

18.3 Sliced Inverse Regression

Sliced inverse regression (SIR) is a dimension reduction method proposed by Duan and Li (1991). The idea is to find a smooth regression function that operates on a variable set of projections. Given a response variable Y and a (random) vector $X \in \mathbb{R}^p$ of explanatory variables, SIR is based on the model:

$$Y = m(\beta_1^\top X, \dots, \beta_k^\top X, \varepsilon), \quad (18.10)$$

where β_1, \dots, β_k are unknown projection vectors, k is unknown and assumed to be less than p , $m : \mathbb{R}^{k+1} \rightarrow \mathbb{R}$ is an unknown function, and ε is the noise random variable with $E(\varepsilon | X) = 0$.

Model (18.10) describes the situation where the response variable Y depends on the p -dimensional variable X only through a k -dimensional subspace. The unknown β_i 's, which span this space, are called *effective dimension reduction directions* (EDR-directions). The span is denoted as *effective dimension reduction space* (EDR-space). The aim is to estimate the base vectors of this space, for which neither the length nor the direction can be identified. Only the space in which they lie is identifiable.

SIR tries to find this k -dimensional subspace of \mathbb{R}^p which under the model (18.10) carries the essential information of the regression between X and Y . SIR also focuses on small k , so that nonparametric methods can be applied for the estimation of m . A direct application of nonparametric smoothing to X is for high dimension p generally not possible due to the sparseness of the observations. This fact is well known as the *curse of dimensionality*, see Huber (1985).

The name of SIR comes from computing the inverse regression (IR) curve. That means instead of looking for $E(Y | X = x)$, we investigate $E(X | Y = y)$, a curve in \mathbb{R}^p consisting of p one-dimensional regressions. What is the connection between the IR and the SIR model (18.10)? The answer is given in the following theorem from Li (1991).

THEOREM 18.1 *Given the model (18.10) and the assumption*

$$\forall b \in \mathbb{R}^p : E(b^\top X | \beta_1^\top X = \beta_1^\top x, \dots, \beta_k^\top X = \beta_k^\top x) = c_0 + \sum_{i=1}^k c_i \beta_i^\top x, \quad (18.11)$$

the centered IR curve $E(X | Y = y) - E(X)$ lies in the linear subspace spanned by the vectors $\Sigma \beta_i$, $i = 1, \dots, k$, where $\Sigma = \text{Cov}(X)$.

Assumption (18.11) is equivalent to the fact that X has an elliptically symmetric distribution, see Cook and Weisberg (1991). Hall and Li (1993) have shown that assumption (18.11) only needs to hold for the EDR-directions.

It is easy to see that for the standardized variable $Z = \Sigma^{-1/2}\{X - E(X)\}$ the IR curve $m_1(y) = E(Z | Y = y)$ lies in $\text{span}(\eta_1, \dots, \eta_k)$, where $\eta_i = \Sigma^{1/2}\beta_i$. This means that the conditional expectation $m_1(y)$ is moving in $\text{span}(\eta_1, \dots, \eta_k)$ depending on y . With b orthogonal to $\text{span}(\eta_1, \dots, \eta_k)$, it follows that

$$b^\top m_1(y) = 0,$$

and further that

$$m_1(y)m_1(y)^\top b = \text{Cov}\{m_1(y)\}b = 0.$$

As a consequence $\text{Cov}\{E(Z | y)\}$ is degenerated in each direction orthogonal to all EDR-directions η_i of Z . This suggests the following algorithm.

First, estimate $\text{Cov}\{m_1(y)\}$ and then calculate the orthogonal directions of this matrix (for example, with eigenvalue/eigenvector decomposition). In general, the estimated covariance matrix will have full rank because of random variability, estimation errors and numerical imprecision. Therefore, we investigate the eigenvalues of the estimate and ignore eigenvectors having small eigenvalues. These eigenvectors $\hat{\eta}_i$ are estimates for the EDR-direction η_i of Z . We can easily rescale them to estimates $\hat{\beta}_i$ for the EDR-directions of X by multiplying by $\hat{\Sigma}^{-1/2}$, but then they are not necessarily orthogonal. SIR is strongly related to PCA. If all of the data falls into a single interval, which means that $\widehat{\text{Cov}}\{m_1(y)\}$ is equal to $\widehat{\text{Cov}}(Z)$, SIR coincides with PCA. Obviously, in this case any information about y is ignored.

The SIR Algorithm

The algorithm to estimate the EDR-directions via SIR is as follows:

1. Standardize x :

$$z_i = \hat{\Sigma}^{-1/2}(x_i - \bar{x}).$$

2. Divide the range of y_i into S nonoverlapping intervals (*slices*) H_s , $s = 1, \dots, S$. n_s denotes the number of observations within slice H_s , and \mathbf{I}_{H_s} the indicator function for this slice:

$$n_s = \sum_{i=1}^n \mathbf{I}_{H_s}(y_i).$$

3. Compute the mean of z_i over all slices. This is a crude estimate \hat{m}_1 for the *inverse regression curve* m_1 :

$$\bar{z}_s = \frac{1}{n_s} \sum_{i=1}^n z_i \mathbf{I}_{H_s}(y_i).$$

4. Calculate the estimate for $Cov\{m_1(y)\}$:

$$\hat{V} = n^{-1} \sum_{s=1}^S n_s \bar{z}_s \bar{z}_s^\top.$$

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

$$\hat{\beta}_i = \hat{\Sigma}^{-1/2} \hat{\eta}_i.$$

REMARK 18.1 *The number of different eigenvalues unequal to zero depends on the number of slices. The rank of \hat{V} cannot be greater than the number of slices–1 (the z_i sum up to zero). This is a problem for categorical response variables, especially for a binary response—where only one direction can be found.*

SIR II

In the previous section we learned that it is interesting to consider the IR curve, that is, $E(X|y)$. In some situations however SIR does not find the EDR-direction. We overcome this difficulty by considering the conditional covariance $Cov(X|y)$ instead of the IR curve. An example where the EDR directions are not found via the SIR curve is given below.

EXAMPLE 18.2 *Suppose that $(X_1, X_2)^\top \sim N(0, \mathcal{I}_2)$ and $Y = X_1^2$. Then $E(X_2|y) = 0$ because of independence and $E(X_1|y) = 0$ because of symmetry. Hence, the EDR-direction $\beta = (1, 0)^\top$ is not found when the IR curve $E(X|y) = 0$ is considered.*

The conditional variance

$$Var(X_1|Y = y) = E(X_1^2|Y = y) = y,$$

offers an alternative way to find β . It is a function of y while $Var(X_2|y)$ is a constant.

The idea of SIR II is to consider the conditional covariances. The principle of SIR II is the same as before: investigation of the IR curve (here the conditional covariance instead of the conditional expectation). Unfortunately, the theory of SIR II is more complicated. The assumption of the elliptical symmetrical distribution of X has to be more restrictive, i.e., assuming the normality of X .

Given this assumption, one can show that the vectors with the largest distance to $Cov(Z | Y = y) - E\{Cov(Z | Y = y)\}$ for all y are the most interesting for the EDR-space. An appropriate measure for the overall mean distance is, according to Li (1992),

$$E \left(\| [Cov(Z | Y = y) - E\{Cov(Z | Y = y)\}] b \|^2 \right) = \quad (18.12)$$

$$= b^T E \left(\| Cov(Z | y) - E\{Cov(Z | y)\} \|^2 \right) b. \quad (18.13)$$

Equipped with this distance, we conduct again an eigensystem decomposition, this time for the above expectation $E \left(\| Cov(Z | y) - E\{Cov(Z | y)\} \|^2 \right)$. Then we take the rescaled eigenvectors with the largest eigenvalues as estimates for the unknown EDR-directions.

The SIR II Algorithm

The algorithm of SIR II is very similar to the one for SIR, it differs in only two steps. Instead of merely computing the mean, the covariance of each slice has to be computed. The estimate for the above expectation (18.12) is calculated after computing all slice covariances. Finally, decomposition and rescaling are conducted, as before.

1. Do steps 1 to 3 of the SIR algorithm.
2. Compute the slice covariance matrix \hat{V}_s :

$$\hat{V}_s = \frac{1}{n_s - 1} \sum_{i=1}^n I_{H_s}(y_i) z_i z_i^T - n_s \bar{z}_s \bar{z}_s^T.$$

3. Calculate the mean over all slice covariances:

$$\bar{V} = \frac{1}{n} \sum_{s=1}^S n_s \hat{V}_s.$$

4. Compute an estimate for (18.12):

$$\hat{V} = \frac{1}{n} \sum_{s=1}^S n_s \left(\hat{V}_s - \bar{V} \right)^2 = \frac{1}{n} \sum_{s=1}^S n_s \hat{V}_s^2 - \bar{V}^2.$$

5. Identify the eigenvectors and eigenvalues of \hat{V} and scale back the eigenvectors. This gives estimates for the SIR II EDR-directions:

$$\hat{\beta}_i = \hat{\Sigma}^{-1/2} \hat{\eta}_i.$$

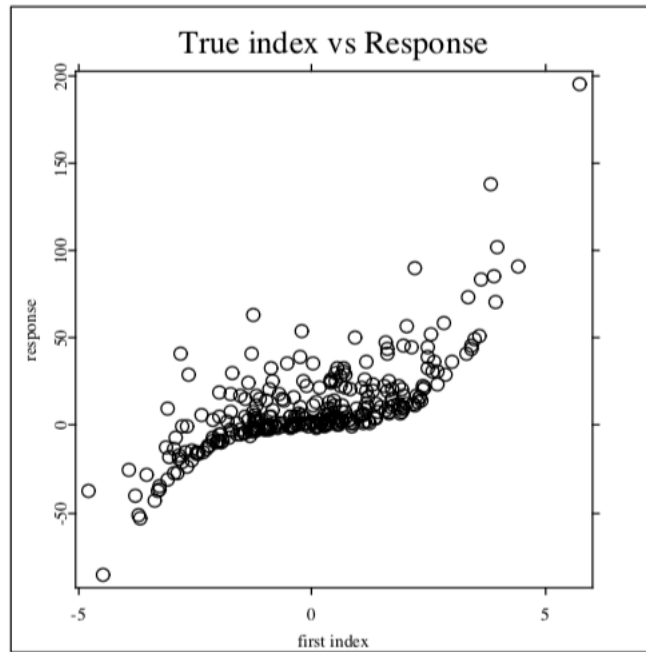


Figure 18.4. Plot of the true response versus the true indices. The monotonic and the convex shapes can be clearly seen. [MVAsirdata.xpl](#)

EXAMPLE 18.3 The result of SIR is visualized in four plots in Figure 18.6: the left two show the response variable versus the first respectively second direction. The upper right plot consists of a three-dimensional plot of the first two directions and the response. The last picture shows $\hat{\Psi}_k$, the ratio of the sum of the first k eigenvalues and the sum of all eigenvalues, similar to principal component analysis.

The data are generated according to the following model:

$$y_i = \beta_1^\top x_i + (\beta_1^\top x_i)^3 + 4(\beta_2^\top x_i)^2 + \varepsilon_i,$$

where the x_i 's follow a three-dimensional normal distribution with zero mean, the covariance equal to the identity matrix, $\beta_2 = (1, -1, -1)^\top$, and $\beta_1 = (1, 1, 1)^\top$. ε_i is standard, normally distributed and $n = 300$. Corresponding to model (18.10), $m(u, v, \varepsilon) = u + u^3 + v^2 + \varepsilon$. The situation is depicted in Figure 18.4 and Figure 18.5.

Both algorithms were conducted using the slicing method with 20 elements in each slice. The goal was to find β_1 and β_2 with SIR. The data are designed such that SIR can detect β_1

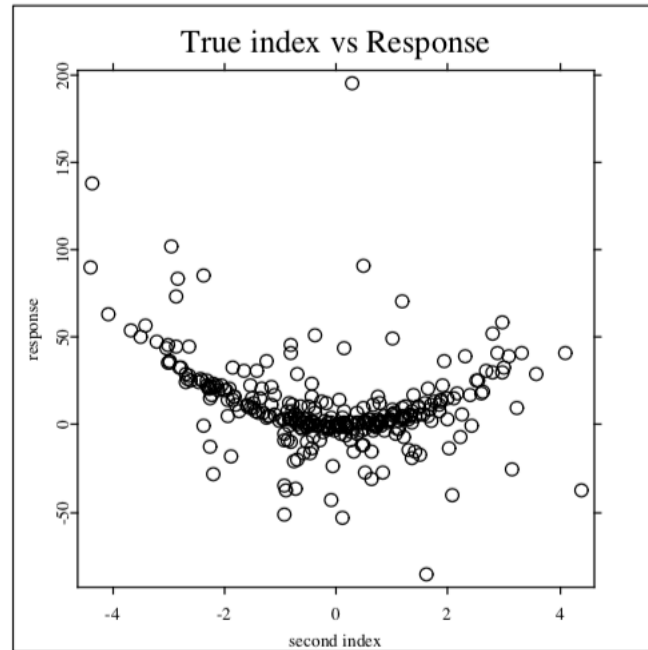


Figure 18.5. Plot of the true response versus the true indices. The monotonic and the convex shapes can be clearly seen. [MVAsirdata.xpl](#)

$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
0.578	-0.723	-0.266
0.586	0.201	0.809
0.568	0.661	-0.524

Table 18.5. SIR: EDR-directions for simulated data.

because of the monotonic shape of $\{\beta_1^\top x_i + (\beta_1^\top x_i)^3\}$, while SIR II will search for β_2 , as in this direction the conditional variance on y is varying.

If we normalize the eigenvalues for the EDR-directions in Table 18.5 such that they sum up to one, the resulting vector is $(0.852, 0.086, 0.062)$. As can be seen in the upper left plot of Figure 18.6, there is a functional relationship found between the first index $\hat{\beta}_1^\top x$ and the response. Actually, β_1 and $\hat{\beta}_1$ are nearly parallel, that is, the normalized inner product

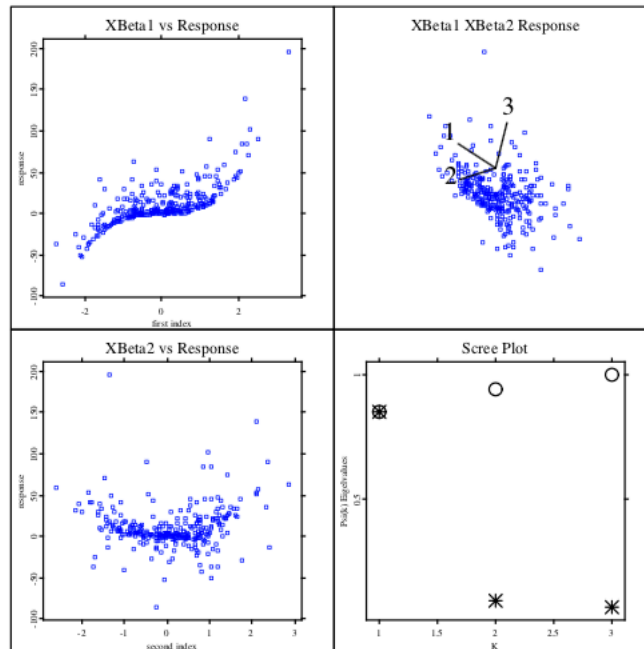


Figure 18.6. SIR: The left plots show the response versus the estimated EDR-directions. The upper right plot is a three-dimensional plot of the first two directions and the response. The lower right plot shows the eigenvalues $\hat{\lambda}_i$ (*) and the cumulative sum (\circ). [MVAsirdata.xpl](#)

$\hat{\beta}_1^\top \beta_1 / \{\|\hat{\beta}_1\| \|\beta_1\|\} = 0.9894$ is very close to one.

The second direction along β_2 is probably found due to the good approximation, but SIR does not provide it clearly, because it is “blind” with respect to the change of variance, as the second eigenvalue indicates.

For SIR II, the normalized eigenvalues are (0.706, 0.185, 0.108), that is, about 69% of the variance is explained by the first EDR-direction (Table 18.6). Here, the normalized inner product of β_2 and $\hat{\beta}_1$ is 0.9992. The estimator $\hat{\beta}_1$ estimates in fact β_2 of the simulated model. In this case, SIR II found the direction where the second moment varies with respect to $\beta_2^\top x$.

In summary, SIR has found the direction which shows a strong relation regarding the conditional expectation between $\beta_1^\top x$ and y , and SIR II has found the direction where the conditional variance is varying, namely, $\beta_2^\top x$.

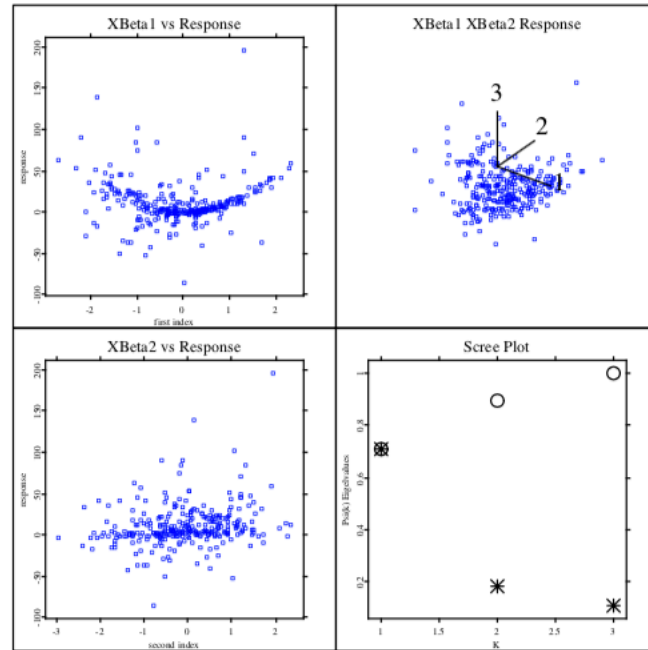



Figure 18.7. SIR II mainly sees the direction β_2 . The left plots show the response versus the estimated EDR-directions. The upper right plot is a three-dimensional plot of the first two directions and the response. The lower right plot shows the eigenvalues $\hat{\lambda}_i$ (*) and the cumulative sum (o).
[MVA_{sir2}data.xpl](#)

$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
0.821	0.180	0.446
-0.442	-0.826	0.370
-0.361	-0.534	0.815

Table 18.6. SIR II: EDR-directions for simulated data.

The behavior of the two SIR algorithms is as expected. In addition, we have seen that it is worthwhile to apply both versions of SIR. It is possible to combine SIR and SIR II (Cook and Weisberg, 1991; Li, 1991; Schott, 1994) directly, or to investigate higher conditional

moments. For the latter it seems to be difficult to obtain theoretical results. For further details on SIR see Kötter (1996).

 Summary
↪ SIR serves as a dimension reduction tool for regression problems.
↪ Inverse regression avoids the <i>curse of dimensionality</i> .
↪ The dimension reduction can be conducted without estimation of the regression function $y = m(x)$.
↪ SIR searches for the effective dimension reduction (EDR) by computing the inverse regression IR.
↪ SIR II bases the EDR on computing the inverse conditional variance.
↪ SIR might miss EDR directions that are found by SIR II.

18.4 Boston Housing

Coming back to the Boston housing data set, we compare the results of exploratory projection pursuit on the original data \mathcal{X} and the transformed data $\hat{\mathcal{X}}$ motivated in Section 1.8. So we exclude X_4 (indicator of Charles River) from the present analysis.

The aim of this analysis is to see from a different angle whether our proposed transformations yield more normal distributions and whether it will yield data with less outliers. Both effects will be visible in our projection pursuit analysis.

We first apply the Jones and Sibson index to the non-transformed data with 50 randomly chosen 13-dimensional directions. Figure 18.8 displays the results in the following form. In the lower part, we see the values of the Jones and Sibson index. It should be constant for 13-dimensional normal data. We observe that this is clearly not the case. In the upper part of Figure 18.8 we show the standard normal density as a green curve and two densities corresponding to two extreme index values. The red, slim curve corresponds to the maximal value of the index among the 50 projections. The blue curve, which is close to the normal, corresponds to the minimal value of the Jones and Sibson index. The corresponding values of the indices have the same color in the lower part of Figure 18.8. Below the densities, a jitter plot shows the distribution of the projected points $\alpha^T x_i$ ($i = 1, \dots, 506$). We conclude from the outlying projection in the red distribution that several points are in conflict with the normality assumption.

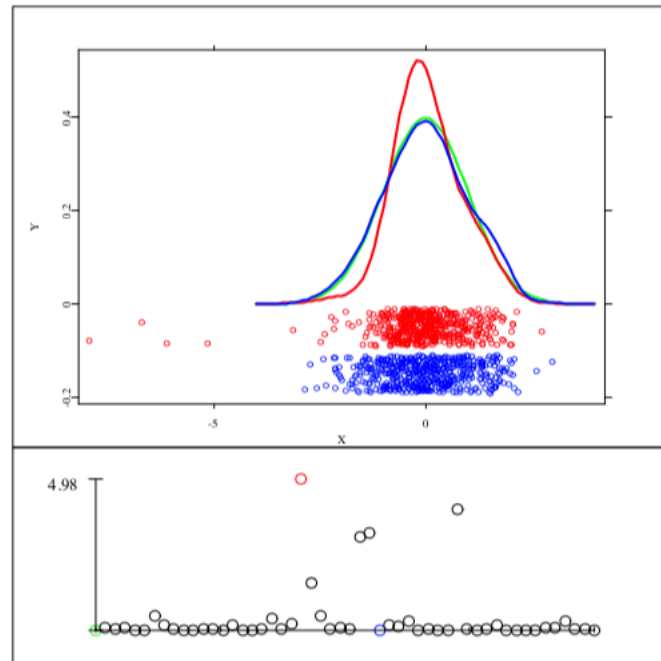


Figure 18.8. Projection Pursuit with the Sibson-Jones index with 13 original variables. [MVAppsib.xpl](#)

Figure 18.9 presents an analysis with the same design for the transformed data. We observe in the lower part of the figure values that are much lower for the Jones and Sibson index (by a factor of 10) with lower variability which suggests that the transformed data is closer to the normal. (“Closeness” is interpreted here in the sense of the Jones and Sibson index.) This is confirmed by looking to the upper part of Figure 18.9 which has a significantly less outlying structure than in Figure 18.8.

18.5 Exercises

EXERCISE 18.1 Calculate the *Simplicial Depth* for the Swiss bank notes data set and compare the results to the univariate medians. Calculate the *Simplicial Depth* again for the genuine and counterfeit bank notes separately.

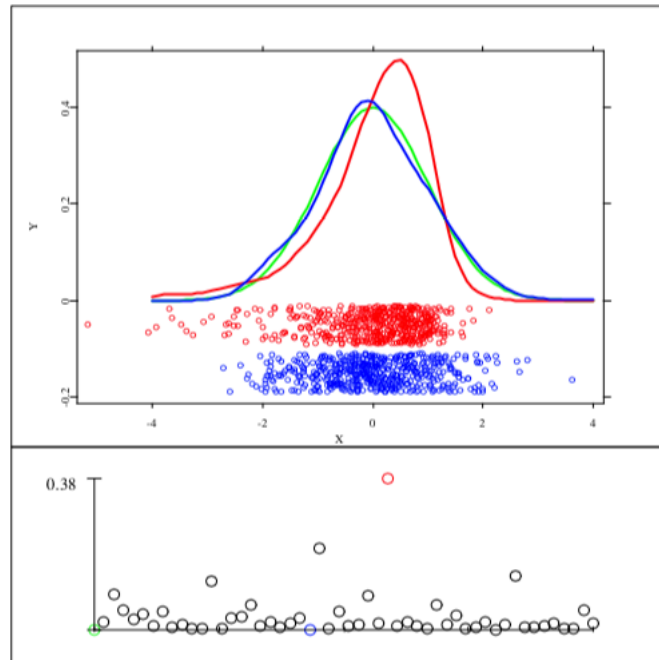


Figure 18.9. Projection Pursuit with the Sibson-Jones index with 13 transformed variables. [MVAppsib.xpl](#)

EXERCISE 18.2 Construct a configuration of points in \mathbb{R}^2 such that $x_{med,j}$ from (18.2) is not in the “center” of the scatterplot.

EXERCISE 18.3 Apply the SIR technique to the U.S. companies data with $Y =$ market value and $X =$ all other variables. Which directions do you find?

EXERCISE 18.4 Simulate a data set with $X \sim N_4(0, I_4)$, $Y = (X_1 + 3X_2)^2 + (X_3 - X_4)^4 + \varepsilon$ and $\varepsilon \sim N(0, (0.1)^2)$. Use SIR and SIR II to find the EDR directions.

EXERCISE 18.5 Apply the Projection Pursuit technique on the Swiss bank notes data set and compare the results to the PC analysis and the Fisher discriminant rule.

EXERCISE 18.6 Apply the SIR and SIR II technique on the car data set in Table B.3 with $Y =$ price.