L_1-Statistical Procedures and Related Topics
IMS Lecture Notes – Monograph Series (1997) Volume 31

A comparison of procedures based on inverse regression

Simon J. Sheather

University of New South Wales, Australia

Joseph W. McKean

Western Michigan University, USA

Abstract: Sliced inverse regression (SIR) was introduced by Li (1991) and Duan and Li (1991) as a dimension reduction technique that determines the number of linear combinations of the predictor variables needed to obtain a parsimonious regression model. It is well known that SIR is not robust to the effects of outliers nor can it always detect symmetric dependence. In this paper, we briefly outline another technique based on inverse regression which potentially overcomes these shortcomings of SIR in an important special case. Finally, we compare the effectiveness of the new technique with that of SIR on some real data sets.

Key words: Dimension reduction, sliced inverse regression, robustness.

AMS subject classification: 62J20, 62G35.

1 Introduction

Regression analysis is arguably one of the most widely used statistical techniques. A regression model expresses the mean of a response variable y as a function, f, of an explanatory variable $\mathbf{x}$, a p-dimensional column vector. Traditional parametric regression methods assume that the functional relationship between y and $\mathbf{x}$ is known apart from some parameters, which must be estimated. When the assumed functional form is correct, a variety of methods (including least squares and robust methods) can be used to estimate the unknown parameters. However, in many applications any parametric model is at best an approximation to the true one and the search for an adequate model becomes increasingly difficult as p, the number of predictors, increases.

Alternatively, nonparametric regression methods estimate the regression function without assuming a particular functional form. Recently, Wand and Jones (1995), Fan and Gijbels (1996) and Simonoff (1996) have provided comprehensive accounts of this field. Many nonparametric regression methods are based on the notion of local smoothing, that is, the estimate of f at any point of interest is based on a smoothed version of y in that region. Thus, the success of local smoothing depends on the existence of sufficiently many data points around each point of interest in the design space to provide adequate information about f. As the dimension of $\mathbf{x}$ increases, larger and larger sample sizes are needed in order to ensure that there are sufficient data points around each point of interest. This problem has been appropriately referred to as the curse of dimensionality (Bellman, 1961). Hastie and Tibshirani (1990, pp. 83, 84) provide a simple, yet effective, example of this phenomenon. A number of approaches have been proposed to cope with the curse of dimensionality. Additive models (see Hastie and Tibshirani, 1990, Chapter 4) approximate f as the sum of nonparametric univariate functions of each of the p predictors. Alternatively, sliced inverse regression (Li, 1991 and Duan and Li, 1991) is a dimension reduction technique that does not rely on a complicated model-fitting process. Sliced inverse regression (SIR) determines the number of linear combinations of the p predictors needed to obtain a parsimonious model for f.

In the next section, we briefly outline a dimension reduction technique based on inverse regression. Finally, in Section 3 we compare the effectiveness of this new technique with that of SIR on some real data sets. We show that the new technique potentially overcomes two of the shortcomings of SIR, namely, a lack of robustness to outliers and a failure to detect some forms of symmetric dependence.

2 Inverse regression methods for dimension reduction

Consider the following general regression model

$$y = f(\beta_1'\mathbf{x}, \ldots, \beta_k'\mathbf{x},\ \epsilon)\,, \tag{1}$$

where f is an unknown arbitrary function, $\mathbf{x}$ is a p-dimensional vector of predictors, and ϵ is an $n \times 1$ vector of errors which is assumed to be independent of $\mathbf{x}$. The integer $k(< p)$ is the number of linear combinations of the predictors that are needed to summarize the dependence of y on $\mathbf{x}$. Li (1991, 1992), Cook and Weisberg (1991) and Schott (1994) provide methods for determining the value of k based on sliced inverse regression.

The simplest nontrivial case occurs when $k = 1$, since then $\beta_1'\mathbf{x}$ contains all the information from $\mathbf{x}$ about y. If $k \geq 2$ then commonly the aim is to reduce k. One possibility, for example, is to seek extra explanatory variables. These extra variables could include interaction terms, polynomial terms and/or dummy variables. Alternatively, the current set of explanatory variables could be changed using transformations. Further discussion of this issue can be found in Cook and Weisberg (1994, Chapter 8).

Thus arguably, an important special case is the decision as to whether $k = 1$ or $k > 1$. In this paper, we focus on this problem.

Under the assumptions that model (1) holds with $k = 1$ and that the distribution of $\mathbf{x}$ is elliptically symmetric, Duan and Li (1991) obtained the following result

$$E(x_j|y) = E(x_j) + \gamma_j\,\kappa(y) \qquad j = 1, \ldots p. \tag{2}$$

This result means that for each predictor x_j, the inverse regression function $E(x_j|y)$ equals the mean of x_j plus some unknown function $\kappa(y)$ times the constant γ_j. A crucial aspect of this result is that $\kappa(y)$ does not depend on j. Hence, a graphical procedure to decide whether $k = 1$ or not is to examine the p plots with x_j on the vertical axis and y on the horizontal axis to see if each plot has the same shape. Such a procedure is advocated by Cook and Weisberg (1994, Chapter 8).

Sheather and McKean (1997) have developed two nonparametric methods for testing whether $k = 1$ or not. The two procedures are based on the following observation. Suppose that model (1) holds with $k = 1$ and that the distribution of $\mathbf{x}$ is elliptically symmetric. Then, for $1 \leq i, j \leq p$

$$\log\left[\left|\frac{E(x_j|y) - E(x_j)}{E(x_i|y) - E(x_i)}\right|\right] = \log\left[\left|\frac{\gamma_j}{\gamma_i}\right|\right], \tag{3}$$

which is independent of y. In practice, the left side of (3) can be replaced by

$$L_{i,j} = \log\left[\left|\frac{x_j - \overline{x}_j}{x_i - \overline{x}_i}\right|\right].$$

Thus, a test of $k = 1$ against the alternative $k > 1$ can be obtained by testing for each combination of i and j whether $L_{i,j}$ is independent of y. Sheather and McKean (1997) have developed the following two tests.

- *Test1:* This test is based on dividing up each plot of $L_{i,j}$ versus y into four quadrants and counting up the number of points in each quadrant. The quadrants are obtained by splitting both $L_{i,j}$ and y into two groups depending on whether they are larger or smaller than their respective medians. If for a given i and j, $L_{i,j}$ does not depend on y then we expect

$n/4$ points to fall in each of the four quadrants. Departures from these expected frequencies are tested using a χ^2 goodness-of-fit test. Sheather and McKean (1997) have found that a significant result on *Test1* may indicate that polynomial terms are missing from the regression model.

- *Test2:* This test of independence is based on a runs statistic. In this case, the statistic used is the number of runs above and below the median value of $L_{i,j}$ in each plot of $L_{i,j}$ versus y. Sheather and McKean (1997) have found that a significant result on *Test2* may indicate that interaction terms are missing from the regression model.

For a more detailed description and discussion of *Test1* and *Test2* see Sheather and McKean (1997).

3 Examples

In this section, we discuss several examples involving real data taken from Cook and Weisberg (1994). In each example, the aim is to decide whether $k = 1$ or $k > 1$ in (1), that is, whether one linear combination of the predictors can adequately summarize the dependence of y on $\mathbf{x}$. The examples have been chosen to illustrate that the inverse regression technique potentially overcomes two of the shortcomings of SIR, namely, a lack of robustness to outliers and a failure to detect some forms of symmetric dependence.

The $R - code$ software supplied with Cook and Weisberg (1994) was used to calculate SIR. In each example the default $R - code$ settings for SIR were used. Rank based regression estimates were calculated using the experimental MINITAB command *rregress*. Once again all the default settings were used.

Example 1 *Ethanol Data*

The data consist of 87 observations obtained from an industrial experiment involving a one-cylinder engine using ethanol as a fuel. The response NOx is a measure of nitric oxide concentration in exhaust emissions. There are two predictors, E and C. E is the equivalence ratio, a measure of the fuel/air mixture while C is the compression ratio.

SIR	$Test1$	$Test2$
0.345	<0.0001	0.0003

Table 1: p-values for testing $k = 1$ against
the alternative $k > 1$ for model (1).

Table 1 summarises the results from SIR as well as tose from *Test1* and *Test2*. As discussed in Cook and Weisberg (1994, pp. 125-127), SIR has failed to detect the symmetric quadratic dependence of E on NOx, which is obvious in plots. On the other hand, *Test1* and *Test2* find very strong evidence that more than one linear combination of E and C is needed to adequately model NOx. In addition, these tests indicate that terms like higher order polynomial terms and interaction terms may be missing from the model.

The following regression model was fit to the data using least squares,

$$NOx = \gamma_0 + \gamma_1 C + \gamma_2 E + \gamma_3 C^2 + \gamma_4 E^2 + \gamma_5 C \cdot E + \epsilon. \tag{4}$$

Table 2 summarises the results. The interaction between C and E is highly significant as is the quadratic term in E. The adjusted R^2 value for model (4) is 84.2% while the corresponding figure is 0.0% when the model without the quadratic and interaction terms is fit.

Parameter	Estimate	t-ratio	p-value
γ_0	-24.26	-15.702	<0.001
γ_1	0.22	1.876	0.864
γ_2	56.76	20.592	<0.001
γ_3	0.00	0.565	0.574
γ_4	-29.79	-21.465	<0.001
γ_5	-0.24	-3.868	<0.001

Table 2: Least squares fit of model (4).

Example 2 *Australian Institute of Sport Data*

The data were obtained from 102 male and 100 female athletes at the Australian Institute of Sport. Interest centers on modeling LBM (lean body mass) as a function of Ht (height in centimetres), Wt (weight in kilograms) and RCC (red cell count). Following Cook and Weisberg (1994, pp. 122-125) we shall analyse the data for the female and male athletes separately.

Female Athletes

Table 3 summarises the results from SIR as well as those from *Test1* and *Test2* for the data on the 100 female athletes. The p-values reported for $Test1$ and $Test2$ are the minimum of those obtained from 3 pairwise tests. The row headed '1 point removed' gives the results when the case corresponding to the largest value of LBM is removed from the data, while the row headed '5 points removed' gives the results when the 5 cases marked with a × in Figure 8.5 of Cook and Weisberg (1994, p. 125) are removed from the data.

	SIR	*Test*1	*Test*2
All data	0.026	0.686	0.005
1 point removed	0.229	0.836	0.008
5 points removed	0.617	0.724	0.007

Table 3: p-values for testing $k = 1$ against the alternative $k > 1$ for model (1).

As discussed in Cook and Weisberg (1994, pp. 124-125), *SIR* is not robust to the effects of outliers. In this case removing just one of the 100 data points produces a 10 fold change in p-value obtained from *SIR*. On the other hand, *Test1* and *Test2* change little when a small number of data points are removed from the data. In addition, *Test2* finds strong evidence that more than one linear combination of *Ht*, *Wt* and *RCC* is needed to adequately model *LBM*. In addition, this test indicates that interaction terms may be missing from the model.

The following regression model was fit to the data using rank based regression,

$$LBM = \gamma_0 + \gamma_1 Ht + \gamma_2 Wt + \gamma_3 RCC + \gamma_4 Ht \cdot Wt + \gamma_5 Ht \cdot RCC + \gamma_6 Wt \cdot RCC + \epsilon. \tag{5}$$

Table 4 summarises the results when model (5) is fit to all 100 data points. In this case, the interaction between *Ht* and *Wt* is highly significant. When the 5 points referred to in Table 3 are removed and model (5) is refit, the p-value for the interaction between *Ht* and *Wt* increases to 0.056. Thus, some but not all of the significance of this interaction term is due to these 5 points.

Parameter	Estimate	t-ratio	p-value
γ_0	-61.400	-2.567	0.012
γ_1	0.719	2.958	0.004
γ_2	0.804	1.529	0.130
γ_3	-0.007	-0.134	0.894
γ_4	-0.006	-2.744	0.007
γ_5	-0.056	-1.518	0.132
γ_6	0.169	1.754	0.082

Table 4: Rank based regression fit of model (5) to all 100 female athletes.

Male Athletes

Table 5 summarises the results from SIR as well as those from *Test1* and *Test2* for the data on the 102 male athletes. The p-values reported for $Test1$ and $Test2$ are the minimum of those obtained from 3 pairwise tests. The row headed '2 points removed' gives the results when the cases corresponding to the two largest values of LBM are removed from the data.

	SIR	$Test1$	$Test2$
All data	0.017	0.950	0.111
2 points removed	0.173	0.831	0.108

Table 5: p-values for testing $k = 1$ against the alternative $k > 1$ for model (1).

These data also illustrate that SIR is not robust to the effects of outliers. In this case removing just two of the 102 data points produces a 10 fold change in p-value obtained from SIR. On the other hand, *Test1* and *Test2* change little when a small number of data points are removed from the data. In addition, neither *Test1* nor *Test2* finds strong evidence that more than one linear combination of Ht, Wt and RCC is needed to adequately model LBM.

The following regression model was fit to the data using rank based regression,

$$LBM = \gamma_0+\gamma_1 Ht+\gamma_2 Wt+\gamma_3 RCC+\gamma_4 Ht\cdot Wt+\gamma_5 Ht\cdot RCC+\gamma_6 Wt\cdot RCC+\epsilon. \tag{6}$$

Table 6 summarises the results when model (6) is fit to all 102 data points. In this case, none of the interactions are significant. When the 2 points referred to in Table 5 are removed and model (6) is refit, once again none of the interaction terms are significant.

Parameter	Estimate	t-ratio	p-value
γ_0	-69.870	-1.072	0.286
γ_1	0.590	1.344	0.182
γ_2	0.716	1.903	0.060
γ_3	6.650	0.544	0.588
γ_4	-0.002	-1.070	0.287
γ_5	-0.062	-0.749	0.228
γ_6	0.058	0.997	0.321

Table 6: Rank based regression fit of model (6) to all 102 male athletes.

In summary, for the athletes data the nonparametric procedure of Sheather and McKean (1997) seems to correctly identify the case (i.e., the male athletes) when a single linear combination of Ht, Wt and RCC adequately models LBM as well as the case (i.e., the female athletes) when more than one linear combination is needed. On the other hand, SIR indicates for both the female and male athletes that more than one linear combination is needed when all the data are used while it indicates the opposite when a small number of outliers have been removed from the data.

References

[1] Cook, R. and Weisberg, S. (1991). Comment on Li (1991). *J. Am. Statist. Assoc.* **86,** 328-332.

[2] Cook, R. D. , and Weisberg, S. (1994). *An Introduction to Regression Graphics.* New York: Wiley.

[3] Duan, N. and Li, K. C. (1991). Slicing regression: A link-free regression method. *Ann. Statist.* **19,** 505-530.

[4] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* London: Chapman and Hall.

[5] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* London: Chapman and Hall.

[6] Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Am. Statist. Assoc.* **86**, 316-342.

[7] Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Am. Statist. Assoc.* **87,** 1025-1039.

[8] Schott, J. R. (1994). Determining the dimensionality for sliced inverse regression. *J. Am. Statist. Assoc.* **89,** 141-148.

[9] Sheather, S. J. and McKean, J. W. (1997). A dimension reduction technique based on inverse regression. Unpublished manuscript.

[10] Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* New York: Springer.

[11] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* London: Chapman and Hall.