Tracking Concept Drift Using a Constrained Penalized Regression Combiner

Li-Yu Wang*, Cheolwoo Park*, Kyupil Yeon**, and, Hosik Choi***

*Department of Statistics, University of Georgia, **Department of Applied Statistics, Hoseo University, ***Department of Applied Information Statistics, Kyonggi University

Abstract

The objective of this work is to develop a predictive model when data batches are collected in a sequential manner. With streaming data, information is constantly being updated and a major statistical challenge for these types of data is that the underlying distribution and the true input-output dependency might change over time, a phenomenon known as concept drift. The concept drift phenomenon makes the learning process complicated because a predictive model constructed on the past data is no longer consistent with new examples. In order to effectively track concept drift, we propose new novel model-combining methods using constrained and penalized regression that possesses a grouping property. The new learning methods enable us to select data batches as a group that are relevant to the current one, reduce the effects of irrelevant batches, and adaptively reflect the degree of concept drift emerging in data streams. We study theoretical properties of the proposed methods and finite sample performance using simulated and real examples. The analytical and empirical results indicate that the proposed methods can effectively adapt to various types of concept drift and show superior performance over existing methods.

Concept Drift

- Consider a sequential regression problem in a data stream which is a series of data batches entering over time continuously
- Data batch has observations, i.e. input-output pairs, and is used for constructing a learning model for predicting the future observations
- Main concern: *concept drift*
  - Any changes of circumstance in a learning process
  - Examples: subprime mortgage crisis, spam classification, and Airplane Delay

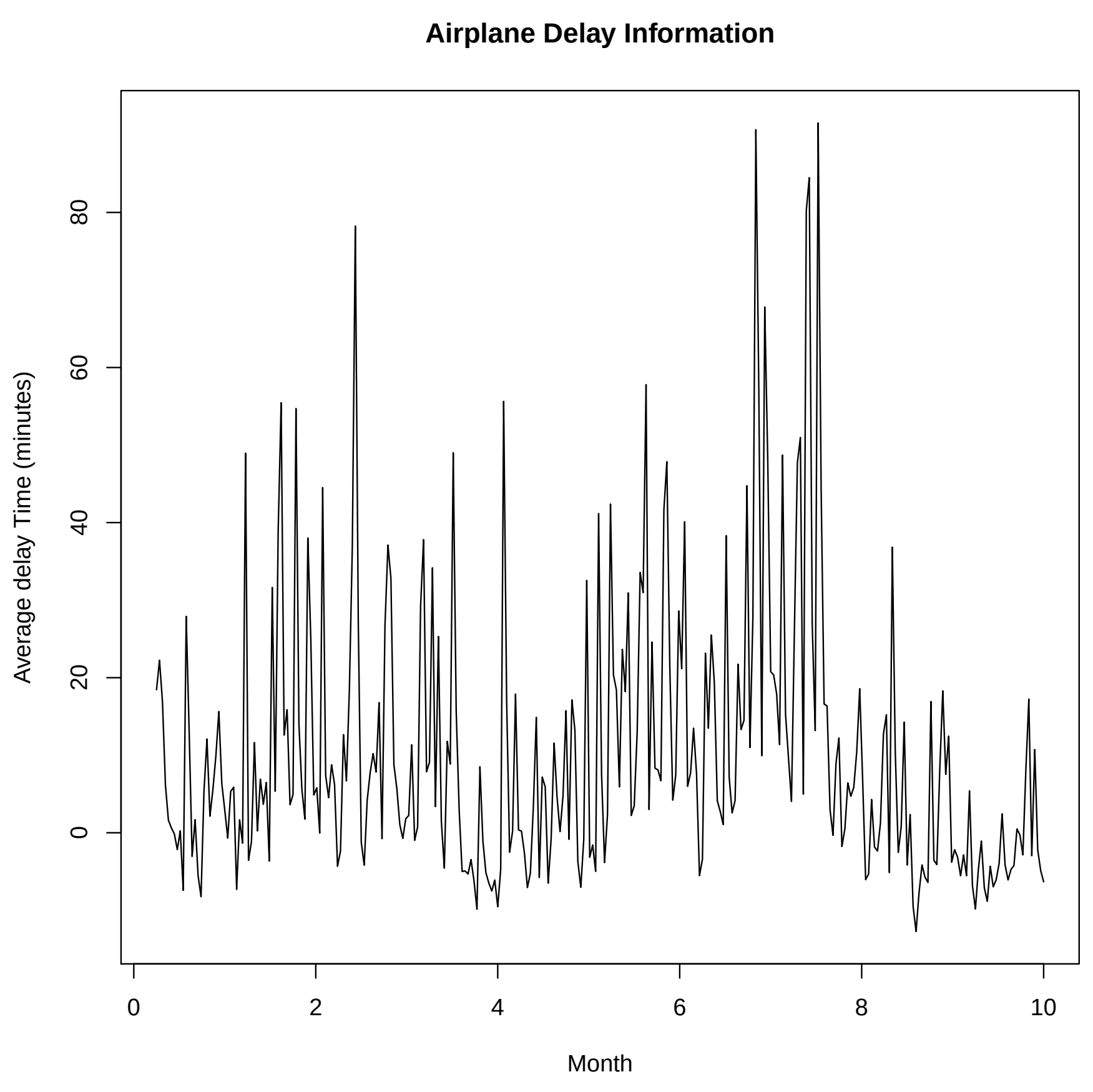


Figure : Abrupt changes in airplane delay minutes

Setting

- Let $\{D_m,\ m = 1, 2, \ldots, M\}$ be a sequence of data batches which consists of input-output pairs
- Suppose observations in D_m are random samples from unknown distributions $F_m(\mathbf{x}, y)$
  - $\mathbf{x} \in \mathbb{R}^p$ is a predictor vector
  - $y \in \mathbb{R}$ is a response
  - $\mathbf{x}$ and y arise from a statistical model

$$y = f_m(\mathbf{x}) + \epsilon,$$

where ϵ is a random error satisfying $E(\epsilon) = 0$ and is independent of $\mathbf{x}$

- the underlying distribution can be different for each batch
- the relationship between response and predictors can change over time

- Prediction obtained from the outputs of base models and the most recent data batch $D_M = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ that best reflects the target concept to be tracked
- For each $\mathbf{x}_i,\ i = 1, \ldots, n$, denote the outputs of base models by $\hat{\mathbf{f}}(\mathbf{x}_i) = (\hat{f}_1(\mathbf{x}_i), \ldots, \hat{f}_{M-1}(\mathbf{x}_i), \hat{f}_M^{(-i)}(\mathbf{x}_i))^\intercal$
- $\hat{f}_M^{(-i)}(\mathbf{x}_i)$ represents a leave-one-out estimate of $\hat{f}_M(\mathbf{x}_i)$ to avoid over-fitting

Data batches	D_1	$\cdots$	D_{M-1}	D_M
Input-output pairs	$(\mathbf{x}_1, y_1)$ $\vdots$ $(\mathbf{x}_n, y_n)$	$\cdots$ $\vdots$ $\cdots$	$(\mathbf{x}_1, y_1)$ $\vdots$ $(\mathbf{x}_n, y_n)$	$(\mathbf{x}_1, y_1)$ $\vdots$ $(\mathbf{x}_n, y_n)$
Base Models	$\hat{f}_1(\mathbf{x}_i)$	$\cdots$	$f_{M-1}(\mathbf{x}_i)$	$\hat{f}_M^{(-i)}(\mathbf{x}_i)$

Methodology

- the design matrix and the response for the model aggregation

$$\mathbf{X} = \begin{pmatrix} \hat{f}_1(\mathbf{x}_1) & \cdots & \hat{f}_{M-1}(\mathbf{x}_1) & \hat{f}_M^{(-1)}(\mathbf{x}_1) \\ \hat{f}_1(\mathbf{x}_2) & \cdots & \hat{f}_{M-1}(\mathbf{x}_2) & \hat{f}_M^{(-2)}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{f}_1(\mathbf{x}_n) & \cdots & \hat{f}_{M-1}(\mathbf{x}_n) & \hat{f}_M^{(-n)}(\mathbf{x}_n) \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- Final predictive model has the form of $f(\mathbf{x}) = \Sigma_{m=1}^M w_m \hat{f}_m(\mathbf{x})$
- Goal: optimally determine the weights $\mathbf{w} = (w_1, \cdots, w_M)$ using $\hat{\mathbf{f}}(\mathbf{x}_i)$

- Proposed method: A constrained penalized aggregation method minimizes the following objective function with respect to w_m for $m = 1, \ldots, M$:

$$R_\gamma^{\mathbf{D}}(\mathbf{w}) = \sum_{i=1}^n \Big(y_i - \sum_{m=1}^M w_m \hat{f}_{im}\Big)^2 + \lambda \|\mathbf{D}\mathbf{w}\|_\gamma^\gamma,$$
$$\text{such that } \sum_{m=1}^M w_m = 1, w_m \geq 0 \qquad (\star)$$

where
- $\|\mathbf{w}\|_\gamma = (\Sigma_{m=1}^M |w_m|^\gamma)^{1/\gamma}$ with $\gamma = 1$ or 2
- $\mathbf{D}$ is a identity matrix or $M' \times M$ pairwise difference matrix where $M' = \binom{M}{2}$
- λ is a tuning parameter
- $\hat{f}_{im}$ is the mth element of the vector $\hat{\mathbf{f}}(\mathbf{x}_i)$
- $\mathbf{D} = I$ and $\gamma = 2$: MC.Ridge1+ (Yeon et al, 2010)
- $\mathbf{D}$ pairwise and $\gamma = 1$: MC.Lasso1+

Important Result

The proposed combiner holds grouping property by assigning the same weights to identical concepts and produces better prediction when there is concept drift by assigning higher weights on relevant batches.

Definition (1)

(A measure of concept drift) For $\mathbf{w}^* = (1/M, \ldots, 1/M)^\intercal$ and $\hat{\mathbf{w}}$ obtained by a combiner, we define the degree of concept drift as the angle between the two vectors given by

$$\eta(\mathbf{w}^*, \hat{\mathbf{w}}) = \cos^{-1}\left(\frac{|\langle \mathbf{w}^*, \hat{\mathbf{w}} \rangle|}{\|\mathbf{w}^*\|_2 \|\hat{\mathbf{w}}\|_2}\right)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors in R^M.

- $\eta(\mathbf{w}^*, \hat{\mathbf{w}})$ is small when there is no or gradual concept drift
- $\eta(\mathbf{w}^*, \hat{\mathbf{w}})$ gets larger when there is considerable concept drift

Theorem (1)

When concept drift is not present, the minimizer of $(\star)$ is given as $\hat{\mathbf{w}} = (1/M, \ldots, 1/M)^\intercal$ for sufficiently large λ.

Theorem (2)

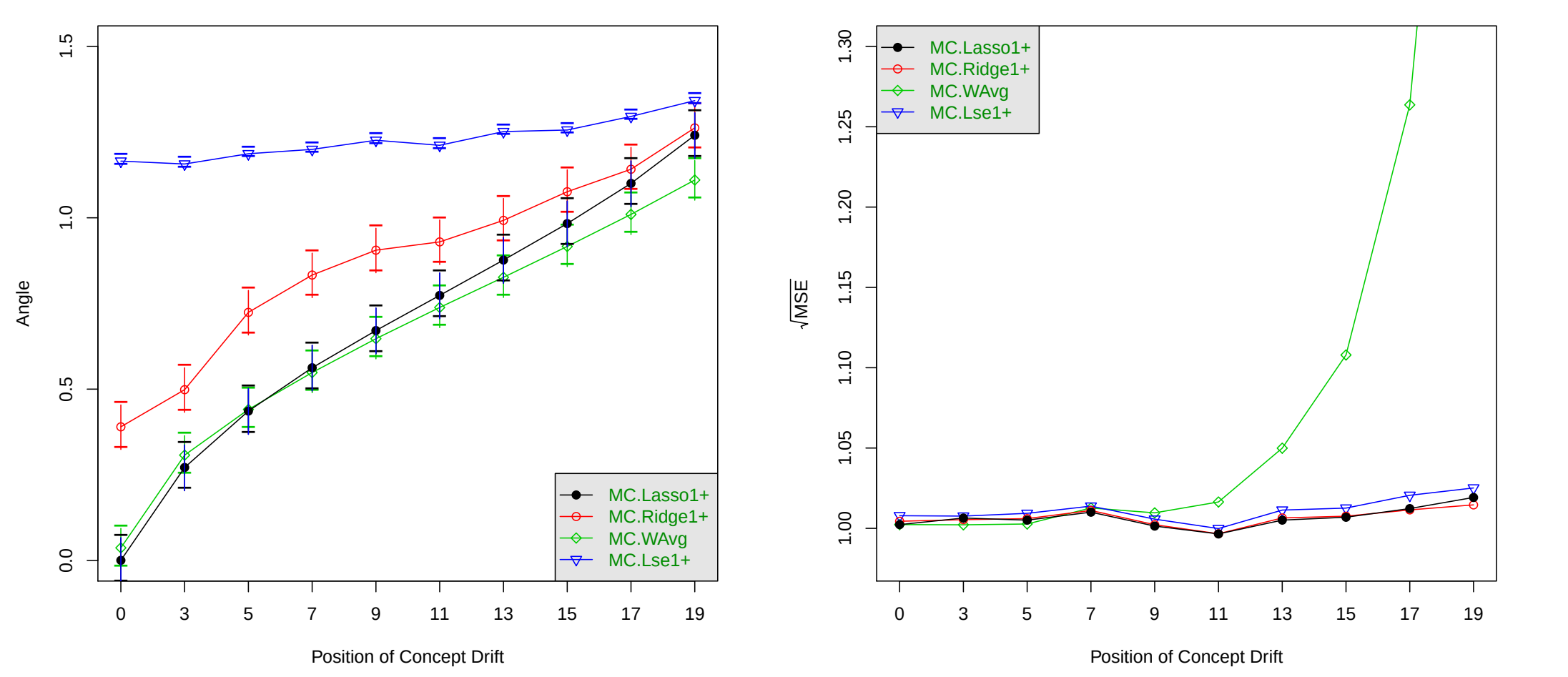
For $\hat{\mathbf{w}}$ obtained from $(\star)$, $\eta(\mathbf{w}^*, \hat{\mathbf{w}})$ in Definition 1 is monotonically decreasing as λ increases.

Simulation

- 20 training sequential data batches $D_1, \ldots, D_{20}$
- Each D_m contains 200 observations independently from $y = \Sigma_{i=1}^{10} \alpha_i X_i + \epsilon$. y is a response variable. $(X_1, \ldots, X_{10})$ are independent predictors from Uniform(0,1). The noise variable ϵ is independently from standard normal distribution
- The concept drift is incorporated by changing the value of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{10})^\intercal$ from a certain time $T = 0, 3, 5, 7, 9, 11, 13, 15, 17, 19,$
  - For example, if $T = 10$, each D_m for $m = 1, \ldots, 9$ is generated with $\boldsymbol{\alpha} = (5, 5, 5, 5, 5, 0, 0, 0, 0, 0)^\intercal$ and the remaining data batches are with $\boldsymbol{\alpha} = (0, 0, 0, 0, 0, 5, 5, 5, 5, 5)^\intercal$
- Test data batch: same data generation process as the last batch D_{20} with $n = 1000$ independent observations

- Tuning parameter λ is selected by 10-fold cross-validation from $\log_{10}(0.5 \times \lambda) = c$ where c varies from -3 to 3 by 0.2 selected based on the lowest RMSE
- Compared with methods
  - MC.WAvg: Weighted average method (Wang et al. 2003)
  - MC.Lse1+ : Least squares method with constraints (Chu et al. 2004)
  - MC.Ridge1+ Ridge regression method (Yeon et al. 2010)

Simulation Results



(a) Angle (b) RMSE

Figure : The left panel (a) shows the angles with standard errors for four methods at each position of concept drift and the right panel. The angle increases as the concept drift point gets closer to current time point (b) shows the RMSE at each position of concept drift.

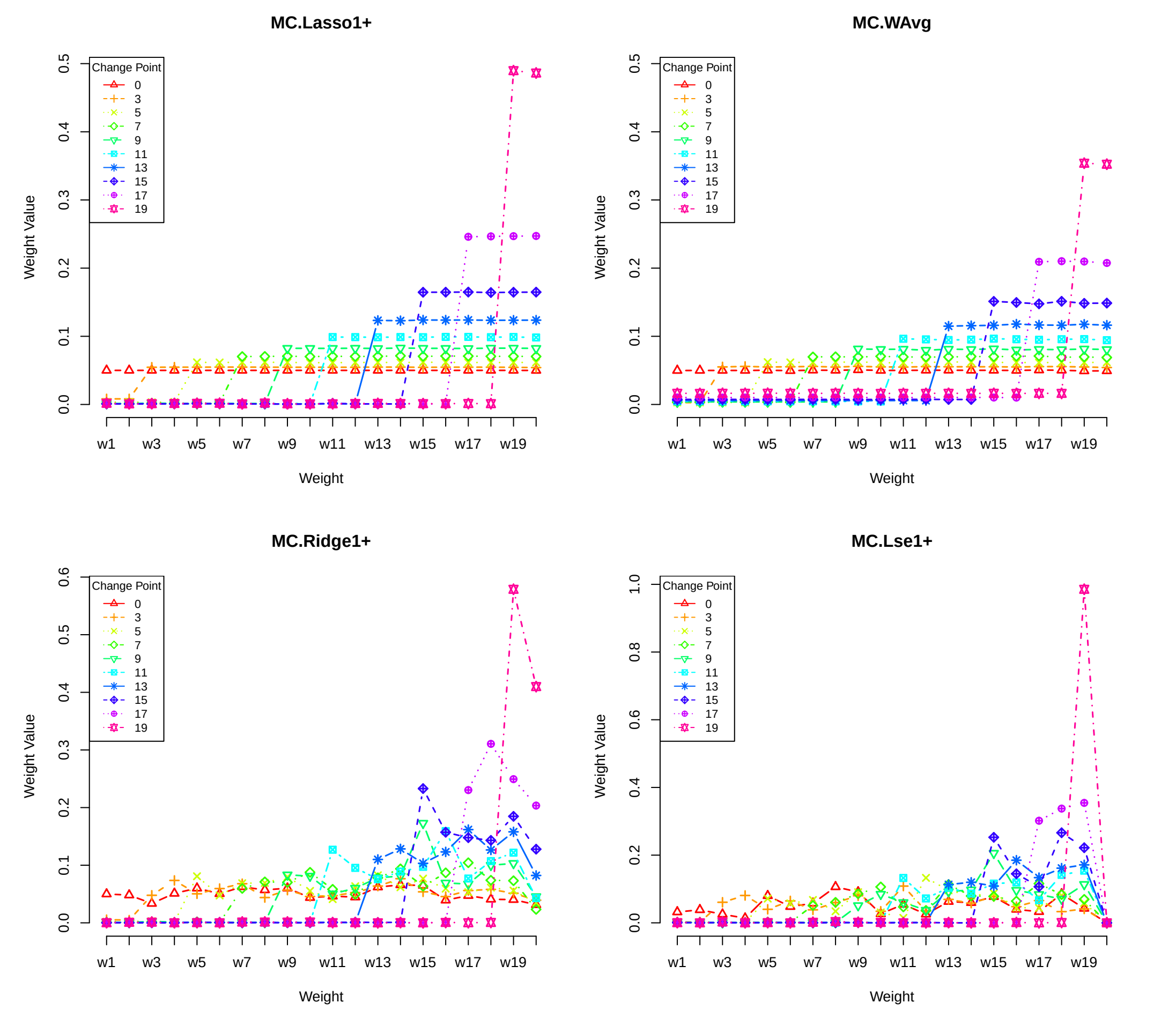


Figure : Estimated weights at each change point for four methods. MC.Lasso1+ shows the grouping property by assigning positive weights to relavent concepts.

References

- Chu, F., Wang, Y., and Zaniolo, C. (2004). Mining noisy data streams via a discriminative model. Discovery Science, pages 47-59
- Wang, H., Fan, W., Yu, P., and Han, J. (2003). Mining concept-drifting data streams usingensemble classifiers. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Knowledge discovery and data mining; KDD-2003, pages 226 - 235
- Yeon, K., Song, M. S., Kim, Y., Choi, H., and Park, C. (2010). Model averaging via penalizedregression for tracking concept drift.Journal of Computational & Graphical Statistics, 19(2):457 - 473