Chapter 3

Sampling Properties of SIR

In this chapter, we discuss the sampling behavior of SIR. First we establish the root-n consistency in Section 3.1. An asymptotic formula which describes how close the SIR directions are to the e.d.r. space is given. Then we derive a chi-square test for determining the number of significant nonzero eigenvalues. This provides an estimate of the reduced dimension K in the dimension reduction model (1.1) of chapter 1. In Section 3.3, we discuss the issue of how many slices should be used. Other sampling aspects of SIR are given in Section 3.4.

3.1 Consistency of SIR.

We shall assume that H is fixed, and the range of Y is partitioned into H intervals, $I_h, h = 1, \cdots$. Slice h consists of cases with $\mathbf{x}_i \in I_h$.

3.1.1 The root n rate.

Let $p_h = P\{y \in I_h\}$, $\mathbf{m}_h = E(\mathbf{x}|y \in I_h)$. Elementary probability theory shows that $\bar{\mathbf{x}}_h$ converges to $\mathbf{m}_h$ at rate $n^{-1/2}$. Let V be the matrix $\Sigma_{h=1}^{H} p_h(\mathbf{m}_h - E\mathbf{x})(\mathbf{m} - E\mathbf{x})'$. It is clear that the $\hat{\Sigma}_\eta$ converges to V at the root n rate. Let $\mathbf{b}_j$ be the jth eigenvector for the eigenvalue decomposition:

$$V\mathbf{b}_j = \lambda_j \Sigma_{\mathbf{x}} \mathbf{b}_j$$

The SIR direction $\hat{\beta}_j$, is seen to converge to the corresponding eigenvector $\mathbf{b}_j$ at the root n rate. Now we use Theorem 2.1 and the simple identity $\mathbf{m}_h = E(E(\mathbf{x}|y)|y \in I_h)$ to see that $\mathbf{b}_j$ will fall in the e.d.r. space.

The case that the range of each slice varies in order to ensure an even distribution of observations is related to the following choice of intervals:

$$I_h = (F_y^{-1}((h-1)/H), F_y^{-1}(h/H)),$$

too many slices?

where $F_y(\cdot)$ is the c.d.f. of y. The root n consistency result still holds.

Q: What happens if $H = \frac{n}{2}$?

A: still O.k. Why ?

Why SIR is still consistent when the number of slices H is large?

Consider the case that $H = n/L$, so that there are L cases in each slice. We shall study what happens if L is fixed when n tends to the infinity. In particular, we want to know what $\hat{\Sigma}_\eta$ converges to first.

Two basic facts are needed. The first one is an ANOVA indentity :

$$\hat{\Sigma}_{\mathbf{x}} = \hat{\Sigma}_\eta + \hat{\Sigma}_e$$

where $\hat{\Sigma}_e$ is average of within-slice covariance :

$$\hat{\Sigma}_e = H^{-1} \sum_{h=1}^{H} \hat{\Sigma}_h$$

$$\Sigma_h = L^{-1} \sum_{i\, in\, slice h} (\mathbf{x}_i - \bar{\mathbf{x}}_h)(\mathbf{x}_i - \bar{\mathbf{x}}_h)'$$

Here recall again

$$\hat{\Sigma}_{\mathbf{x}} = n^{-1} \sum (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

$$\hat{\Sigma}_\eta = H^{-1} \sum_{h=1}^{H} (\bar{\mathbf{x}}_h - \bar{\mathbf{x}})(\bar{x}_h - \bar{\mathbf{x}})'$$

From this identity, it is enough to find out what the average within-slice covariance, $\hat{\Sigma}_e$, converges to.

The second fact is that in order to get an unbiased estimate of any covariance, we should not use the sample size itself. The correct denominator should be the sample size minus one. Of course this does not matter if the sample size is large. But for the situation considered here, it is critical because within each slice, there are only a fixed number of cases L,which could even be as small as 2.

Using the second fact and the law of large number (as H tends to infinity), it is intuitively clear that $\hat{\Sigma}_e$ converges to

$$E(\frac{L-1}{L} cov(\mathbf{x}|Y)) = \frac{L-1}{L} E(cov(\mathbf{x}|Y))$$

Now the population version of ANOVA states that

$$Cov(\mathbf{x}) = cov(E(\mathbf{x}|Y)) + E(cov(\mathbf{x}|Y))$$

To simply the notations, we can rewrite this as

$$\Sigma_{\mathbf{x}} = \Sigma_\eta + \Sigma_e$$

Since it is clear that $\hat{\Sigma}_{\mathbf{x}}$ converges to $\Sigma_{\mathbf{x}}$, putting things together, we see that $\hat{\Sigma}_\eta$ $(= \hat{\Sigma}_{\mathbf{x}} - \hat{\Sigma}_e)$ coverges to

$$\Sigma_{\mathbf{x}} - \frac{L-1}{L}\Sigma_e = L^{-1}\Sigma_{\mathbf{x}} + \frac{L-1}{L}\Sigma_\eta$$

Now recall the population SIR :

$$\Sigma_\eta b_i = \lambda_i \Sigma_{\mathbf{x}} b_i$$

It follows that

$$(L^{-1}\Sigma_{\mathbf{x}} + \frac{L-1}{L}\Sigma_\eta)b_i = (L^{-1} + \frac{L-1}{L}\lambda_i)\Sigma_{\mathbf{x}} b_i$$

So we see that asymptotically, the eignevctors from the SIR algorithm remain the same, even if the number of slices H increase in a way that within each slice there are only a fixed number of cases.

The eignevalues does not converge to the corresponding population values though. So if L is small, it may make sense to estimate the true eignvalues by solving

$$\hat{\lambda}_i = L^{-1} + \frac{L-1}{L}\lambda_i$$

However, this may lead to a negative value. So the problem is not completely resolved. Nevertheless, an interesting message is that if the eignevalues from the SIR output are smaller than the reciprocal of the number of cases per slice, then the corresponding SIR components would not be significant.

columns of $H = 10$, in Tables 3.2 and 3.3. (the conclusions are similar for other H's). For $\bar{\lambda}_{(8)}$, the numbers are close to the rescaled χ^2 values. Thus guided by the χ^2, not very often we will falsely conclude that the third component is real (or mistakenly claim that there are at least 3 components in the data).

Turning to $\bar{\lambda}_{(9)}$, we expect the numbers to be larger than what are given by using the rescaled χ^2 that falsely assumes only one component in the model. For the rational function model with $\sigma = .5$, this is clearly so, as we see that the 1% quantile of $\bar{\lambda}_{(9)}$ is close to the 99% quantile of the rescaled χ^2. Thus in this case, we correctly infer that there are at least 2 components in the model in each of the 100 replicates. As confirmed by the corresponding $R^2(\hat{\beta}_2)$ reported in Table 3.3, high value of $\bar{\lambda}_{(9)}$ leads to good performance of $\hat{\beta}_2$ as an e.d.r. direction. On the other hand, the distribution of $\bar{\lambda}_{(9)}$ for the quadratic model with $\sigma = 1$ shows a substantial overlap with the rescaled χ^2. This is reflected in the relatively lower average and higher standard deviation of $R^2(\hat{\beta}_2)$ in Table 3.2. But a positive point is that by comparing $\bar{\lambda}_{(9)}$ with the rescaled χ^2, we realize that our data do not strongly support the claim that the second component is real.

Finally, $\bar{\lambda}_{(10)}$ is well above the associated χ^2, assuring the high average and the low standard deviation of $R^2(\hat{\beta}_1)$ in all cases.

3.2.1 Chi-squared test.

As argued before, in order to be really successful in picking up all K dimensions for reduction, the inverse regression curve can not be too straight. In other words the first K eigenvalues for V must be significantly different from zero compared to the sampling error. This can be checked by the companion output eigenvalues.

The asymptotic distribution of the average of the smallest $p - K$ eigenvalues, denoted by $\bar{\lambda}_{(p-K)}$, for $\hat{V}$ can be derived, based on perturbation theory for finite dimensional spaces (Kato 1976, chapter 2). For normal $\mathbf{x}$, we have the following result.

Theorem 3.1. *If $\mathbf{x}$ is normally distributed, then $n(p - K)\bar{\lambda}_{(p-K)}$ follows a χ^2 distribution with $(p - K)(H - K - 1)$ degrees of freedom asymptotically.*

3.2.2 Eigenvalues and the assessment of K.

An outstanding dilemma facing all data-analysts is that the more you screen, the more you may find. Good or bad ? While it is desirable to discover as many patterns as possible so one can have a better chance to develop a new theory, this also increases the chance of a false alarm. It is helpful to know whether an observed pattern is spurious or not. Yet this is by no means an easy task, and there is not much discussion on this issue in the literature. For our problem, how many components SIR finds are really there ? The output eigenvalues in the eigenvalue decomposition step of SIR are helpful in answering this question.

First observe that following from Theorem 3.1, we see that theoretically the smallest $p - K$ eigenvalues have to be 0. But in order to be really successful in picking up all K dimensions for reduction, the inverse regression curve can not be either degenerated or close to being degenerated. In other words the first K eigenvalues for the covariance matrix must be significantly different from zero compared to the sampling error. This can be checked by using the companion output eigenvalues of SIR. In the last section, we have derived the asymptotic distribution of the total of the smallest $p-K$ eigenvalues. We may use that result to give a conservative assessment of the number of components in the model.

For $j = 0, 1, 2, ...,$ we define

$$\textit{P-value}_j = P\{\chi^2_{(p-j)(H-j-1)} \geq n(n-p)\bar{\lambda}_{(p-j)}\}$$

This sequence of *P-values* can be used to indicate how many components are found by SIR. A simple forward selection procedure is to start with $j = 0$. If *P-value*$_j$ is less than say .05, then we may claim that there are at least $j+1$ components. Go to the next j till we fail to make the claim. Of course, we may have many other selection procedures to use. Mallows(1973, Technometrics) pointed out the merit of inspecting the plot of the whole sequence of the C_p measures. His point applies to our case too. Thus a sudden jump from a small P-value to a large P-value serves as a better indication of where to stop. We also found the sequence of eigenvalues themselves are often good indicators of how many components found by SIR are worth of close inspection. Typically a value more than .25 is noteworthy. Later on, we shall interprete eigenvalues as R-squared values for multiple linear regression of some suitable transformations of y against $\mathbf{x}$.

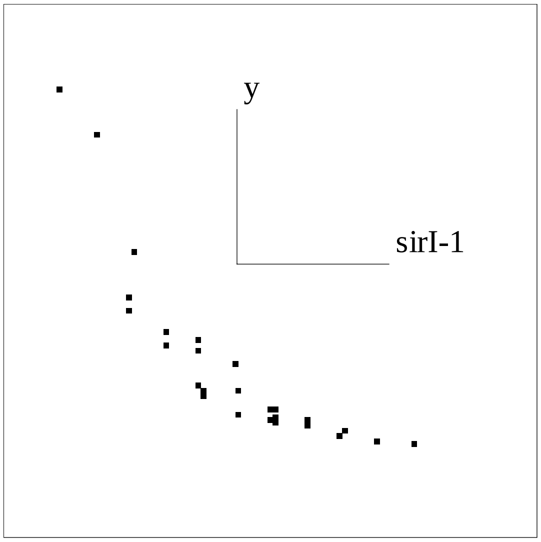

Figure 4.1: SIR view for Worsted yarn Data

the pattern is linear. Note that SIR is invariant under monotone transformation of y. We will come back to the transformation aspect of SIR later on.

4.2 Variable selection.

Boston Housing data has thirteen regressors. This means that each SIR direction would have 13 coefficients. An immediate concern how to interpret so many coefficients properly ? This general issue is discussed in this section.

First of all, Each variable has its own unit so a small coefficient does not mean that the corresponding variable should be ignored. A quick remedy is to report the result after standardizing each variable to have the same variance(=1). But this may not be enough.

In order to obtain a parsimonious description for the estimated e.d.r. space, it is appropriate to select a small subset of regressors for conducting SIR. Just like the variable selection in multiple linear regression, there are several ways of doing it. The following is one simple recommendation.

(1). Conduct SIR with all regressor variables included. Let $\hat{b}_1$, .., $\hat{b}_k$ be the estimated e.d.r. directions.

(2). Then find a projection from a small subset S_1 of regressors, denoted by $\hat{b}'_{s1}\mathbf{x}$, which is still reasonably close to the first projection $\hat{b}'_1\mathbf{x}$ with ,say, an R-square value of 90% (which amounts to about 18.5^o difference between the two projection angles) or better. This can be done by either forward or backward selection procedure in multiple linear regression by treating $\hat{b}'_1\mathbf{x}$ as y.

(3). After S_1 is selected, we then check if using variables from S_1 is good enough to approximate the second projection $\hat{b}'_2\mathbf{x}$ or not. If not, we should enlarge it and continue to the next projection. Let S be the final set of variables selected.

(4). Apply SIR again, this time using only the variables in S.

(5). If necessary, go through the variable selection procedure (2)-(4) again.

Note it is a good practice to compare the plot found from the reduced variables with the original one. If substantial difference is found, then some caution should be taken.

4.3 Boston housing data.

We first apply SIR to the Boston Housing Data (described in Chapter 1). with H, the number of slices, ranging from 10 to 30. The result, based on $H = 15$, is reported in Figures 4.2(a)-(d). As we rotate the cloud along the y-axis, it looks like a helix or slide. A further inspection of the eigenvalues , .82, .48, .20, .08, .05, $\cdots$, reveals that there are three significant components. Figure(4.3) provides the scatter-plot matrix for y and these three projection variables.

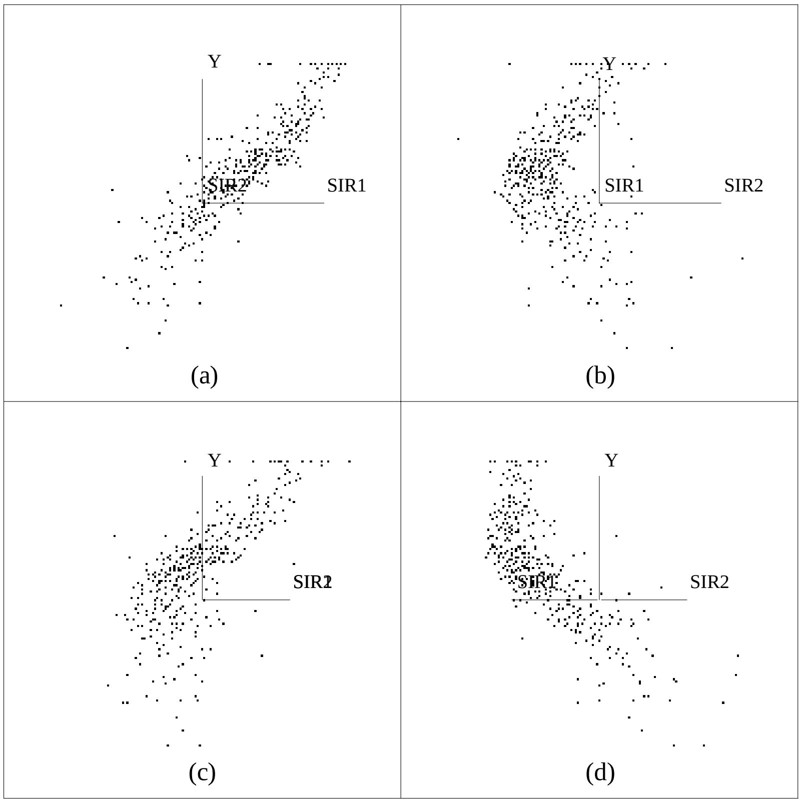

Figure 4.2: SIR view for ~~Worsted yarn Data~~ Boston Housing

4.3.1 Crime rate.

When we carry out the variable selection procedure as described above (forward variable selection is used), the result is not illuminating. For the first projection, we can identify the main contributor to be x_{13}, proportion of poor, with x_6, average number of rooms in houses, as a close runner-up. Primary contributors of the second projection variable are harder to identify. The top candidate x_1, crime rate, leads seven other competing variables only marginally.

We take a closer look at the relationship of crime rate with other variables by inspecting scatterplots. As observed in Chapter 1, a special group of cases with high crime rate stand out from the others. All cases in this group share the same value in each of the following 5 variables : $x_2, x_3, x_9, x_{10}, x_{11}$.

4.3.2 The low crime rate group.

Excluding this high crime rate group, there are 374 cases remaining. We run SIR on these and now there are only two components significant. The pictures are similar to but sharper than the ones obtained from the whole sample. We are able to identify x_6 as the primary contributor of the first component. For the second component, x_1 and x_{13} are the top candidates. We then run SIR again, with x_1, x_6, x_{13} as the regressor variables this time (Figures (4.3(a)-(d)). The first component $\hat{b}_1'\mathbf{x}$ is clearly due to x_6, which has a correlation higher than .99 with $\hat{b}_1'\mathbf{x}$. The second component, can be described roughly as $x_1 + 30x_{13}$ adjusted by the first component for orthogonalization. These two SIR variates are nonlinearly correlated; see Figure 4.4.

Other values of H, ranging from 10 to 30, have provided essentially the same view. The logarithm transformation used to obtain Y is borrowed from Harrison and Rubinfeld (1978). This is not necessary because SIR is invariant under the monotone transformation of Y. SIR would still find the same projections if the original scale were used. In Figure 4.5 the original scale of house price is used. This can be compared with Figures 4.3. It appears that the logarithm transformation is unnecessary.

4.3.3 Intrepretation.

In this study, SIR identifies two key factors of different nature and provides a graphical summary. The variable x_6, average number of rooms, is a physical factor, which may reflect the construction cost and the practical utility of a house to some degree. It affects the physical condition of a house. The other variable $x_1 + 30x_{13}$, the crime rate and percentage of the poor, is a socio-economic factor. It reflects the desirability of the house's neighborhood which in turn affects the area's land value. SIR reveals the nonlinear association between these two factors. The importance of the physical factor is also confirmed by other methods; for example, the straightforward linear regression, the more complicated model fitting of Harrison and Rubinfeld, and ACE of Breiman and Friedman. In fact, in each of these studies, the physical variable is always the leading factor which accounts for the highest percentage of variation in the prediction equation. Because of this consistency from different studies,

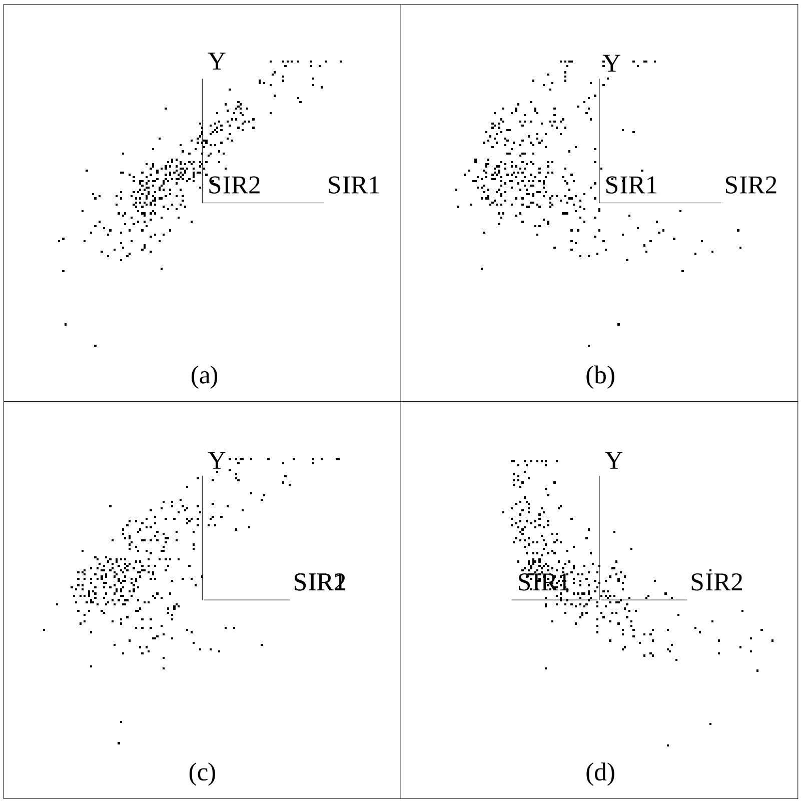

Figure 4.3: SIR's view on the subsample that excludes a high crime rate group with only x_1, x_6, x_{13} as the regressor variables. Y is the logrithm of the median value of owner-occupied house.

one might naively be forced to conclude that x_6 is the dominant factor. To challenge this simple-minded statement, we should resort to the helix type of nonlinear confounding pattern exhibited by the three-dimensional SIR plot. The second factor which appears equally important from the SIR plots, cannot be found from the other studies because their models have precluded structure like the one we found here a priori.

It is usually hard to draw any decisive conclusion from a single study. If the same helix shape of distribution also exists in data from other cities, for example, then the finding would be much more noteworthy. The graphics found here, however, is not available from linear regression or other methods. The exposure of the helix type data cloud offers an alarming diagnosis for methods aiming at the approximation the regression surface, which are sensitive to nonlinear confounding. We shall turn to this point again in Chapter 10.

Finally, we have run SIR with x_1, x_6, x_{13} and x_5 as the regressors. It turns out that the two components found are essentially the same as those without x_5, and that the correlation coefficient for the corresponding components is higher than .99 for each of the two directions found by SIR. This suggests that x_5, nitrogen oxide concentration, does not show a significant

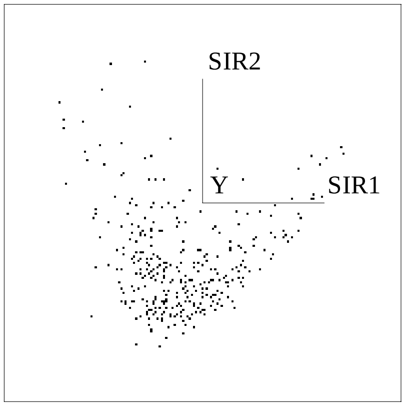

Figure 4.4: The scatterplot of the first two components in Figures 4.3 (a)-(d)

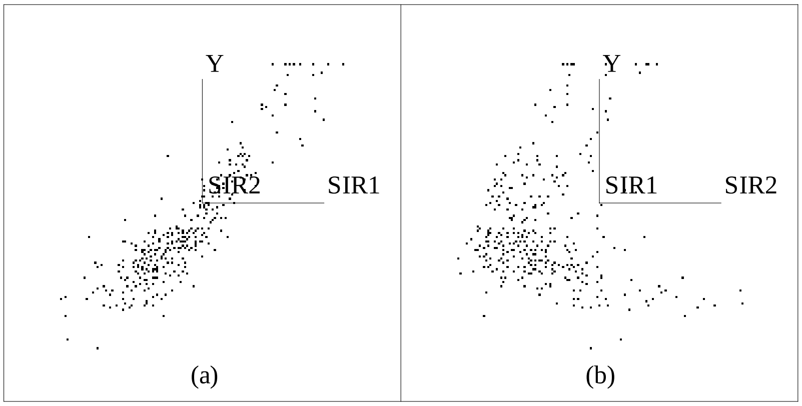

Figure 4.5: SIR's view on the subsample that excludes a high crime rate group with only x_1, x_6, x_{13} as the regressor variables. Y is the original scale median value of owner-occupied house.

role in affecting the relative housing prices in the low crime rate areas.

4.4 Structure removal.

Quite often, finer structure in the data can only be detected after the main structure is removed.

Example 4.4.1 Ozone data. We take a data set from Breiman and Friedman (1985), the data for studying the atmospheric ozone concentration in the Los Angeles basin. We use the daily measurement of ozone concentration in Upland as the output variable y and want to find its relationship with eight meteorological variables (see Table 8.4). There are $n = 330$ observations in the study. First, we apply SIR to the data and find one significant component.

This component is almost identical to the $\hat{b}'_{ls}\mathbf{x}$, the component found by the linear least squares fitting. For certain slice sizes, we can find a marginally significant second component as well, but we decide to ignore that. We then use a forward selection method to find out the important variables contributing to the first component. Three variables x_1, x_2, x_6 are found that explain more than 99% of the total variation of the first component. We run SIR again using only x_1, x_2, x_6 as the input variables. The scatterplot of y against $\hat{b}'_s\mathbf{x}$, the first component found. We use 30 slices here for SIR, but other choices yield almost identical scenes. The correlation between $\hat{b}'_{ls}\mathbf{x}$ and $\hat{b}'_s\mathbf{x}$ is above .99 as well.

A quadratic trend is visible from the SIR plot. After fitting a quadratic polynomial :

$$y = c_0 + c_1 u_1 + c_2 u_1^2 + \epsilon,$$

where u_1 denotes the variable $\hat{b}'_s\mathbf{x}$.

We finally apply SIR again to the residual and found one significant component, which gives the view Figure 4.6. An interesting triangle pattern of heterogeneity is seen.

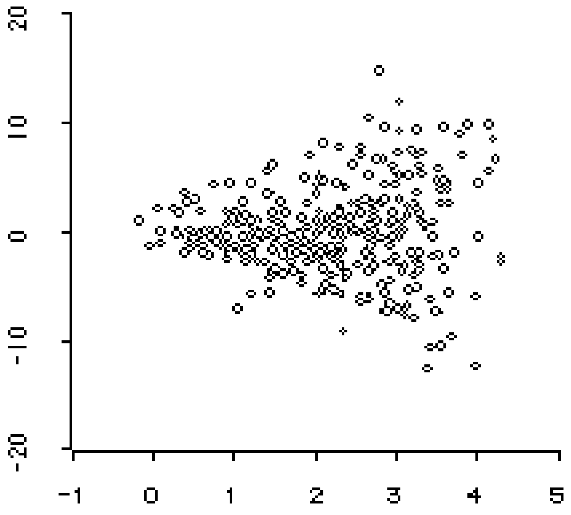

Figure 4.6: Ozone data. Residuals against the direction found by SIR

4.5 OTL push-pull circuit.

Most regression analysis techniques deal with data which are empirically collected. But there are many other cases in which the relationship between the input and output variables can be derived from physical/mathematical laws. The response is deterministic and is indeed already given. There are no data to analyze.

Figure 4.* depicts an OTL push-pull circuit used by a TV manufacturing company in Hanzou City of China(Chen et. al. 1985).