Cluster-wise regression models via a quasi-linear function

Keio University (Japan) Kenichi Hayashi
The institute of Statistical Mathematics (Japan) Shinto Eguchi

Suppose that there are multiple heterogeneous subgroups in a dataset. This would be a natural assumption for many fields of application in the "Big data" era such as biology, marketing, psychology, etc. Consider a dataset $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i); i = 1, \dots, n\}$ and a regression of y_i's on the space of $\boldsymbol{x}_i$'s. Assume that there are several heterogeneous groups in $\mathcal{D}_n$ and the regressors vary among them. Then, conventional linear regression models cannot approximate the relationship between the response and features. These result in not only poor prediction performance but also misleading interpretation of analyses. Cluster-wise linear regression models, involving information on clusters of each cases, have been applied in various contexts (e.g., O'Driscoll et al., 2012; Suk et al., 2014).

In this study, we propose a novel extension of cluster-wise regression models $\mathrm{E}\left[Y|\boldsymbol{x}\right] = g^{-1}(\mu(\boldsymbol{x};\boldsymbol{\theta}))$, where g is the link function and $\mu(\boldsymbol{x};\boldsymbol{\theta})$ is a combination of regressors for each cluster written as

$$\mu(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{\tau}\log\left(\sum_{k=1}^{K} p_k(\boldsymbol{x})\exp\left(\tau\mu_k(\boldsymbol{x})\right)\right). \tag{1}$$

Here, $\tau \in \mathbb{R}$ is a pre-specified parameter, $\mu_k(\boldsymbol{x})$ is a regressor for the kth cluster, and $\boldsymbol{\theta}$ is a parameter vector that consists of the parameters of μ_k's. We usually consider the linear regression function $\mu_k(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{\beta}_k$, hence $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_K^\top)^\top$. $p_k(\boldsymbol{x})$ in (1) is the point association probability indicating the probability that point $\boldsymbol{x}$ belongs to the kth cluster. Since the distribution of features for each cluster is often unknown and different, we need to detect clusters without any assumptions about them. For this purpose, we propose the application of a clustering method by Rose et al. (1990) to estimate the point association probabilities by minimizing the free energy $-\frac{1}{\omega}\sum_{i=1}^{n}\log\left(\sum_{k=1}^{K}\exp(-\omega||\boldsymbol{x}_i - \boldsymbol{c}_k||^2)\right)$, where $\omega > 0$ and $\boldsymbol{c}_k$ is the centroid of the kth cluster.

When $\tau \to 0$, the model (1) reduces to the weighted mean of K regressors: $\mu(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k(\boldsymbol{x})\mu_k(\boldsymbol{x})$. On the other hand, when $\tau \to \infty$, the (1) yields the maximum score $\max_{k=1,\dots,K}\mu_k(\boldsymbol{x})$. In general, the value of the association probabilities ranges in the interval $[0, 1]$. As a special case, however, they take only one or zero when $\omega \to \infty$, implying hard clustering. Hence, the proposed model includes the method proposed in DeSarbo et al. (1989).

The algorithm and some numerical examples are given to show how well the proposed method works.

References

[1] DeSarbo, W.S., et al. (1989). *Psychometrika*, **54**, 707–736.
[2] O'Driscoll, D.M., et al. (2012). *Sleep*, **35**, 1269–1275.
[3] Rose K., et al. (1990). *Physical Revidw Letters*, **65**, 945–948.
[4] Suk, H., et al. (2014). *PLoS ONE*, **9** (2): e87056.