are obviously depend on the unknown location of change points $\boldsymbol{\tau}$. If $\boldsymbol{\tau}$ is fixed, the MLE $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ can be obtained by general iterative methods like iterative weighted least squares. Moreover, the MLE $\hat{\phi}$ can be calculated with common method like the Pearson statistics by using the $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$.

Since $\boldsymbol{\tau}$ is unknown in actuality, some iterative search method is need that estimating $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and ϕ under possible combination of fixed $\boldsymbol{\tau}$. It seems to be intuitive to use the grid search over all possible $\boldsymbol{\tau}$, but there is a problem from the perspective of computational effort. That is, the order of a grid search for the known number of change point d is $O(n^d)$, and this method is not realistic when the number of samples n or segments d is large.

3. Hierarchical splitting

The HS algorithm is a repeated method like classification and regression tree. As first step of the algorithm, all learning samples $\mathcal{L}$ is split to two segments based on the splitting rule "$x_{i1} \leq \tau'$?". To determine the splitting rule "$x_{i1} \leq \tau'$?", that is to estimate the change point τ', all possible splits are evaluated and a split that maximizes the sum of the log likelihood of both segments is selected as the optimal one.

As second step, the next splitting rule "$x_{i1} \leq \tau''$?" is determined under the assumption that the splitting rule "$x_{i1} \leq \tau'$?" is remained. That is, all learning samples $\mathcal{L}$ is split to three segments based on the two splitting rule "$x_{i1} \leq \tau'$?" and "$x_{i1} \leq \tau''$?". The second rule "$x_{i1} \leq \tau''$?" is determined which maximizes the sum of the log likelihood of three segments from all possible splits.
By repeating this procedure, we construct the model with $d-1$ change points. The stopping rule for the algorithm is defined by the known number of change points or information criteria like Akaike information criterion (AIC) or BIC. The HS algorithm used in this study is described as follows:

1. The initial set of change points is given by $T_1 = \{-\infty, +\infty\}$.
2. **For** $k \leftarrow 2$ **to** the known number of segments d, or predefined max search number of segments d' **do**
3. Find the set of possible change points $T'_k = \{(T_{k-1}, \tau')\}$ which segments the data to k segments under the assumption that the splitting rule T_{k-1} is given.
4. Define the optimal set of k change points by

$$T_k = \arg \max_{(T_{k-1}, \tau') \in T'_k} l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\phi} | (T_{k-1}, \tau'), \boldsymbol{y}).$$

5. **end**
6. If the number of segments d is unknown case, the optimal set of change points is estimated by using AIC or BIC.
7. The estimated linear predictor model is given by

$$\hat{\eta}_i^{HS} = \sum_{k=1}^{d} I(\hat{\tau}_{k-1} < x_{i1} \leq \hat{\tau}_k) \boldsymbol{x}'_i \hat{\boldsymbol{\beta}}_k + \boldsymbol{z}'_i \hat{\boldsymbol{\alpha}}, \tag{1}$$

where $I(\cdot)$ represents the indicator function, and $(\hat{\tau}_0 = -\infty < \hat{\tau}_1 < \hat{\tau}_2 < \cdots, \hat{\tau}_{d-1} < \hat{\tau}_d = +\infty)$ are values obtained by rearranging the elements in the T_d in ascending order.

4. Bagging

To construct a model with better prediction accuracy, we will consider the use of the bagging algorithm (Breiman (1996)). The bagging algorithm is a representative method in the parallel ensemble methods which construct a set of base models and combine them. As stated in Zhou (2012), for regression problem, the degree of the improvement of the mean squared error by bagging depends on the instability of the base learnears. Since the variance of estimators of change points given by HS algorithm are expected to large, as discussed in above, the instability of the obtained model will also increase. Therefore, the bagging algorithm expected to work effectively in the construction of models which includes the change points.
The bagging algorithm used in this study is described as follows:

1. **For** $b \leftarrow 1$ **to** the predefined iterative number B **do**
2. Construct a set of bootstrap samples $\mathcal{L}^{(b)}$ by sampling with replacement from $\mathcal{L}$.
3. Estimate a linear predictor model $\hat{\eta}_i^{HS(b)}$ by using HS algorithm based on $\mathcal{L}^{(b)}$:

$$\hat{\eta}_i^{HS(b)} = \sum_{k=1}^{d^{(b)}} I(\hat{\tau}_{k-1}^{(b)} < x_{i1} \leq \hat{\tau}_k^{(b)}) \boldsymbol{x}'_i \hat{\boldsymbol{\beta}}_k^{(b)} + \boldsymbol{z}'_i \hat{\boldsymbol{\alpha}}^{(b)},$$

where $d^{(b)}$ is the known number of change points d, or if the number of change points is unknown case, it is estimated by using AIC or BIC based on $\mathcal{L}^{(b)}$.

4. **end**

5. The estimated linear predictor model is given by

$$\hat{\eta}_i^{Bag} = \frac{1}{B}\sum_{b=1}^{B}\left\{\sum_{k=1}^{d^{(b)}} I(\hat{\tau}_{k-1}^{(b)} < x_{i1} \leq \hat{\tau}_k^{(b)})\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_k^{(b)} + \boldsymbol{z}_i'\hat{\boldsymbol{\alpha}}^{(b)}\right\}. \tag{2}$$

Although it is possible to use the estimated linear predictor model (2) directly for prediction, it is difficult to interpret it. That is, in the model obtained by HS algorithm, the point estimates of location of change points are obviously $\hat{\tau}_1, \hat{\tau}_2, \cdots, \hat{\tau}_{d-1}$ in (1). On the other hand, the point estimates of the change points in (2) are not clear. To estimate the change points based on (2), we consider to use the mean values ($\hat{\tau}_k^*$) or medians ($\tilde{\tau}_k^*$) of change points for all base models when the number of change points is known.
If the bootstrap distribution used in bagging is mimicked the population distribution well, the estimator $\hat{\tau}_k^*$ is expected to follow a normal distribution asymptotically. On the other hand, $\tilde{\tau}_k^*$ is expected to be robust against extreme estimates. Because it is predicted that the variance of estimators of change points given by HS algorithm becomes large, there is a possibility of obtaining an extremely estimated value. $\tilde{\tau}_k^*$ is expected to deal with this problem. In order to clarify the notation, we express the estimate of τ_k which is given by HS algorithm on $\mathcal{L}$ as $\hat{\tau}_k^{HS}$.

5. Simulations

We present some simulation results to compare the HS algorithm and the algorithm based on bootstrap method described in previous sections. We compare the algorithms in terms of the prediction accuracy of the obtained model, and bias and variance of the estimator of the change point. We used the data generated from the Poisson regression model with two change points. The response are generated from the following model:

$$\eta_i = g(\mu_i) = \begin{cases} 1 + 0.4x_i + 0.2z_i, & -\infty < x_i \leq 3 \\ 3 - 0.3x_i + 0.2z_i, & 3 < x_i \leq 6 \\ 0.3x_i + 0.2z_i, & 6 < x_i \leq +\infty \end{cases},$$

where $g(\cdot)$ is the log link function. The explanatory variables x_i and z_i follow the uniform distribution between 0 and 10.
The number of learning samples n used are 100 and 300. We set the restriction of the number of learning samples included in a segment as 10 for efficient calculation. When the number of change points is unknown setting, we used the max search number of segments d' as 6. The iteration number of bootstrap B is set to 500. Simulations are repeated 300 times in every data group.
To compare the model accuracy for two algorithms, we used the mean square error of the prediction ($MSE(\hat{\mu})$) for test data. The number of test samples is set to 1000. The results of simulations are listed in Table 1. The table lists the average values and standard deviations in all simulations of $MSE(\hat{\mu})$ for both cases where the number of change points is known or unknown. As expected, the prediction accuracy of the model given by the bagging algorithm is better than that given by the HS algorithm for all simulation settings. The average values and the standard deviations of $MSE(\hat{\mu})$ for $\hat{\eta}^{Bag}$ is lower than $\hat{\eta}^{HS}$.
The model obtained by using BIC has greatest accuracy in the three patterns (d known, AIC, BIC). This result is somewhat strange, because the accuracy of the model obtained in the case where d is unknown is higher than the case where d is known. The reason for this, there is a possibility that the estimation of the change points is largely incorrect for $\hat{\eta}^{HS}$. This will be discussed in the next simulation result.
For $\hat{\eta}^{Bag}$, this result seems to be due to the diversity of the base models. That is, when the number of change point is known case, the base models which construct the model (2) have the same number of segments. On the other hand, when the number of change point is unknown case, the base models have several number of segments. As the result, the diversity of the base models included in the estimated model when d is unknown cases is higher than in known cases.
To compare the three change point estimators $\hat{\tau}_k^{HS}$, $\hat{\tau}_k^*$, and $\tilde{\tau}_k^*$, we used the average values and standard deviations in all simulations of the estimators for the number of change points is known case. The results are shown in Table 2.

Table 1: The simulation results of comparison of model prediction accuracy for HS algorithm and bagging algorithm when the true model contains change points. The values in the table represent the average values of $MSE(\hat{\mu})$ in 300 simulations. The value in parentheses represents the standard deviations of $MSE(\hat{\mu})$ in the simulations.

n	algorithm	d known	d unknown	
			AIC	BIC
100	$\hat{\eta}^{HS}$	9.17	7.58	7.14
		(3.65)	(4.28)	(3.78)
	$\hat{\eta}^{Bag}$	5.56	5.65	5.00
		(2.15)	(2.50)	(2.17)
300	$\hat{\eta}^{HS}$	6.54	3.25	1.92
		(1.81)	(1.34)	(1.06)
	$\hat{\eta}^{Bag}$	3.61	2.12	1.63
		(1.05)	(0.77)	(0.68)

Table 2: The simulation results of comparison of three change point estimators. The values in the table represent the average values of the estimators in 300 simulations. The value in parentheses represents the standard deviations of the estimators in the simulations.

n	estimator	$\tau_1 = 3$		$\tau_2 = 6$	
100	$\hat{\tau}_k^{HS}$	3.29	(0.78)	5.15	(0.84)
	$\hat{\tau}_k^{*}$	3.24	(0.39)	5.31	(0.41)
	$\tilde{\tau}_k^{*}$	3.33	(0.56)	5.24	(0.68)
300	$\hat{\tau}_k^{HS}$	3.41	(0.63)	5.13	(0.82)
	$\hat{\tau}_k^{*}$	3.33	(0.30)	5.13	(0.37)
	$\tilde{\tau}_k^{*}$	3.38	(0.46)	5.12	(0.71)

The average values of estimates for $\hat{\tau}_k^{HS}$, $\hat{\tau}_k^{*}$, and $\tilde{\tau}_k^{*}$ is slightly biased. Especially for τ_2, the differences between the average values of estimates and the true value are about 0.8 for all estimators. The standard deviation of $\hat{\tau}_k^{*}$ is the smallest, then $\tilde{\tau}_k^{*}$ is the middle, and $\hat{\tau}_k^{HS}$ is the largest. The difference of the standard deviations between $\hat{\tau}_k^{*}$ and $\hat{\tau}_k^{HS}$ is almost twice for all patterns.
As a result of the whole simulations, depending on the model, there are cases where the obtained estimates of change point have an obviously bias. The bias tends to be pulled in the direction of another true change point. The empirical distribution of $\hat{\tau}_k^{*}$ tends to have the unimodal distribution with small standard deviation. On the other hand, the empirical distributions of $\hat{\tau}_k^{HS}$ and $\tilde{\tau}_k^{*}$ have the same shape, but the standard deviation of $\tilde{\tau}_k^{*}$ tends to be smaller than $\hat{\tau}_k^{HS}$. The histogram of the empirical distribution and more details of the results will be announced at the conference.

6. Conclusions

As a standard approach for multiple change points analysis, the application of the HS algorithm is widely used. The algorithm is easy to execute and the computational efficiency is good. However, there is the risk that the estimated change points by the algorithm does not become the MLEs, and as a result its variance increases and it has no consistency and asymptotic normality. To deal with this problem, we focused on the application of bootstrap method based on the HS algorithm in GLM with piecewise different coefficients. Especially, we studied the convenience of the method from the two viewpoints: improvement of the prediction accuracy by bagging, and reduction of the standard error of the estimator of the change point.
As the first main result, the prediction accuracy of the model obtained by bagging algorithm is almost certainly higher than the model obtained by HS algorithm. As a little surprising result, the model obtained by bagging algorithm when the number of change points is estimated in each base model is more accurate

than the model obtained when the number of change points is known case. The reason for this is considered to be due to the diversity of the basic models, and therefore application of further development of algorithm such as VR-Tree ensemble (Liu et al. (2008)) can be considered.
Second, there is little difference between the average values of estimators of the change points obtained by HS algorithm and bootstrap method. Depending on the setting of the true model, both estimators have bias, but the standard error of the estimator obtained by bootstrap method is smaller than it by HS algorithm.

References

Ulm K. (1991). A statistical method for assessing a threshold in epidemiological studies, Statistics in Medicine. **10**, 341-349.

Hawkins D. M. (1977). Testing a sequence of observations for a shift in location, Journal of the American Statistical Association. **72**, 180-186.

Worsley K. J. (1979). On the likelihood ratio test for a shift in location of normal populations, Journal of the American Statistical Association. **74**, 365-367.

Inclán C. (1993). Detection of multiple changes of variance using posterior odds, Journal of Business and Economics Statistics. **11**, 289-300.

Chen J., & Gupta A. K. (1997). Testing and locating variance changepoints with application to stock prices, Journal of the American Statistical Association. **92**, 739-747.

Breiman L., Friedman J. H., Olshen R. A., & Stone C. (1984). Classification and Regression Trees. Wadsworth, California.

Chen J., & Gupta A. K. (2012). Parametric Statistical Change Point Analysis, 2nd Edition. Birkhäuser, New York.

Csörgő M., & Horváth L. (1997). Limit Theorems in Change-Point Analysis. John Wiley & Sons, New York.

Akaike H. (1973). Information theory and an extension of the maximum likelihood principle, in proceedings of the 2nd International Symposium on Information Theory, Petrov B. N., and Csáki F. (Eds.), Budapest, 267-281.

Breiman L. (1996). Bagging predictors, Machine Learning. **24**, 123-140.

Zhou Z. H. (2012). Ensemble Methods Foundations and Algorithms. Chapman and Hall/CRC Press, Boca Raton, Florida.

Liu, F. T., Ting. K. M., Yu, Y., & Zhou, Z. H. (2008). Spectrum of variable-random trees., Journal of Artifical Intelligence Research. **32**, 355-384.