Communications of the Korean Statistical Society
2011, Vol. 18, No. 3, 267–275

DOI: 10.5351/CKSS.2011.18.3.267

A Short Note on Empirical Penalty Term Study of BIC in K-means Clustering Inverse Regression

Ji Hyun Ahn[a], Jae Keun Yoo[1,a]

[a]Department of Statistics, Ewha Womans University

Abstract

According to recent studies, Bayesian information criteria(BIC) is proposed to determine the structural dimension of the central subspace through sliced inverse regression(SIR) with high-dimensional predictors. The BIC may be useful in K-means clustering inverse regression(KIR) with high-dimensional predictors. However, the direct application of the BIC to KIR may be problematic, because the slicing scheme in SIR is not the same as that of KIR. In this paper, we present empirical penalty term studies of BIC in KIR to identify the most appropriate one. Numerical studies and real data analysis are presented.

Keywords: Bayesian information, inverse regression, multivariate regression, K-means clustering.

1. Introduction

The goal of sufficient dimension reduction(SDR) in regression of $Y|\mathbf{X} \in \mathbb{R}^p$ replaces the original p-dimensional many-valued or continuous predictors $\mathbf{X}$ by a lower-dimensional linear projection predictor without loss of information about the conditional distribution of $Y|\mathbf{X}$. That is, SDR pursues to find $\boldsymbol{\alpha} \in \mathbb{R}^{p\times q}$ such that

$$Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}, \tag{1.1}$$

where $\perp\!\!\!\perp$ stands for independence and $q \leq p$.

Equation (1.1) directly implies that the two conditional distributions of $Y|\mathbf{X}$ and $Y|\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}$ are the same. Subsequently, we attain the dimension reduction of $\mathbf{X}$ through usage of $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}$. Then a subspace spanned by the columns of such $\boldsymbol{\alpha}$ is called a dimension reduction subspace, and SDR typically seeks for the intersection of all dimension reduction subspaces. Then the intersection is called the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$. By its construction, $\mathcal{S}_{Y|\mathbf{X}}$ has the minimal dimension and is unique, if it exists. The existence of $\mathcal{S}_{Y|\mathbf{X}}$ is guaranteed under various mild conditions such as the open and convex support of $\mathbf{X}$. The true dimension and an orthonormal basis matrix of $\mathcal{S}_{Y|\mathbf{X}}$ will be denoted as d and $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}$. Then the lower-dimensional linear projection predictor $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ is called *sufficient predictors*.

One of the popular SDR methods to estimate $\mathcal{S}_{Y|\mathbf{X}}$ should be a methodology of sliced inverse regression (SIR; Li, 1991). The SIR estimates the structural dimension and an orthonormal basis of $\mathcal{S}_{Y|\mathbf{X}}$ through constructing a sample version of $E(\mathbf{X}|Y)$ by slicing Y. For the dimension determination,

For the corresponding author Dr. Jae Keun Yoo, this work was supported by Basic Science Research Program through the National Research Foundation of Korea(KRF) funded by the Ministry of Education, Science and Technology (2010-0003189).

[1] Corresponding author: Assistant Professor, Department of Statistics, Ewha Womans University, 11-1 Daehyun-Dong Seodaemun-Gu, Seoul 120-750, Korea. E-mail: peter.yoo@ewha.ac.kr

weighted chi-squared tests(WCT) are usually done. The WCT, however, has two major deficits. One is that the WCT cannot properly handle the significance level during the entire dimension tests, and the other is that it is limited to regressions of $n < p$. To overcome these deficits, Bayesian information criteria is proposed by Zhu *et al.* (2006), who recommend a proper form of penalty term in SIR.

For a multivariate regression of $\mathbf{Y} \in \mathbb{R}^r|\mathbf{X}$, the definition of the central subspace still holds. However, usual application of SIR in such regressions is problematic, because the number of slices increases exponentially. To avoid this problem, Setodji and Cook (2004) proposed K-means clustering inverse regression(KIR), which constructs slices through K-means clustering of $\mathbf{Y}$. Once $\mathbf{Y}$ is clustered, SIR is applied in a typical way. Therefore, in KIR, we can use the BIC for the dimension determination. Since the slicing schemes in SIR and KIR are different, however, we cannot directly use the penalty term recommended by Zhu *et al.* (2006).

This paper conducts empirical studies of penalty terms of the BIC for KIR for choosing the most appropriate one, which gives the most robust results in dimension estimation regardless of true multivariate regression models. For this, we consider two types of regressions. One is a case of $n > p$, which is a common type of regression, and the other is that of $n < p$. In the latter case, we have found no literature studies as of present.

2. Inverse Regressions and Bayesian Information Criteria

2.1. Sliced inverse regression

One of the most popular SDR methods to recover $\mathcal{S}_{Y|\mathbf{X}}$ is sliced inverse regression (SIR; Li, 1991). The SIR constructs a subspace $\mathcal{S}\{E(\mathbf{X}|Y)\}$, which is a subspace spanned by $E(\mathbf{X}|Y)$ by varying Y.

Then linearity condition that $E(\mathbf{X}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X})$ is linear in $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ guarantees that $\mathcal{S}\{E(\mathbf{X}|Y)\} \subseteq \mathcal{S}_{Y|\mathbf{X}}$. The linearity condition is common in SDR literature. If $\mathbf{X}$ has an elliptically contoured distribution, the condition is automatically satisfied. In the case that the linearity condition does not hold, $\mathbf{X}$ can often be one-to-one transformed to satisfy this condition. By assuming coverage condition of $\mathcal{S}\{E(\mathbf{X}|Y)\} = \mathcal{S}_{Y|\mathbf{X}}$, the SIR recovers $\mathcal{S}_{Y|\mathbf{X}}$ exhaustively. The coverage condition seems mild and has reasonable approximation in SDR literature according to Cook and Ni (2005).

Since it is known that $\mathcal{S}\{E(\mathbf{X}|\mathbf{Y})\} = \mathcal{S}[\mathbf{M}_{\mathrm{SIR}} := \mathrm{cov}\{E(\mathbf{X}|Y)\}]$, the sample version of $\mathrm{cov}\{E(\mathbf{X}|Y)\}$ is usually constructed to estimate $\mathcal{S}_{Y|\mathbf{X}}$ as follows :

(1) If Y is categorical, each category forms slices. If Y is many-valued or continuous, divide the observed range of Y into h slices J_h.

(2) Compute sample means of $\hat{\mathbf{X}}$ within each slice, $\bar{\mathbf{X}}_s = 1/n_s \sum_{Y_s \in J_s} \mathbf{X}_i$, $s = 1, \ldots, h$, where n_s is the number of observations within J_s.

(3) $\hat{\mathbf{M}}_{\mathrm{SIR}} := \widehat{\mathrm{cov}}\{E(\mathbf{X}|Y)\} = \sum_{s=1}^{h} f_s(\bar{\mathbf{X}}_s - \bar{\mathbf{X}})(\bar{\mathbf{X}}_s - \bar{\mathbf{X}})^{\mathrm{T}}$, where $f_s = n_s/n$ and $\bar{\mathbf{X}} = (1/n) \sum_{i=1}^{n} \mathbf{X}_i$.

Once $\hat{\mathbf{M}}_{\mathrm{SIR}}$ is constructed, then the true structural dimension d and basis $\boldsymbol{\eta}$ of $\mathcal{S}_{Y|\mathbf{X}}$ are estimated by spectral decomposition of $\hat{\mathbf{M}}_{\mathrm{SIR}}$. In the population level, the true rank of $\mathbf{M}_{\mathrm{SIR}}$ is equal to d, and hence d can be estimated by identifying non-zero eigenvalues of $\hat{\mathbf{M}}_{\mathrm{SIR}}$ throughout testing the following sequence of hypotheses:

(1) Start testing $H_0 : d = m$ versus $H_1 : d > m$ with $m = 0$.

(2) If H_0 is rejected, then add 1 to m. Repeat (1).

(3) Repeat steps 1 and 2 for the first time H_0 is not rejected. Then the hypothesized d set $\hat{d}$.

This test procedure requires a statistic for $H_0 : d = m$, which will be discussed in a later subsection. Once d is determined, $\hat{\boldsymbol{\eta}}$ is constructed from the eigenvectors corresponding to non-zero eigenvalues of $\hat{\mathbf{M}}_{\text{SIR}}$.

2.2. K-means inverse regression

With multi-dimensional responses $\mathbf{Y} = (Y_1, \ldots, Y_r)^{\text{T}}$, the number of slices increase exponentially. Therefore, for high dimensional responses, the usual slicing scheme faces the curse of dimensionality. For example, letting $r = 4$, the minimum total number of slices will be $2^4 = 16$, and this minimum number of slices may not be effective for a small sample of size 100 or less. Although SIR can be implemented in this example, we cannot expect a reliable estimate for $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$.

In terms of methodology, the slicing procedure is only required to estimate the inverse mean $E(\mathbf{X}|\mathbf{Y})$. To construct slices more effectively, we cluster the responses $\mathbf{Y}$ by a K-means clustering algorithm and use the clusters as slices. Once the clusters are formed, SIR can be applied in a typical way. Setodji and Cook (2004) call this approach K-means inverse regression(KIR).

2.3. Weighted chi-squared tests and Bayesian information criteria

Usual statistics for testing $H_0 : d = m$ discussed in Section 2.1 is as follows:

$$\hat{\Lambda}_m = n \sum_{j=m+1}^{p} \hat{\lambda}_j, \quad m = 0, 1, \ldots, \min(p-1, h-1),$$

where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{\min(p,h-1)} \geq \hat{\lambda}_{\min(p,h-1)+1} = \cdots = \hat{\lambda}_p = 0$ are the ordered eigenvalues of $\hat{\mathbf{M}}_{\text{SIR}}$.

According to Bura and Cook (2001), the statistics $\hat{\Lambda}_m$ tends in distribution to χ^2 or weight sum of independent χ^2 with one degree of freedom. The test statistics are usually used in SIR for the dimension determination. In the sequential dimension tests with $\hat{\Lambda}_m$, the choice of the nominal level at each step is essential in estimating d. However, the nominal level at each step does not control the nominal level of the entire test procedure. In addition, the asymptotics of $\hat{\Lambda}_m$ fails with $n < p$, and hence the weighted chi-squared tests are not possible in such case.

To avoid these deficits, Zhu *et al.* (2006) proposed Bayesian information criteria(BIC) to estimate d. First we construct $\hat{\boldsymbol{\Omega}} = \hat{\mathbf{M}}_{\text{SIR}} + \mathbf{I}_p$. Let $\hat{\theta}_1 \geq \cdots \geq \hat{\theta}_p \geq 1$ be the ordered eigenvalues of $\hat{\boldsymbol{\Omega}}$. Define τ as the number of $\hat{\theta}_j > 1$. Then Bayesian information based on this is as follows:

$$G(m) = \frac{n}{2} \sum_{j=1+\min(\tau,m)}^{p} \left(\log \hat{\theta}_j + 1 - \hat{\theta}_j\right) - \frac{C_n m(2p - m + 1)}{2}, \quad m = 0, 1, \ldots, \min(p-1, h-1), \quad (2.1)$$

where C_n is a penalty term.

Then an estimate $\hat{d}_{\text{BIC}}$ of d is the maximizer m of $G(m)$, that is,

$$G\left(\hat{d}_{\text{BIC}}\right) = \max_{0 \leq m \leq (p-1, h-1)} G(m).$$

Under mild conditions, $\hat{d}_{\text{BIC}}$ is consistent. It is clear that $\hat{d}_{\text{BIC}}$ depends on the choice of the number of slices and C_n. Zhu *et al.* (2006) discuss that $\hat{d}_{\text{BIC}}$ is quite robust to the number of slices and suggest $C_n = (h/n)(0.5 \log n + 0.1 n^{1/3})/2$.

3. Adaptation of BIC in KIR

The quantity $G(m)$ given in (2.1) clearly depends on the forms of the penalty C_n. We investigate C_n more explicitly. According to Zhu *et al.* (2006), C_n must satisfy the following two conditions for $\hat{d}$ to be a consistent estimate of d:

$$\text{for } t > 0 \text{ and } 2t > s, \quad \text{(a)} \lim_{n\to+\infty} C_n/n^{1-s} = 0; \qquad \text{(b)} \lim_{n\to+\infty} C_n/n^{1-2t} = \infty.$$

If we select $C_n = O(n^a)$, the number a should be selected from $1 - 2t < a < 1 - s$, equivalently $4s < a < 1 - s$, where $O(n^a)$ stands for usual Big-O notations, that is, $C_n = O(n^a)$ means that $\lim_{n\to+\infty} |C_n|/|n^a| \leq M$ with a constant M.

Zhu *et al.* (2006) divide C_n into two parts: $C_n = c^{-1}W_n$. For the constant c, Zhu *et al.* (2006) selected the average sample sizes per slice n/h. It is known that the WCT depends on the number of slices. Therefore, the choice of n/h for c is very reasonable, because the BIC should also account for the slice size impact in the structural dimension estimation.

For W_n, various candidates can be considered as forms of n^a with $0 < 4s < a < 1 - s$ for fixed s. In BIC context, one of the popular choices for n^a is $\log n$ (Schwarz, 1978). Besides of $\log n$, Zhu *et al.* (2006) also consider $n^{1/3}$ and several linear combinations of the two. From these candidates, Zhu *et al.* (2006) found the most appropriate C_n for SIR through simulation studies.

In KIR, we cannot expect the same results as SIR, because slicing schemes of KIR and SIR are clearly different. Usual slicing scheme in SIR is to divide response variables into h slices to have almost equally samples per each slice. However, K-means clustering algorithm does not normally construct clusters to have almost equally samples per cluster. In addition, in KIR, we restrict the minimum samples per cluster to two for each cluster to be informative in estimation of $E(\mathbf{X}|\mathbf{Y})$. Therefore, it is straightforward that Zhu *et al.*'s suggestion of $C_n = (h/n)(0.5\log n + 0.1n^{1/3})/2$ for SIR may not be a best one in KIR.

Here we consider the most appropriate C_n for KIR with two cases of $n > p$ and $n < p$. The construction of sample kernel matrices for the KIR and implementation of BIC are not restricted to either $n > p$ and $n < p$. Again C_n may be not the same in both the cases.

Following the guidance of Zhu *et al.* (2006), $\log n$, $n^{1/3}$, and their linear combinations will be considered as choices for W_n. And, for c, we will consider average sample sizes per slices, n/h and the medians of sample sizes of clusters, $\mathrm{med}(h)$. Based on the above suggestions, the following candidates for C_n will be considered for $n > p$ and $n < p$.

- Case 1: $n > p$:

(1.1a) $C^{1a}_{n>p} = (h/n)\left(0.1\log n + 0.5n^{1/3}\right)$; (1.1b) $C^{1b}_{n>p} = \mathrm{med}(h)^{-1}\left(0.1\log n + 0.5n^{1/3}\right)$

(1.2a) $C^{2a}_{n>p} = (h/n)\left(0.1\log n + 0.5n^{1/3}\right)/2$; (1.2b) $C^{2b}_{n>p} = \mathrm{med}(h)^{-1}\left(0.1\log n + 0.5n^{1/3}\right)/2$

(1.3a) $C^{3a}_{n>p} = (h/n)\left(0.5\log n + 0.1n^{1/3}\right)$; (1.3b) $C^{3b}_{n>p} = \mathrm{med}(h)^{-1}\left(0.5\log n + 0.1n^{1/3}\right)$

(1.4a) $C^{4a}_{n>p} = (h/n)\left(0.5\log n + 0.1n^{1/3}\right)/2$; (1.4b) $C^{4b}_{n>p} = \mathrm{med}(h)^{-1}\left(0.5\log n + 0.1n^{1/3}\right)/2$.

- Case 2:$n < p$:

(2.1a) $C^{1a}_{n<p} = (h/n)\log n$; (2.1b) $C^{1a}_{n<p} = \mathrm{med}(h)^{-1}\log n$

(2.2a) $C^{2a}_{n<p} = (h/n)n^{1/3}$; (2.2b) $C^{1a}_{n<p} = \mathrm{med}(h)^{-1}n^{1/3}$

(2.3a) $C^{3a}_{n<p} = (h/n)\left(\log n + n^{1/3}\right)/2$; (2.2c) $C^{1a}_{n<p} = \mathrm{med}(h)^{-1}\left(\log n + n^{1/3}\right)/2$.

We will not consider $\log n$ and $n^{1/3}$ multiplied by 0.5 or 0.1 for $n < p$, because preliminary numerical studies showed that they made C_ns too small and usually underestimated the true structural dimension.

From various simulated multivariate regression models, we will see which candidates give the most robust results. To report the numerical study results, in the case of $n > p$, we will compute the percentages of the decisions that $\hat{d} \leq d$, $\hat{d} = d$ and $\hat{d} \geq d$. We will consider C_n to give the most frequencies of $\hat{d} = d$ as a best one. In the case of $n < p$, initial interest is typically to reduce the dimension of $\mathbf{X}$ enough to develop proper statistical models. In the underestimation of d, the dimension-reduced predictors are not sufficient to have equal information to the original predictors $\mathbf{X}$, while there is, at least, no information loss about $\mathbf{Y}|\mathbf{X}$ in the overestimation. Therefore, underestimation of d is more problematic than its overestimation. Then it should be important to compute the percentages of the decisions that $\hat{d} \geq d$ in summarizing the numerical studies for $n < p$. The percentages of the decisions that $\hat{d} \geq d$ can be interpreted as the ratio of the number of correctly identified sufficient predictors to the number of true sufficient predictors. In this sense, the percentages of $\hat{d} \geq d$ can be considered as a version of true positive rate(TPR), which is a measure commonly used in biomedical literature. We will report TPR in the case of $n < p$.

According to various numerical studies, for $n > p$, the penalty term of $C^{3a}_{n>p} = (h/n)^{-1}(0.5 \log n + 0.1n^{1/3})$ provides the most robust asymptotic results in the estimation of d, while the penalty term of $C^{1a}_{n<p} = (h/n) \log n$ seems the best among the others for $n < p$.

4. Numerical Studies and Data Analysis

From the preliminary simulation results (not reported), good number of clusters should be between 5 and 7. In all simulations, we used 6 clusters.

4.1. $n > p$

To construct Model 1, 10-dimensional predictors $\mathbf{X} \in \mathbb{R}^{10}$ were randomly sampled from $N(0, \mathbf{\Sigma})$, where all diagonal elements of $\mathbf{\Sigma}$ are equal to one and all the other off-diagonal elements are equal to 0.5. Random error vectors $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_4)^{\mathrm{T}}$ were independently generated from $N(0, 1)$ and $\varepsilon \perp\!\!\!\perp \mathbf{X}$. Based on the variable configurations, we consider the following regression model introduced in Setodji and Cook (2004) with indicating $1_4 \in \mathbb{R}^{10} = (1, 1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$.

Model 1

$$Y_1 = c_1 1_4^{\mathrm{T}}\mathbf{X} + c_2 \exp\left(c_3 1_4^{\mathrm{T}}\mathbf{X}\right)\varepsilon_1;$$

$$Y_2 = c_1 1_4^{\mathrm{T}}\mathbf{X} + c_2 \exp\left\{c_3\left(2 - 31_4^{\mathrm{T}}\mathbf{X}\right)\right\}\varepsilon_2;$$

$$Y_3 = c_1 1_4^{\mathrm{T}}\mathbf{X} + c_2 \exp\left(2c_3 1_4^{\mathrm{T}}\mathbf{X}\right)\varepsilon_3;$$

$$Y_4 = c_1 1_4^{\mathrm{T}}\mathbf{X} + c_2 \exp\left\{c_3\left(1 - 1_4^{\mathrm{T}}\mathbf{X}\right)\right\}\varepsilon_4.$$

In Model 1, the central subspace $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ is spanned by 1_4 regardless of the actual numbers of c_1, c_2 and c_3, and hence the structural dimension is equal to one.

The purpose of Model 1 is to see how well the true structural dimension is estimated under the existence of heteroscedasticity. The forms of heteroscedasticity in Model 1 are quite popular, and Cook and Weisberg (1983) developed a methodology to test this.

From Model 1, we considered the following three different scenarios: (1) homoscedastic linear model with $c_1 = c_2 = 1$ and $c_3 = 0$; (2) heteroscedastic linear model with $c_1 = 0.1$, $c_2 = 1$ and

Table 1: Dimension estimation for Model 1 in Section 4.1: homoscedastic linear model

	WCT	$C^{1a}_{n>p}$	$C^{1b}_{n>p}$	$C^{2a}_{n>p}$	$C^{2b}_{n>p}$	$C^{3a}_{n>p}$	$C^{3b}_{n>p}$	$C^{4a}_{n>p}$	$C^{4a}_{n>p}$
$\hat{d} = 0$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\hat{d} = 1$	94.0	96.8	98.6	40.2	45.2	97.0	98.8	41.4	45.0
$\hat{d} \geq 2$	6.0	3.2	1.4	69.8	54.8	3.0	1.1	58.6	55.0

Table 2: Dimension estimation for Model 1 in Section 4.1: heteroscedastic linear model

	WCT	$C^{1a}_{n>p}$	$C^{1b}_{n>p}$	$C^{2a}_{n>p}$	$C^{2b}_{n>p}$	$C^{3a}_{n>p}$	$C^{3b}_{n>p}$	$C^{4a}_{n>p}$	$C^{4a}_{n>p}$
$\hat{d} = 0$	0.07	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{d} = 1$	88.60	98.60	98.20	46.00	50.40	97.40	98.60	44.40	56.20
$\hat{d} \geq 2$	3.80	1.20	1.80	54.00	49.60	2.60	1.40	55.60	43.80

Table 3: Dimension estimation for Model 1 in Section 4.1: heteroscedastic linear model with constant mean

	WCT	$C^{1a}_{n>p}$	$C^{1b}_{n>p}$	$C^{2a}_{n>p}$	$C^{2b}_{n>p}$	$C^{3a}_{n>p}$	$C^{3b}_{n>p}$	$C^{4a}_{n>p}$	$C^{4a}_{n>p}$
$\hat{d} = 0$	7.8	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0
$\hat{d} = 1$	88.0	97.6	97.0	46.2	49.0	97.8	98.2	46.0	53.0
$\hat{d} \geq 2$	4.2	2.4	2.8	53.8	51.0	2.2	1.8	54.0	47.0

$c_3 = 0.1$; (3) heteroscedastic linear model with constant mean with $c_1 = 0$, $c_2 = 1$ and $c_3 = 0.1$. In the simulations, Model 1 was iterated 1000 times with $n = 100$ per each case. The dimension estimation results are summarized in Tables 1–3. In the tables, WCT indicates weighted chi-squired tests and is reported for comparison purpose. For the WCT, nominal level 5% was used. In the WCT, if the percentages of $\hat{d} = 2$ is close to 95%, the WCT performs very well. Using the BIC, the percentages should be close to 100% for good estimation of the true structural dimension.

According Tables 1–3, among 8 candidates for C_n for the BIC, $C^{1\bullet}_{n>p}$ and $C^{3\bullet}_{n>p}$ clearly dominate $C^{2\bullet}_{n>p}$ and $C^{4\bullet}_{n>p}$, and hence the suggestion given by Zhu *et al.* (2006) for SIR, which is $C^{4a}_{n>p}$, does not hold any more in KIR.

It is hard to say which one among $C^{1\bullet}_{n>p}$ and $C^{3\bullet}_{n>p}$ is the best, and hence any of the four penalty terms seems usable in practice. For homoscedastic linear models (Tables 2–3), both BIC and WCT worked very well, but, if heteroscedasticity exits, BIC gives better dimension estimation results.

Another simulation model, Model 2, focuses on various types of mean functions along with non-normal predictors. To construct predictors for Model 2, we generate the following variables: $V_i \overset{iid}{\sim}$ Uniform$(-4, 4)$, $i = 1, 2, 3$, $W_1 \sim 0.5N(0, 4) + 0.5N(0, 16)$ and $W_2 \sim$ Uniform$(-4, 4)$, and all W_is and W_js are independent. Then 10-dimensional predictors are randomly generated as follows: $X_1 = W_1$, $X_2 = V_1 + (1/2)W_2$, $X_3 = -V_1 + (1/2)W_2$, $X_4 = V_2 + V_3$, $X_5 = V_2 - V_3$ and $(X_6, \ldots, X_10)^{\mathrm{T}} \overset{iid}{\sim} N(0, 4)$ independent of $(X_1, \ldots, X_5)$. Next we define that $\eta_1 \in \mathbb{R}^{10} = (1, 0, \ldots, 0)^{\mathrm{T}}$ and $\eta_2 \in \mathbb{R}^{10} = (0, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$. Then Model 2 is constructed as follows:

Model 2

$$Y_1 = \left(4 + \eta_1^{\mathrm{T}}\mathbf{X}\right)\left(\eta_2^{\mathrm{T}}\mathbf{X} + 2\right) + 0.5\varepsilon_1;$$

$$Y_2 = \eta_1^{\mathrm{T}}\mathbf{X} + 0.5\left(\eta_2^{\mathrm{T}}\mathbf{X}\right)\varepsilon_2;$$

$$Y_3 = \eta_2^{\mathrm{T}}\mathbf{X} + \left(\eta_2^{\mathrm{T}}\mathbf{X}\right)^2 + 0.5\varepsilon_3;$$

$$Y_4 = 0.5\varepsilon_4,$$

where $(\varepsilon_1, \ldots, \varepsilon_4)^{\mathrm{T}} \overset{iid}{\sim} N(0, 1) \perp\!\!\!\perp \mathbf{X}$.

Table 4: Dimension estimation for Model 2 in Section 4.1

	WCT	$C^{1a}_{n>p}$	$C^{1b}_{n>p}$	$C^{2a}_{n>p}$	$C^{2b}_{n>p}$	$C^{3a}_{n>p}$	$C^{3b}_{n>p}$	$C^{4a}_{n>p}$	$C^{4a}_{n>p}$
$\hat{d} = 0$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\hat{d} = 1$	28.2	16.2	29.0	0.4	1.8	12.4	28.4	0.0	3.0
$\hat{d} = 2$	67.2	82.6	70.6	61.4	76.0	86.4	70.6	59.0	72.2
$\hat{d} \geq 3$	4.6	1.2	0.4	38.2	22.2	1.2	1.0	41.0	24.8

Table 5: Dimension estimation for Model 3 in Section 4.2

	$C^{1a}_{n<p}$	$C^{1b}_{n<p}$	$C^{2a}_{n<p}$	$C^{2b}_{n<p}$	$C^{3a}_{n<p}$	$C^{3b}_{n<p}$
TPR	99.8	80.0	93.5	88.2	95.5	81.0
$\hat{d} = 2$	74.5	56.0	73.0	59.0	72.0	57.5

In Model 2, the central subspace $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ is spanned by the two columns of η_1 and η_2, and hence the true structural dimension is equal to two. Model 2 has linear mean, constant mean and non-linear mean with heteroscedasticity. In the simulations, Model 2 was iterated 1000 times with $n = 100$ per each case. Since Model 2 has more complex mean structure, we expect that the percentages of $\hat{d} = 2$ should be worse than Model 1. We summarize the simulation results in Table 4.

In Model 2, we can confirm that $C^{1a}_{n>p}$ and $C^{3a}_{n>p}$ clearly dominate $C^{2\bullet}_{n>p}$ and $C^{4\bullet}_{n>p}$ in the estimation of the correct decision of $\hat{d} = 2$. And, the behaviors of $C^{1\bullet}_{n>p}$ and $C^{2\bullet}_{n>p}$ are similar to $C^{3\bullet}_{n>p}$ and $C^{4\bullet}_{n>p}$ respectively. One notable thing is types of misspecification of d. The penalty terms of $C^{1\bullet}_{n>p}$ and $C^{3\bullet}_{n>p}$ tends to underestimate d, which is a similar pattern to WCT, while $C^{2\bullet}_{n>p}$ and $C^{4\bullet}_{n>p}$ often overestimate d. According to Table 4, the usage of BIC give a superior estimation of d than WCT. For Model 2, the most frequencies of $\hat{d} = 2$ happens with $C^{3a}_{n>p}$. Models 1 and 2 represent the characteristic behaviors in dimension estimation we observed in other simulations. Based on these results, we conclude that $C^{3a}_{n>p}$ should be the most appropriate penalty in KIR with $n > p$.

4.2. $n < p$

For constructing a regression of $n < p$, we randomly selected 200-dimensional predictors $\mathbf{X} = (X_1, \ldots, X_{200})$ from $X_i \overset{iid}{\sim} N(0, 1)$, $i = 1, \ldots, 200$. Then we independently generated random errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_4)^{\mathrm{T}}$ from $N(0, 1)$ and $\boldsymbol{\varepsilon} \perp\!\!\!\perp \mathbf{X}$. In the simulations, Model 3 was iterated 1000 times with $n = 100$ per each case.

Model 3

$$Y_1 = X_1 + X_1 * X_2 + 0.1\varepsilon_1;$$

$$Y_2 = X_1 + 0.1 \exp(X_2)\varepsilon_2;$$

$$Y_3 = X_1 + 0.1\varepsilon_3;$$

$$Y_4 = X_2 + 0.1\varepsilon_4.$$

In Model 3, the central subspace $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ is spanned by the two columns of $(1, 0, \ldots, 0)$ and $(0, 1, \ldots, 0)$, and hence the true structural dimension is equal to two. Model 3 has linear mean, constant mean and non-linear mean with heteroscedasticity. TPR for Model 3 along with the percentages of the correct decisions that $\hat{d} = 2$ is given in Table 5. According to Table 5, usage of the average sample sizes of clusters always gives better TPR than that of the median sample sizes of clusters including the percentages of $\hat{d} = 2$. In addition, all three of $\log n$, $n^{1/3}$ and $0.5(\log n + n^{1/3})$ gives almost equal

Table 6: Dimension estimation for Minneapolis school data in Section 4.3

	$H_0 : d = 0$	$H_0 : d = 1$	$H_0 : d = 2$	$H_0 : d = 3$
WCT	0.000	0.043	0.241	0.659
BIC	−8.796	−4.739	−5.043	−5.854

performances. However, $\log n$ seems slightly better than the other two. Since we observed similar results from other simulations, one can use $C^{1a}_{n<p}$ as a good choice of penalties for BIC in regressions of $n < p$.

4.3. Data analysis-Minneapolis school

To illustrate a methodology introduced in the previous sections, we use data on the performance of students in $n = 63$ Minneapolis schools studied by Cook (1998). The four dimensional responses $\mathbf{Y}$ consists of the percentages $P_{(\cdot)}$ of students in a school scoring above (A) and below (B) average on standardized fourth and sixth grade reading comprehension tests, $\mathbf{Y} = (P_{A4}, P_{B4}, P_{A6}, P_{B6})^{\mathrm{T}}$. Subtracting either pair of grade specific percentages from 100 gives the percentage of students scoring about average on the test. We used the nine predictors in the dataset: (1) the percentage of children receiving Aid to Families with Dependent Children(AFDC), (2) the average percentage of children in attendance during the year (Attend), (3) the percentage of children not living with both biological parents (B), (4) the number of children enrolled in the school (Enrol), (5) the percentage of adults in the school area who completed high school(HS), (6) the percentage of minority children in the area (Minority), (7) the percentage of children who started in a school, but did not finish there (Mobility) (8) the percentage of persons in the area below the federal poverty level(PL), (9) the pupil-teacher ratio(PTR).

The nine predictors were transformed to square-root to induce the linearity conditions. For the dimension determination, the Bayesian informations with $C^{3a}_{n>p}$ and p-values for the WCT up to $H_0 : d = 3$ are reported in Table 6. The BIC decides that $\hat{d} = 1$, while the WCT concludes that $\hat{d} = 2$ with level 5%.

The p-value for $H_0 : d = 1$ is quite close to the nominal level, and we may conclude that $\hat{d} = 1$ without further investigation. To obtain useful information for deciding between $\hat{d} = 1$ and $\hat{d} = 2$, however, the following simulation is considered. Define that $\mathbf{X}_0 = \hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}$ is the estimated sufficient predictor from KIR with $d = 1$. We constructed new data sets from the model of $Y^*_{k_i} = f_k(\mathbf{X}_{0_i}) + \sigma_k\varepsilon_{k_i}$, $i = 1, 2, \ldots, 63$, $k = 1, 2, 3, 4$, where f_k is a LOWESS smooth of Y_k against $\mathbf{X}_0$ using 0.7 as the tuning parameter, $\sigma^2_k = 62^{-1}\sum^{63}_{i=1}\{Y_{k_i} - f_k(\mathbf{X}_{0_i})\}^2$ and the ε_{k_i}'s are independent standard normal random variables. These data sets were generated 1000 times in this way, and the true null hypothesis $d = 1$ at nominal level 5% was tested. The percentages of $\hat{d} = 1$ were 80.0% for the WCT and 86.3% for the BIC. Clearly, the BIC estimates the true dimension slightly better than the WCT, we concluded that $\hat{d} = 1$.

5. Discussions

In this paper, we perform empirical studies to choose the most appropriate penalty term of BIC in K-means clustering inverse regression(KIR). For this, we consider two types of regressions. One is a case of $n > p$, which is a common type of regression, and the other is that of $n < p$. In the two types of regressions, we suggest two different quite good penalty terms regardless of regression models.

Zhu *et al.* (2006) provides a guideline about the penalty terms in BIC using sliced inverse regression for regressions of $n > p$. Since KIR is different from sliced inverse regression in slicing scheme, we may expect that the penalty guidances recommended by Zhu *et al.* (2006) for SIR cannot be directly applied to KIR. The studies confirm that we had better use different penalties in KIR from Zhu *et al.*'s recommendation in SIR. In case of regressions of $n < p$, we newly recommend another penalty term for BIC, which is different from that for regressions of $n > p$.

Usual weighted chi-squared tests for dimension determination in KIR are limited in $n > p$, and hence KIR is clearly not fully useful in practice. Usage of BIC does not require a condition of $n > p$, However, there is no clear guideline about what penalty terms should be used. In this paper we suggest two penalty terms in the BIC throughout various empirical studies depending on relation of n and p, and it is expected that our work provides one solution to answer the increasing demand of high-dimensional data analysis.

Acknowledgements

The authors are grateful to two referees for many helpful comments.

References

Bura, E. and Cook, R. D. (2001). Extended sliced inverse regression: The weighted chi-squared test, *Journal of the American Statistical Association,* **96**, 996–1003.

Cook, R. D. (1998). *Regression Graphics,* Wiley, New York.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach, *Journal of the American Statistical Association,* **100**, 410–428.

Cook, R. D. and Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression, *Biometrika*, **70**, 1–10.

Li, K. C. (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association,* **86**, 316–342.

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Mathematical Statististics*, **30**, 461–464.

Setodji, C. M. and Cook, R. D. (2004). K-means inverse regression, *Technometrics*, **46**, 421–429.

Zhu, L., Miao, B. and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates, *Journal of the American Statistical Association,* **101**, 630–643.

Received March 2011; Accepted April 2011