Sliced Regression in R

Hang Weiqiang

January 29, 2017

1 Introduction

Dimension reduction is very efficient in high dimensional data processing. It is a method of reducing redundant information of original data. In most cases, it replace original covariate variables x with a few linear combinations $\beta_1^\top x, \cdots, \beta_d^\top x$, in which all the information contained in x for response variable y or $E(y|x)$ is retained. In dimension reduction, we have a response variable $y \in \mathbb{R}^1$, and p-dimensional covariate variable $x \in \mathbb{R}^p$. Let $\mathbf{B} \in \mathbb{R}^{p\times d}$, and $\mathcal{S}(\mathbf{B})$ be the linear space spanned by the column vector of $\mathbf{B}$. We call $\mathcal{S}(\mathbf{B})$ is central subspace[3] if

$$y \perp\!\!\!\perp x | \mathbf{B}^\top x.$$

if the intersection of all central subspace is still a central subspace, then it is called central space(CS), which is denoted by $\mathcal{S}_{y|x}$[3]. We assume that central space exists with the basis of $\mathbf{B}_0 \in \mathbb{R}^{p\times d_0}$ for some $0 < d_0 < p$. Then, all the information contained in x about y is included in $\mathbf{B}_0^\top x$.

In some cases, such as nonparametric model, conditional mean $E(y|x)$ is more of interest. We want to find all information x contained in $E(y|x)$. Following the definition in [4], we call $\mathcal{S}(\mathbf{A})$ is central mean subspace if

$$y \perp\!\!\!\perp E(y|x) | \mathbf{A}^\top x.$$

Similarly, if the intersection of all central mean subspace is still a central mean subspace, central mean space(CMS) can be defined and denoted by $\mathcal{S}_{E(y|x)}$. Further discussion about the uniqueness and existence of CMS can be found in [4]. As can be seen, $\mathcal{S}_{E(y|x)}$ should be subset of or equal to $\mathcal{S}_{y|x}$.

In this paper, we will give a short description of the method used in the package MAVE.

2 Central mean space estimation

2.1 The Initial estimate

When estimating central mean space, we follow the model in [6], assuming that

$$y = g(\mathbf{A}_0^\top x) + \varepsilon$$

where g is an unknown smooth link function, $\mathbf{A}_0$ is a $p \times d_0$ orthogonal matrix, namely $\mathbf{A}^\top \mathbf{A} = I_{d_0}$ for some $d_0 < p$ and $E(\varepsilon|x) = 0$. MAVE and OPG methods are proposed to find $\mathbf{A}_0$ in[6]. In package MAVE, OPG and MAVE method are implemented, but a little different from the original. The difference is to make the algorithm of find CMS can be fused with that of estimating CS, so that the code will be easier to develop. The main approach to estimate central mean space is by estimating the derivative of conditional expectation $E(y|x)$, which is given by

$$\frac{\partial E(y|x)}{\partial x} = \frac{\partial g(\mathbf{A}_0^\top x)}{\partial x} = \mathbf{A}_0 \nabla g(\mathbf{A}_0^\top x).$$

Then if $E[\nabla g(\mathbf{A}_0^\top x) \nabla g^\top (\mathbf{A}_0^\top x)]$ is of full rank, then $\mathcal{S}_{E(y|x)}$ can be estimated completely by d_0 eigenvectors of $E\left[\partial E(y|x)/\partial x \left(\partial E(y|x)/\partial x\right)^\top\right]$.

In order to estimate the derivative of conditional expectation, local least squared method is used[5]. Let X be the $n \times p$ design matrix with X_i be the ith random sample, Y be $n \times 1$ response matrix and Y_i is the ith response data. The value of $(E(y|x), \partial E(y|x)/\partial x)$ at X_i can be estimated by $(\hat{a}_i, \hat{b}_i)(\hat{a}_i \in \mathbb{R}^1, \hat{b}_i \in \mathbb{R}^p)$, , which is obtained by minimizing the following least squared functions,

$$n^{-1} \sum_{i=1}^{n} \{Y_i - a_i - b_i^\top X_i\}^2 K_{h_0}(X_{ij})$$

where $X_{ij} = X_i - X_j$ and $K_{h_0}(\cdot)$ is kernel function with bandwidth h_0. Further discussion on the kernel function and the selection of bandwidth can be found in [5, 6, 7]. Then we construct the following matrix to recover $\mathbf{A}_0$

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} \hat{b}_i \hat{b}_i^\top.$$

Then the basis of $\mathcal{S}_{E(y|x)}$ can be estimated by the largest d_0 eigenvectors of $\hat{\Sigma}$.

2.2 The refined estimate

The estimation of $\hat{A}$ can be further refined following the idea of MAVE[6]. The key is updating the kernel weight for every iteration. In OPG method, using the estimate from the eigenvectors of $\hat{\Sigma}$ as the initial estimate, for every iteration, given $\mathbf{A}_{(t)}$, the next estimate $(\hat{a}_i^{(t+1)}, \hat{b}_i^{(t+1)})$ can be obtained by minimizing

$$n^{-1}\sum_{i=1}^{n}\{Y_i - a_i - b_i\}^2 K_{h(t)}(\mathbf{A}_{(t)}^\top X_{ij}).$$

The next $\mathbf{A}_{(t+1)}$ can be estimated by the d_0 largest eigenvectors of $\hat{\Sigma}_{(t+1)}$, which is given by

$$\hat{\Sigma}_{t+1} = \sum_{i=1}^{n} \hat{b}_i \hat{b}_i^\top.$$

For MAVE method, given $\mathbf{A}_{(t)}$, the next estiamte $\mathbf{A}_{(t+1)}$ is obtained by minimizing

$$n^{-1}\sum_{i=1}^{n}\{Y_i - a_i - d_i^\top \mathbf{A}^\top X_{ij}\}^2 K_{h(t)}(A_{(t)}^\top X_{ij})$$

where $\mathbf{A} \in \mathbb{R}^{p\times d_0}$ with $\mathbf{A}^\top\mathbf{A} = I_{d_0}$. The difference between OPG and MAVE is MAVE restrict b_i in OPG inside $\mathcal{S}(\mathbf{A})$, which will make the result more accurate.

2.3 Cross-validation

In most cases, d_0 is unknown, so we need to find a method to evaluate the estimated central (mean) space of different dimensions and find the best one. In MAVE package, cross-validation is used. In each iteration, the dataset is divided into training set and test set randomly. The size of test set is around $n - n^{2/3}$ to make the selection more consistent. Prediction error based on each central (mean) space is calculated. Further discussion about the consistency of the selected dimension can be found in [6]

3 Central space estimation

Since MAVE and OPG estimate the reduced dimensions by conditional expectation of Y given X, some information of Y given X is lost. Therefore, MAVE and OPG is not capable of finding central space exhaustively, but little change can be done to make these methods work. Following the idea of SIR, we divide the span of Y into some slices. Let $-\infty = s_0 < s_1 < \cdots < s_H = +\infty$,

$y_k = I_{(y<s_k)}$ and $Y_{ik} = I_{(Y_i<s_k)}$. By Prop. 2 in [7], if the slices are sufficiently dense, $S_{y|x}$ will coincide with the CMS of $(y_1, \cdots, y_H)$. We can use MAVE or OPG to estimate the CMS of $(y_1, \cdots, y_H)$, then CS of y can be obtained.

4 Kernel sliced inverse regression

Sliced inverse regression is proposed by [1]. Under certain condition on the conditional expectation about x, the centered conditional expectation $E(x|y) - E(x) \in S(\mathbf{B}_0)$, where $\mathbf{B}_0$ is the basis matrix of $\mathcal{S}_{y|x}$. However, there is no guarantee that SIR can exploit the central space exhaustively in some cases[2]. The main step for SIR is as follows:

1. Standardize design matrix X to $\tilde{X}$, $\tilde{X} = \hat{\Sigma}_{xx}(X - EX)$, $\hat{\Sigma}_{xx} = cov(X)$, such that $E(\tilde{X}) = 0$ and $cov(\tilde{X}) = I_p$. Divide the range of the value of Y into H slices, $S_1, \cdots, S_H$.

2. Let $\hat{p}_h$ be the frequency of Y_i falling into S_h, namely $\hat{p}_h = 1/n \sum_{i=1}^{n} I_{\{Y_i \in S_h\}}$, and $\hat{m}_h \in \mathbb{R}^p$ be the sample mean of the data in S_h, namely $\hat{m}_h = 1/(n\hat{p}_h) \sum_{i=1}^{n} I_{\{Y_i \in S_h\}} \tilde{X}_i$.

3. Construct the weighted covariance matrix of conditional mean $E(y|x)$: $\hat{V} = \sum_{h=1}^{H} \hat{p}_h \hat{m}_h \hat{m}_h^\top$.

4. Let the d_0 largest eigenvectors of $\hat{V}$ be $\eta_k (k = 1, \cdots, d_0)$. The basis of the estimated central space is $\beta_k = \hat{\eta}_k \hat{\Sigma}_{xx}^{-1/2} (k = 1, \cdots, d_0)$.

Kernel version of sliced inverse regression is using kernel method to estimate the conditional mean $E(x|y)$, which in [1] is computed by simply averaging the sample in each slice. This method will make the estimation more accurate and make the division of the range of Y unnecessary.

References

[1] Li, K. C. (1991). Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 86(414), 316-327.

[2] Cook, R.D. and Weisberg, S.(1991). Discussion of Li(1991). Journal of the American Statistical Association, 86, 328-332.

[3] Cook, R.D.(1998), Regression Graphics. New York: Wiley

[4] Cook, R. D., and Li, B. (2002). Dimension reduction for conditional mean in regression. Annals of Statistics, 455-474.

[5] Fan, J., and Gijbels, I. (1996). Local Polynomial Modelling and Its Applications, New York: Chapman and Hall.

[6] Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(3), 363-410.

[7] Wang, H., and Xia, Y. (2008). Sliced regression for dimension reduction. Journal of the American Statistical Association, 103(482), 811-821.