Task 3 : Bandwidth choice function

Martin Guy and Hannes Leskelä

2 November 2016

Introduction

The goal of this exercise is to implement bandwidth choice functions using predictive mean square error (PMSE) and cross-validation (CV). We had to implement **Leave-one-out CV** (LOOCV), **5-fold CV**, **10-fold CV** and **Generalized CV** (GCV). We used the `loess` estimator for **5CV** and **10CV**, and `locpolreg` for **LOOCV** and **GCV**. The reason for choosing locpolreg was to use the smoothing matrix S to avoid the computational cost of fitting N different nonparametric regressions.

We will then try our functions on the `Boston Housing` dataset and compare the results with `h.select` (package **sm**) and `dpill` (package **KernSmooth**).

Methods

LOOCV and GCV

In order to reduce compational cost, we get the smoothing matrix S from the `locpolreg` function. We can compute the cost for LOOCV this way:

$$PMSE_{CV} = \frac{1}{N}\sum_{i=1}^{N}(\frac{y_i - \hat{y}_i}{1 - s_{ii}})^2$$

This is much more faster than fitting N different nonparametric regressions.

For generalized CV, we have this formula:

$$PMSE_{CV} = \frac{1}{N}\sum_{i=1}^{N}(\frac{y_i - \hat{y}_i}{1 - \mu})^2$$

where $\mu = \frac{1}{N}\sum_{i=1}^{N}(s_{ii})$ is the mean of the trace of the smoothing matrix S.

5-fold and 10-fold CV

Here, we just created a function making the k-fold CV, then apply it for $k = 5$ and $k = 10$.

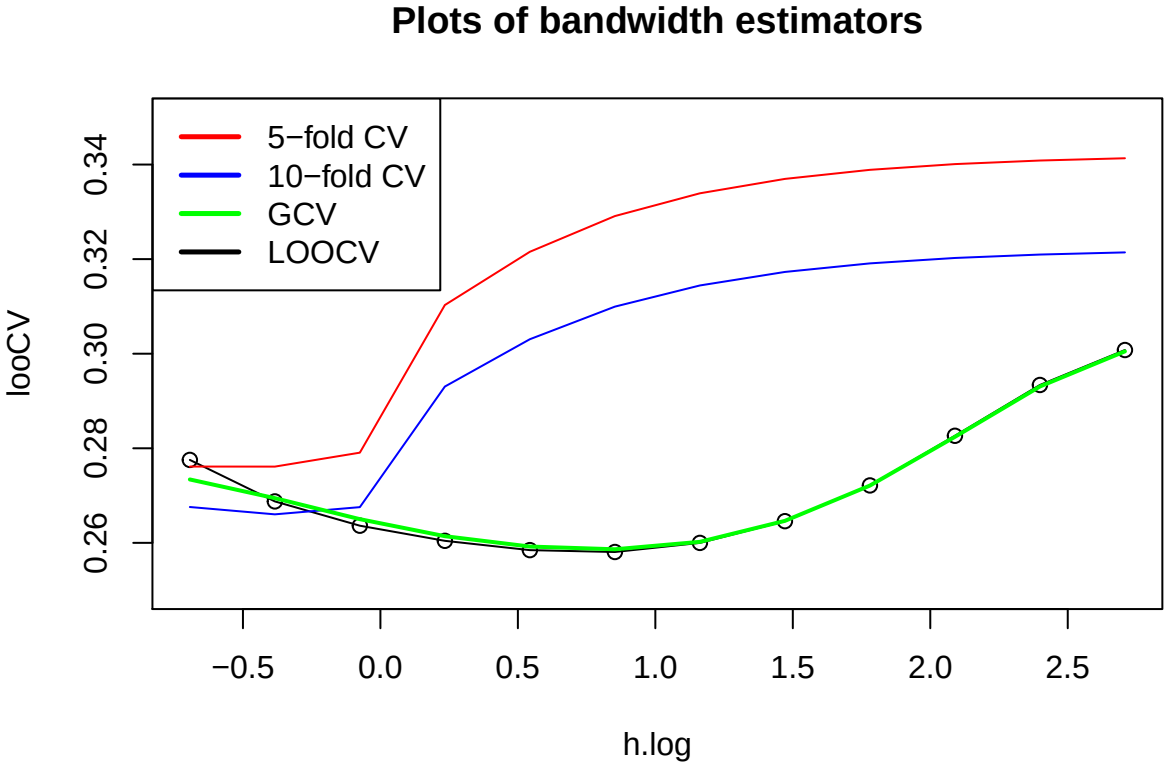
Regarding how to generate the k-folds, we are currently using an in-sequence approach where we leave out a segment of the data instead of splitting it pseudo-randomly. This can affect the result of the estimated "best" bandwidth a lot since a bigger sequential gap might lead to a worse estimate for the section that is left out. A better way would be to randomly sample the folds, but when we tried to do so the method failed and we had to resort to this subpar method.

Results and Discussion

The different cross-validations means that we will sacrifice a lot of our training data in favor of the validation set, which implies that the **5-** and **10-fold CV** might be underfitting quite a lot compared to **N-fold**. However, the **N-fold** is quite costly computational wise, and might overfit. This can be seen here, where the error keeps increasing with increasing sizes in the validation set:

Method	best h
h.select	1.986876
dpill	1.499037
Leave-one-out CV	1.986876
5-folds CV	1.499037
10-folds CV	1.986876
Generalized CV	1.499037

Table 1: Results of the bandwidth choice for different methods



However, the general form of the logplots are consistent, decreasing initially until we find an optimal value for the bandwith, then increasing as we move to bigger values. This is similar to when one increase the regularization of a model, thus introducing more and more bias, and gives a visual confirmation that the bandwith is also controlling the bias-variance ratio.

Conclusion

If we do a majority vote, both 1.499 and 1.986 are "good" values for the bandwidth. We could make a better estimation of the bandwidth now by taking new h candidates between these two values.

Having validation data is good to estimate the bandwidth choice for our estimators. However, when we have not so much datas, we need to seperate a part of our data for validation. One can think this changes the estimation of the bandwidth, but usually, real datas are big datas and there is no need of considering this problem.