ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 141, 1997

*Thomas Kötter**

ASYMPTOTIC RESULTS FOR SLICED INVERSE REGRESSION

Abstract. It is well known that nonparametric regression techniques do not have good performance in high dimensional regression. However nonparametric regression is successful in one- or low-dimensional regression problems and is much more flexible than the parametric alternative. Hence, for high dimensional regression tasks one would like to reduce the regressor space to a lower dimension and then use nonparametric methods for curve estimation.

A possible dimension reduction approach is Sliced Inverse Regression (L i 1991). It allows to find a base of a subspace in the regressor space which still carries important information for the regression. The vectors spanning this subspace are found with a technique similar to Principal Component Analysis and can be judged with the eigenvalues that belong to these vectors. Asymptotic and simulation results for the eigenvalues and vectors are presented.

Key words: dimension reduction, inverse regression, linear projections.

1. INTRODUCTION

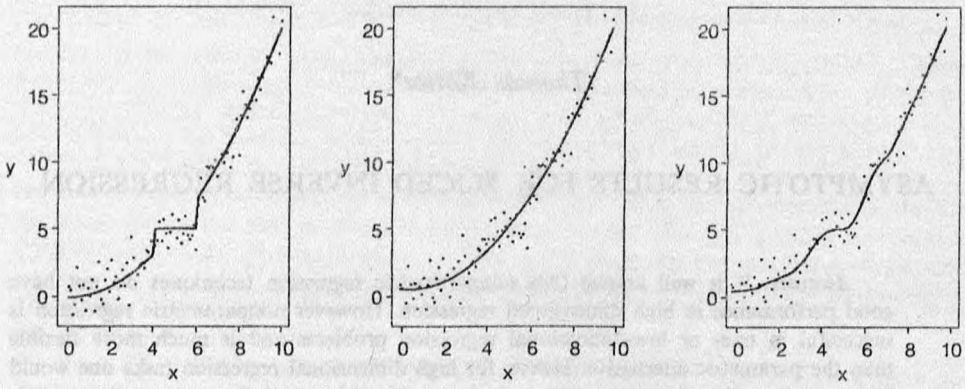
In this paper we will discuss some properties of a certain dimension reduction method. First the question arises: Why should the dimensionality be reduced? The reason is that we can do nonparametric regression in low dimensional spaces but not in high dimensions. And, of course, we want to do nonparametric regression.

Parametric regression has the crucial drawback that it can only fit a predefined model which has to be selected before. However, if this model is the true one, the properties of the estimates are good and well known.

Nonparametric regression allows the data to speak for themselves. There are not pressed into a coreselett like a predefined model. Hence, it is much more flexible than parametric regression.

* Humboldt-Universität zu Berlin, Wirtschaftswissenchaftliche Fakultät, Institut für Statistik und Ökonometrie, Spandauer Str. 1, D-10178 Berlin, e-mail: thomas@wiwi.huberlin.de.

Example:



The left plot shows the data with the true model. They are generated by $y_i = 5 + \varepsilon_i$, if $y_i \in [4, 6]$, and $y_i = x^2 + \varepsilon_i$; else, ε_i are standard normal. The data have a small plateau around $x = 5$ which cannot be found by the parametric fit of the model $y = a + b + cx^2$ as shown in the second plot. The Goodness-of-Fit criterion $s_{\hat{y}}^2 / s_y^2$ is close to one (0.966) although the plateau feature was not detected. The third plot consists of the data and a nonparamtric smoother (lowess, Cleveland 1979).

But also nonparametric regression has a crucial drawback. Due to local averaging, that is the main aspect of nonparametric methods, the performance in high dimensional spaces cannot be very good.

Example: (P. J. Huber) Assume we have a uniform distribution on a 10-dimensional unit ball with radius 1. Then 5% of the data lie in a ball with a radius of $0.05^{1/10} = 0.7411$. It is not possible to gather local features in this space except we have a huge dataset.

Now, the question is what we can do if we have the following situation.

$$Y = m(\beta_1^T X, \ldots, \beta_K^T X, \varepsilon) \tag{1}$$

with: $1 \leqslant K \leqslant d$, K – unknown, m: $IR^{K+1} \longrightarrow IR$ unknown, Y is a random variable, X is a IR^d random vector, ε is a random variable with $E[\varepsilon | x] = 0$.

As we do not know the dependence of Y on X it is not reasonable to choose a parametric approach. On the other hand d might be too large to use a nonparametric method directly (e.g. smoothing).

Hence, we want to reduce the dimensionality and then use a flexible nonparametric regression algorithm. Here we will focus on the first task.

Remarks to model (1):

1. We do not allow redundancy in the representation of m so without loss of generality we can assume that the β_i, $i = 1, \ldots, K$, are linear independent.

2. Neither the length nor the direction of the β_i, $i = 1, ..., K$, are identifiable. Only the space which is spanned by these vectors can be identified.

3. Model (1) includes models of the shape $y = \sum_{i=1}^{K} m_i(\beta_i^T x) + \varepsilon$ where $m_i: IR \longrightarrow IR$, but it is more general. Of course, here the directions of the β_i, $i = 1, ..., K$, can be identified.

4. The β_i', s are called effective dimension reduction directions (edr--directions).

5. Conditioned on $\beta_1^T X, ..., \beta_K^T X$, X and Y are independent.

6. Y depends on X only through $\beta_1^T X, ..., \beta_K^T X$, i.e. $F_{Y/\beta_1^T X, ..., \beta_K^T X} = F_{Y/X}$.

Sliced Inverse Regression (SIR) is able to work with model (1). It delivers d directions β_i and due to the algorithm d (eigen-) values λ_i with which the importance of the single β_i can be judged.

Furthermore, it is possible to establish asymptotic normality of

$$\hat{\Psi}_K = \frac{\sum_{i=1}^{K} \hat{\lambda}_i}{\sum_{i=1}^{d} \hat{\lambda}_i}$$

which can be interpreted as the ratio of the variance which is declared by the first K edr-directions.

So this statistic $\hat{\Psi}$ helps us to find the number of β_i which has to be taken into account i.e. how big K is.

2. SLICED INVERSE REGRESSIONS

Theorem. Given the model (1) and the assumption

$$\forall b \in IR^d \text{ gilt: } E[b^T X | \beta_1^T X = \beta_1^T x, ..., \beta_k^T X = \beta_k^T x] = c_0 + \sum_{i=1}^{K} c_i \beta_i^T X \qquad (2)$$

the centered inverse regression curve $E[X | Y = y] - E[X]$ lies in the linear subspace spanned by vectors $\sum_{XX} \beta_i$, $1, ..., K$.

Sketch of the proof:

Without loss of generality $E[X] = 0$

It is sufficient to show that $\forall b \in IR^d$:

$$b^T \sum_{XX} \beta_i = 0 \Rightarrow b^T E[X | Y = y] = 0$$

With the abbreviation $E[X|y] := E[X|Y=y]$ and using $E[X] = E[E[X|Y=y]]$.

$$E[X|y] = E[E[X|y|\beta_1^T x, ..., \beta_K^T x, y]|y]$$
$$= E[E[X|\beta_1^T x, ..., \beta_K^T x]|y]$$

Further it is

$$E[b^T X|\beta_1^T x, ..., \beta_K^T x] = 0$$
$$\Leftrightarrow E[E^2[b^T X|\beta_1^T x, ..., \beta_K^T x]] = 0$$

Finally

$$E[E^2[...]] = E[E[...]E[...]]$$
$$= E[E[E[b^T X|\beta_1^T x, ..., \beta_K^T x X^T b|\beta_1^T x, ..., \beta_K^T x]]$$
$$= E[(c_0 + \sum_{i=1}^{K} c_i \beta_i^T X) X^T b]$$
$$= E[c_0 X^T b] + \sum_{i=1}^{K} c_i \beta_i^T \sum_{XX} b$$
$$= 0 + 0$$

because of the assumption. QED

Remark. The assumption (2) made above is equivalent to the fact that the distribution of X is elliptical symmetric (Cook, Weisberg 1991). It can be weakened as Hall and Li (1993) showed. Another approach to find interesting subspaces is SIR II, which investigates the inverse covariance structure (Cook, Weisberg 1991; Li 1991). The implementation and application of the SIR algorithms can be found in Kötter (1995).

Corollar. Let Z be the standardized random vector with $Z = \sum_{XX}^{-1/2}(X - E[X])$. Then $E[Z|y]$ lies in the space which is spanned by $\eta_i = \sum_{XX}^{1/2} \beta_i$.

Now it is easy to see that from $b^T \eta_i = 0$ it follows that $E[b^T Z|y] = 0$ and that the conditional covariance $Cov[E[Z|y]]$ is degenerated to each direction orthogonal to the η_i.

So an algorithm to find edr-directions is to standardize X then to estimate $E[Z|y]$ and $Cov[E[Z|y]]$. Conduct a eigenvalue/eigenvector decomposition, choose the eigenvectors to the largest eigenvalues and scale back to the original scale. This retransformed eigenvectors are estimators for the edr-directions.

3. ALGORITHM

First some notations: X, Y and Z are data matrices, not random vectors. The observations are in the rows. Single observations are signed by small letters. The sample of size n is $\{x_i, y_i\}_{i=1}^n$.

$$x_i = (x_{i1}, \ldots, x_{id})^T,$$
$$X = (x_1, \ldots, x_d)^T,$$
$$Y = (y_1, \ldots, y_n)^T,$$
$$\bar{X} := \frac{1}{n} 1_n 1_n^T X, \qquad 1_n := \underbrace{(1, \ldots, 1)^T}_{n},$$
$$\hat{\Sigma}_{XX} := \frac{1}{n-1}(X^T X - n\bar{X}\bar{X}^T)$$

Estimate the edr-directions with

1. Standardize the x values:

$$z_i := \hat{\Sigma}_{XX}^{-1/2}(x_i - \bar{x}) \text{ or } Z := (X - \bar{X})\hat{\Sigma}_{XX}^{-1/2}$$

2. Divide the range of y_i in S non overlapping slices H_s, n_s denotes the number of observations within slice S_s.

$$n_s := \sum_{i=1}^{n} I_{H_s}(y_i)$$

3. Compute the mean of z_i over all slices.

$$\bar{z}_s := \frac{1}{n_s} \sum_{i=1}^{n} z_i I_{H_s}(y_i)$$

4. Calculate the weighted covariance matrix.

$$\hat{V} := n^{-1} \sum_{i=1}^{n} n_s \bar{z}_s \bar{z}_s^T$$

5. Identify the eigenvalues $\hat{\lambda}_i$ and eigenvectors $\hat{\eta}_i$ of $\hat{V}$.

6. Transform the standardized edr-directions $\hat{\eta}_i$ back to the original scale. Now the estimates for the edr-directions are given by:

$$\hat{\beta}_i := \hat{\Sigma}_{XX}^{-1/2} \hat{\eta}_i$$

3.1. Costs of Computation

The following table shows the costs of different steps of the algorithm. In the costs column the terms are the order of the O function.

Costs	Cause
nd	Mean $\bar{x}$
nd^2	Covariave Σ_{XX}
d^3	$\sum_{XX}^{-1/2}$
$nd + nd^2$	Standardize the matrix X to Z
Sn	Computation of n_s and $\bar{z}_s$
Sd^2	$\hat{V}$
d^3	Eigendecomposition of $\hat{V}$
d^3	Rescaling to the edr-directions β_i

The sum of the costs is of order $O(nd^2 + Sn + d^3)$. As we discuss later it is convenient to choose $S = O(n)$, so the sum is dominated by n^2 if d is constant. It can be reduced to $O(n\log(n))$ if the data are sorted before slicing. Sorting needs $O(n\log(n))$, then slicing the only $O(n)$.

This is a very good behaviour regarding the sample size n. Other nonparametric methods often have to be treated very tricky to achieve rates below $O(n^2)$ (e.g. WARPing by kernel density estimation).

4. STATISTICAL PROPERTIES

It is possible to find a $\sqrt{S}$-consistent estimate for $Cov[E[X|y]]$. With $\hat{V}$ calculated from the algorithm, define

$$\hat{V}^*: = \frac{n_s}{n_s - 1}\hat{V} - \frac{1}{n_s - 1}I_d$$

This estimator is $\sqrt{S}$-consistent for $Cov[E[X|y \in H_s]]$ and as S goes to infinity for $Cov[E[X|y]]$.

It is easy to see that it is necessary that $S = O(n)$ to achieve $\sqrt{n}$-consistency for the estimates. In other words the number of elements within each slice should be constant. In the following we assume that $n_s = n/S$.

4.1. Asymptotic normality

Some asymptotic results can be derived:

- asymptotic normality of $uvec(\hat{V})$

$$uvec(A): = (a_{11}, \ldots, a_{1d}, a_{22}, \ldots, a_{2d}, a_{33}, \ldots, a_{dd})^T$$

- asymptotic normality of the vector $\sqrt{S}(\hat{\lambda}_1, \ldots, \hat{\lambda}_d)_{i=1}^d$
- asymptotic normality of $\hat{\Psi}_K$.

An important condition to show this asymptotics is that $\hat{\Sigma}_{XX}$ and Z have to be independent. In applications the data set has to be split. With one part $\hat{\Sigma}_{XX}$ is estimated, the other part is standardized by using this $\hat{\Sigma}_{XX}$.

Terms of the shape $Cov[\tilde{\sigma}_{aj}\tilde{\sigma}_{bk}, \tilde{\sigma}_{cl}\tilde{\sigma}_{dm}]$ (with $\tilde{\sigma}_{ij}$ is an element of $\hat{\Sigma}_{XX}^{-1/2}$) appear within the computations of the asymptotic covariance matrix. Unfortunately, these terms are of the same order ($O_p(1/\sqrt{S})$) as the asymptotic covariance itself. In order to overcome this problem, the computation of two independent estimates for $\hat{\Sigma}_{XX}^{-1/2}$ has to be done.

4.2. Main Idea of the Proofs

– since the slices are disjoint the elements of $\hat{V}$ v_{ij} can be written as a sum of S independent terms. This yields to asymptotic normality.

– with the *Cramer-Wold-device* the asymptotic distribution of $uvec(\hat{V})$ can then be shown,

– since the eigenvalues are continuous in the elements of the matrix (Theorem by *Wielandt-Hoffmann*, Wilkinson 1965) thus the eigenvalues are also $\sqrt{S}$-consistent.

– the asymptotic distribution of $(\lambda_i)_{i=1}^d$ can be derived by taking a connection between the asymptotic distribution of the characteristic polynom of $\hat{V}(|\hat{V}-\lambda I|)$ and the eigenvalues λ_i.

– the asymptotic normality of $\hat{\Psi}_K$ can then be shown by using the same technique as for principal component analysis (Mardia et al. 1979).

4.3. Asymptotic Expectation and Covariance

In this section only the formulae for the asymptotic expectation and covariance for the random vector $uvec(\hat{V})$ are given. For the latter the computation is long and tedious (Kötter 1990).

Expectation. As the above mentioned estimate for $uvec(Cov[E[Z|y]])$ is $\sqrt{S}$-consistent, the asymptotic expectation is $Cov[E[Z|y]]$.

Covariance. The asymptotic covariance structure of $uvec(Cov[E[Z|y]])$ is:

$$\lim_{S\to\infty} Cov[\hat{v}_{ab}, \hat{v}_{cd}] = \frac{1}{S}\sum_{j=1}^{d}\sum_{k=1}^{d}\sum_{l=1}^{d}\sum_{m=1}^{d} E[\tilde{\sigma}_{aj}\tilde{\sigma}_{bk}, \tilde{\sigma}_{cl}\tilde{\sigma}_{dm}]$$

$$\{E[x_p^j x_{p'}^k x_q^i x_{q'}^m] + E[x^j]E[x^k]E[x^l]E[x^m] - \frac{1}{n_s}E[x^l]E[x^m]$$

$$(Cov[x^j, x^k] - Cov[E[x^j|y]E[x^k|y]] + n_s E[E[x^j|y]E[x^k|y]]) - \frac{1}{n_s} E[x^j]E[x^k]$$
$$(Cov[x^l, x^m] - Cov[E[x^l|y]E[x^m|y]] + n_s E[E[x^l|y]E[x^m|y]])$$
$$+ \left((1 - \frac{1}{n_s}) Cov[E[x^j|y], E[x^k|y]] + \frac{1}{n_s} Cov[x^j, x^k] \right)$$
$$\times \left((1 - \frac{1}{n_s}) Cov[E[x^l|y], E[x^m|y]] + \frac{1}{n_s} Cov[x^l, x^m] \right) \Big\}$$

where $\tilde{\sigma}_{ij}$ is the (i, j) elements of $\Sigma_{XX}^{-1/2}$.

Define $\Sigma^* = n_s^2/(n_s - 1)^2 \times \lim_{s \to \infty} Cov[uvec(\hat{V})]$ then the following asymptotic result holds:

$$\sqrt{S}((\lambda_i - \hat{\lambda}_i)_{i=1}^d) \sim AN(0, D\Sigma^* D^T)$$

with $D := (uvec([\hat{V}^* - \lambda_i I_d]^{-1} | \hat{V}^* - \lambda_i I_d | / D'_V(\lambda_i))_{i=1}^d \in IR^{d \times d(d+1)/2}$

Furthermore, with $\Sigma = D\Sigma^* D^T$ the asymptotic distribution of $\hat{\Psi}$ is given by:

$$\hat{\Psi}_K := \frac{\Sigma_{i=1}^K \hat{\lambda}_i}{tr(\hat{V}^*)} \sim AN(\Psi_K, B\Sigma B^T)$$

with $B := (\partial \hat{\Psi}_K / \partial \hat{\lambda}_1, \ldots, \partial \hat{\Psi}_K / \partial \hat{\lambda}_d)$.

5. NUMERICAL EXAMPLE

The data are generated by the model

$$y_i = e^{(x_1 + x_2 + x_3)/2} \times (x_1 - x_2 - x_3) + \varepsilon_i$$

ε is standard normal distributed. Sample size $n = 200$.

SIR gives with 10 elements in each slice for the der-directions:

−0.3496	0.9327	−0.0419
0.6759	0.2750	−0.7049
0.6488	0.2333	0.7081

The eigenvalues were (0.6032, 0.3317, 0.1012) and the corresponding $\hat{\Psi} = (0.5822, 0.9023, 1)$, i.e. over 90% of the variance is declared by the first two edr-directions.

The third edr-direction $\hat{\beta}_3$ is nearly parallel to the vector which is orthogonal to the design plane $span((1, 1, 1)^T, (1, -1, -1)^T)$. The normal inner product of $e_3 = (0, -1, 1)^T/\sqrt{(2)}$ is $\beta_3^T e_3 / \| \beta_3 \| \| e_3 \| = 0.99912$.

Additionally, the third eigenvalue $\lambda_3 = 0.1012$ is much smaller than the first two ones; SIR performs very well in this example.

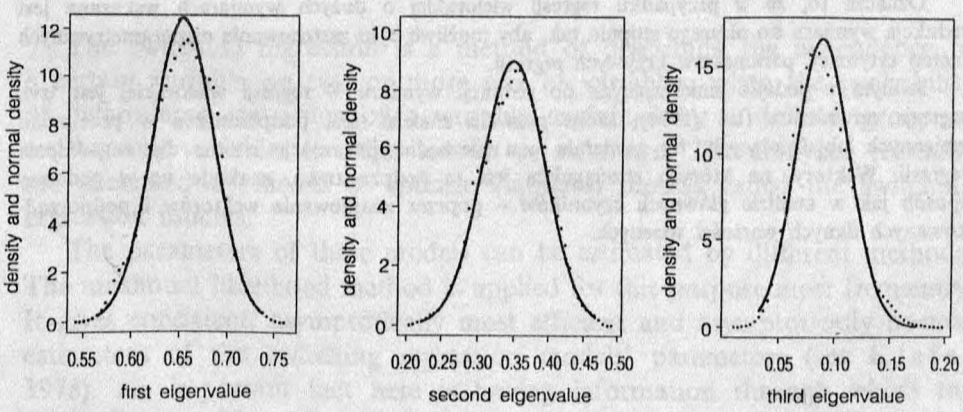
The simple setting in of estimates into the asymptotic formulae yields to estimates for the variance of $\hat{\Psi}$ which seem to be very sensitive to the generation of the subsamples and the subsample sizes. Here some work remains which has to be done in the future, how to estimate the asymptotic Covariance of $E[Z|y]$.

6. SIMULATION

With the same model as above, data were 500 times generated and SIR was conducted. In the following plots you see the smoothed density of each eigenvalue. They are very similar to the normal density which is also plotted (with the same mean and variance). It is remarkable that the variances of the eigenvalues are quite small.

Mean and variance of the simulation:

$$\bar{\lambda} = (0.6541,\ 0.3539,\ 0.0886)^T \quad Var[\lambda] = (0.0010,\ 0.0019,\ 0.0006)^T$$



REFERENCES

Cleveland W. S. (1979): *Robust locally weighted regression and smoothing scatterplots*, „Journal of the American Statistical Association", 74, p. 829–836.

Cook R. D., Weisberg S. (1991): *Comments on Sliced Inverse Regression for Dimension Reduction*, „Journal of the American Statistical Association", 86, p. 328–332.

Hall O., Li K. C. (1993): *On almost linearity of low dimensional projections from high dimensional data*, „Annals of Statistics", 21, No 2, p. 867–889.
Li K. C. (1991): *Sliced Inverse Regression for Dimension Reduction*, „Journal of the American Statistical Association", 86, p. 316–327.
Kötter T. (1990): *Regression mitt unbekannter Linkfunktion*, „Diplomarbeit, Fachbereich Statistik, Universität Dortmund.
Kötter T. (1995): *An Asymptotic Result for Sliced Inverse Regression*, „Computational Statistics", (to appear).
Mardia K. V., Kent J. T., Bibby J. M. (1979): *Multivariate Analysis*, „Academic Press", London.
Wilkinson J. H. (1965): *The Algebraic Eigenvalue Problem*, „Oxford University Press".

Thomas Kötter

ASYMPTOTYCZNE REZULTATY DLA „SLICED INVERSE REGRESSION"

Jest rzeczą wiadomą, że techniki regresji nieparametrycznej nie funkcjonują właściwie w przypadku regresji wielowymiarowej. Jednakże są to techniki działające skutecznie w przypadku regresji jednowymiarowej bądź o małej liczbie wymiarów, a ponadto są bardziej elastyczne niż ich parametryczne odpowiedniki.

Oznacza to, że w przypadku regresji wielorakiej o dużych wymiarach wskazana jest redukcja wymiaru do niższego stopnia tak, aby możliwe było zastosowanie nieparametrycznych metod estymacji parametrów krzywych regresji.

Jednym z podejść zmierzających do redukcji wymiaru w regresji wielorakiej jest tzw. regresja odwrócona (Li (1991), która pozwala znaleźć taką podprzestrzeń w przestrzeni zmiennych objaśniających, by zawierała ona niezbędne informacje istotne dla zagadnienia regresji. Wektory, na których rozciągnięta jest ta podprzestrzeń, znajduje się w podobny sposób jak w analizie głównych czynników – poprzez znajdowanie wektorów i podporządkowanych danych wartości własnych.