**Appendix**

Methods using local smoothing : Additive models, ACE, projection pursuit regression, MARS

**1. One dimensional nonparametric regression model :**

$$Y = g(x) + \epsilon$$

**Methods** : kernel smoothing, least squares by splines, wavelets, etc. LOWESS

Statistical ideas : balance between bias and variance ; cross-validation; model selection

**In terms of transformation** : $g(x) = E(Y|x)$ is the optimal transformation $t(\cdot)$ such that

$$min_{t(.)} E(y - t(x))^2$$

where the minimization is over any transformation of $x$. Equivalently, the squared correlation coefficient (called R-squared) between $Y$ and the transformed $x$ is maximized

**2. Additive model : (Hastie and Tibshirani, 1986, Statisical Science, 297-318**

$$Y = g_1(x_1) + g_2(x_2) + \cdots + g_p(x_p) + \epsilon$$

**Methods** : same as 1, with iterations. Thus beginning with $x_1$, a best smoother $\hat{g}_1(x_{1)}$ is obtained by regressing $Y$ against $x_1$ nonparametrically. Then take the residual as $Y$ and regress against $x_2$. Again find the residual and apply the same smoothing procedure to $x_3$ and continue till $x_p$ is fitted. After that, update $\hat{g}_1$ by regressing $Y - \sum_{i=2}^{p} g_i(x_i)$ against $x_1$. Continue the updating for each function many times till some criterion of convergence is satisfied.

**In terms of transformation:** find transformation for each coordinate so that the squared multiple correlation coefficient (called R-squared) between $Y$ and the transformed regressors is maximized.

**3. ACE (alternating conditional expectation) , also referred to as alternating least squares**(Breiman and Friedman 1985, JASA, 580-597):

$$f(Y) = g_1(x_1) + g_2(x_2) + \cdots + g_p(x_p) + \epsilon$$

Methods : (1) $p = 1$. First regress $Y$ against $x$ as in (1) to obtain a smooth $\hat{g}(x)$. Then reverse the role and regress $\hat{g}(x)$ against $Y$ to get a smooth $\hat{f}(Y)$. Multiply a constant to $\hat{f}(Y)$ so that the variance of $\hat{f}(x)$ is normalized to one. Regress $\hat{f}(Y)$ against $x$ to update $\hat{g}(x)$ and then use that to update $\hat{f}(Y)$ again. Iterate till some convergence criterion is reached.

(2) $p > 1$. Combine the steps used in (1) and the steps used in additive modeling.

**In terms of transformation:** find transformations for both regressors and response variable so that the multiple correlation is maximized.

Comparison with SIR :

For Boston Housing data, ACE, after variable selection yields an additive equation involving four regressors.

"The two terms that enters most strongly involve the number of rooms squared and the logrithm of the fraction of the population that is of lower status. ... The remaining two variables that enter into this model are pupil-teacher ratio and property tax rate. ... "

"

In contrast, SIR identifies three variables : crime rate, number of room, and the fraction of the population in lower status. The crime rate and the fraction of the poor is further combined into one sigle variable, yielding a two component model, one being the house size indicator and the other being a social environment indicator. These two factors interact in a nonlinear way, exhibiting a helical-looking data pattern.

Important variables found by both SIR and ACE : room size and percentange of the poor . Yet, ACE assumes that the effects from these two variables are additive (or at least approximately so). This assumption is NOT implied by SIR.

Crime rate is not identified as a key variable by ACE. SIR found this factor to be doubly important : (1) as a stratification variable in identifying a lower-crime rate cluster (2) as one of the three key variable in the final model.

**Projection pursuit regression (Friedman and Stuetzle 1981), JASA 817-823** :

$$Y = g_1(\beta_1'\mathbf{x}) + g_2(\beta_2'\mathbf{x}) + \cdots + g_k(\beta_k'\mathbf{x}) + \epsilon$$

Methods : nonlinear least squares + smoothing.

(1) $k = 1$. Start with regressing Y on $\mathbf{x}$ linearly to get $\hat{\beta}$. Treat $\hat{\beta}$ as $x$ and apply the one-dimensional regression method to estimate $g_1$. Then treating the estimated $g_1$ as being fixed, use nonlinear least squares to update $\hat{\beta}$. Iterate several times.

(2)$k > 1$. This requires a sweeping through each regressor iteratively in a way similar to how additive modeling generalizes the one-dimensional nonparametric regression.

Remark 1: a global searching method for $\beta_i$ was proposed in the original article of Friedman and Stueztle.

Remark 2 : There is an interesting connection with a robustness property of multiple linear regression due originally to Brillinger(1977, Biometrika, 509-515; 1983 in "A Festschrift for Erick L.Lehmann", Belmont, CA: Wadsworth 97-114) and later extended by Li and Duan(1989, Ann. Stat. 1009-1052).

# Chapter 6

# Transformation and SIR

In this chapter, we shall draw connections between SIR and multiple linear regression. This is based on Chen and Li(1998).

## 6.1 Dependent variable transformation.

In this section, we shall derive SIR from the viewpoint of tranformation on the dependent variable $Y$. This derivation is descriptive in nature.

Transformation has become one of the routine steps in regression analysis. For experienced data analysts, an inspection on the scatterdiagrams, or on plots of the residuals may often lead to some suitable transformations. For high dimensional data, however, we have many scatterdiagrams to inspect. Moreover, in many cases, a common transformation for simplifying the analysis may not be possible. A transformation suitable for one plot may not be good for another.

Contrary to these subjective eyeballs-based transformation methods, Box and Cox(1964) formulated the problem rigorously as the estimation of the power parameter $\lambda$ in the power transformation family:

$$\begin{aligned} T(Y,\lambda) &= \alpha + \beta' \mathbf{x} + \epsilon \\ T(Y,\lambda) &= (Y^{\lambda} - 1)/\lambda, 0 \leq \lambda \leq \infty. \end{aligned} \tag{1.1}$$

One hope is that this family may be flexible enough to incorporate many reasonable transformations suggested by human eyes and to achieve the multiple purpose of linearizing the regression , stabilizing the variance, and achieving the normality.

While the Box-Cox transformation model is a good approximation of many data sets, it is clearly deficient for the application of 3-D graphing. Indeed, if the Box-Cox transformation model is correct, then there is no pressing need to project $\mathbf{x}$ on more than one directions. Finding a good estimate of $\beta$ and project $\mathbf{x}$ on the estimated $\beta$ direction seems informative enough.

In this chapter, transformation will be used in a way different from its traditional role of being a mechanism for improving the goodness of model fitting. It will serve as an intermediate tool for finding interesting projections of $\mathbf{x}$. We shall consider a direction $b$ interesting

for viewing if the resulting scatterplot of $Y$ against the projected variable $b'\mathbf{x}$, may suggest a transformation on $Y$ to obtain a good linear fit. A commonly used measure of goodness of fit, the R-squared, is adopted here. For any direction $b$, let the associated "optimal" transformation be $T_b(Y)$, i.e, $T_b(Y)$ achieves

$$R_b^2 = \max corr\ (T(Y), b'\mathbf{x})^2 \tag{1.2}$$

where the maximum is taken over all transformations $h$, and $corr$ stands for the correlation coefficient.

Now we propose $R_b^2$ as the index in searching for the optimal projections. This index reasonably reflects the degree of interestingness hidden in the scatterdiagram. Of course, this does not mean that one has to find the optimal transformation by inspecting the scatterdiagram with eyes. Neither did we claim that all interesting aspects about the scatterdiagram can be fully captured by this single index; otherwise we may need only the index but not the graph. Yet it is believable that a high value of the maximum R-squared may allow a lot of interesting features to occur, including blurring curves, heteroscedasticity, and clusters.

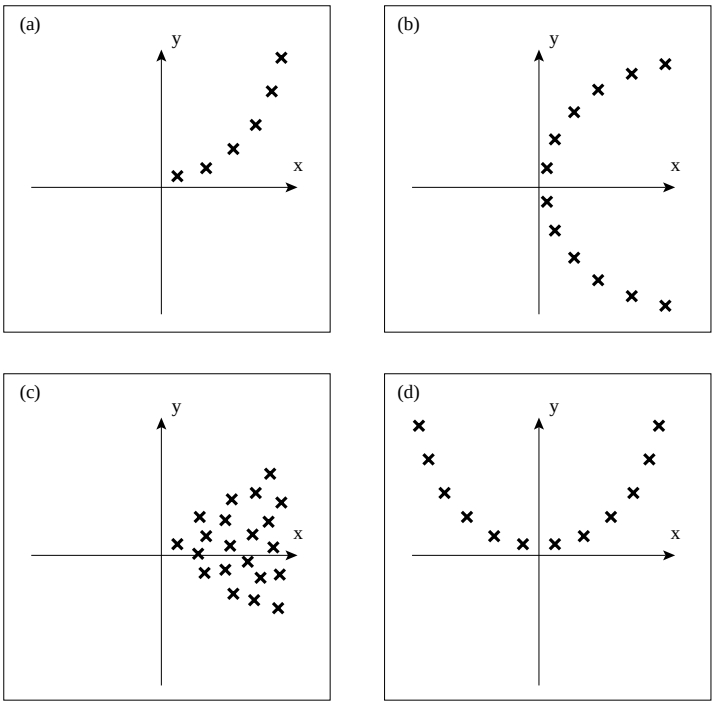


Figure 6.1: Transformation Helps Linearize the Regression for (a), (b), (c), but not (d)

Figures 6.1(a)-(d) show some typical situations. Transformation on $Y$ can help increase the R-squared value substantially in Figure 6.1(a)-(c) ( in (b) and (c), take the absolute value, for example). It does not help in Figure 6.1(d), however.

What transformation will optimize the R-squared value ? The answer is

$T_b(y) = E(b'\mathbf{x}|Y = y)$. To see this, first assume that $E\mathbf{x} = 0$ for simplicity. Then

$$\begin{aligned} corr(T(Y), b'\mathbf{x})^2 &= \frac{[ET(Y)b'\mathbf{x}]^2}{varT(Y)var(b'\mathbf{x})} \\ &= \frac{(ET(Y)T_b(Y))^2}{varT(Y)var(b'\mathbf{x})} \\ &= corr(T(Y), T_b(Y))^2 \frac{varT_b(Y)}{var(b'\mathbf{x})} \end{aligned}$$

It is now clear that the maximum is achieved when the correlation coeficient in the last expression is eqal to 1 or -1, showing that our answer is correct.

With our index, the first projection is a direction $b_1$ that maximizes $R_b^2$ over all vectors $b$. After finding $b_1$, we then look to those directions uncorrelated to $b_1$ for the the second maximization direction. Then we can plot $Y$ against $b_1'\mathbf{x}$ and $b_2'\mathbf{x}$ by, say, the 3-D rotating plot for visualization.

We may continue the above maximization process to obtain a set of vectors, $b_1, ..., b_p$, satisfying the conditions

$$\begin{aligned} &cov(b_i'\mathbf{x}, b_j'\mathbf{x}) = 0, \quad for\ i \neq j \\ &R_{b_i}^2 = \max_b R_b^2, \end{aligned} \tag{1.3}$$

where the maximum is taken over all vectors $b$ satisfying $cov(b'\mathbf{x}, b_j'\mathbf{x}) = 0$, for $j = 1, ..., (i-1)$.

Theorem 1.1 below characterizes the $b_i's$ and establishes the connection with sliced inverse regression. Recall the definition of inverse regression curve

$$\eta(y) = E(\mathbf{x}|Y = y).$$

and its covariance $\Sigma_\eta = cov\,[\eta(Y)]$

**Theorem 6.1.** *The vectors constructed from the maximization problem (1.3), $b_i$, $i = 1, .., p$, are the same as the eigenvectors for the eigenvalue decomposition of the covariance matrix $\Sigma_\eta$ with respect to $\Sigma_\mathbf{x}$; i.e.,*

$$\begin{aligned} &\Sigma_\eta b_i = \lambda_i \Sigma_\mathbf{x} b_i, i = 1, ..., p, \\ &\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \end{aligned}$$

**Proof.** Without loss of generality, we assume that $E\mathbf{x}$=0. First, it can be verified that for any direction $b$, the "optimal" transformation is $T_b(y) = E(b'\mathbf{x}|Y = y) = b'\eta(y)$. Then a simple conditional expectation argument leads to

$$\begin{aligned} cov(T_b(Y), b'\mathbf{x}) &= E[T_b(Y)(b'\mathbf{x})] \\ &= E[T_b(Y)E(b'\mathbf{x}|Y)] \\ &= b'E(\eta(Y)\eta(Y)')b \\ &= b'\Sigma_\eta b. \end{aligned} \tag{1.4}$$

It follows that

$$R_b^2 = \frac{b'\Sigma_\eta b}{b'\Sigma_{\mathbf{x}} b} \tag{1.5}$$

Therefore the eigenvalue decomposition of $\Sigma_\eta$ with respect to $\Sigma_x$ solves the maximization problem (1.3), completing the proof. □

The spectrum decomposition problem stated in this theorem is exactly the same as the one proposed in sliced inverse regression.

An equivalent way of defining interesting directions for projections can be phrased in the following. For a transformation $T(Y)$ of the dependent variable $Y$, consider the linear least squares fit by $\mathbf{x}$; namely,

$$\min_{a\in R, b\in R^p} E(T(Y) - a - b'\mathbf{x})^2 \tag{1.6}$$

Denote the minimizer by $a(T), b(T)$. Consider again the *R-squared*:

$$\frac{Var\ (a(T) - b(T)'\mathbf{x})}{Var\ T(Y)} = [corr(T(Y), b(T)'\mathbf{x})]^2 \tag{1.7}$$

Let $T_1$ be any "optimal" transformation that maximizes (1.7). Subject to being orthogonal to $T_1$ in the sense that $cov(T(Y), T_1(Y)) = 0$, we again maximize (1.7) to find $T_2(Y)$. Continue in the similar fashion until we find $p$ optimal orthogonal transformations $T_1, ...., T_p$. The following theorem shows that the regression slope vectors, $b(T_i), i = 1, \cdots, p$, are the solutions, $b_i, i = 1, ..., p$, for the maximization problem (1.3).

**Theorem 6.2.** *The regression slope vectors , $b(T_i), i = 1, ..., p$, for the optimal transformations $T_i, i = 1, ..., p$ are the solutions $b_i, i = 1, ..., p$ for (1.3). On the other hand, $T_i(y) = E(b_i'\mathbf{x}|Y = y), i = 1, ..., p$, maximize (1.7).*

**Proof.** Our strategy is to show that the two maximization problems, (1.3) and the maximization of (1.7), can be translated into a common double maximization problem of the form (1.9) below.

First observe that the least squares solution to (1.3) is also a solution to the maximization problem:

$$\max_b corr(T(Y), b'\mathbf{x})^2 \tag{1.8}$$

Thus $T_1(\cdot)$ solves

$$\max_{T(\cdot)} \max_b\ corr(T(Y), b'\mathbf{x})^2 \tag{1.9}$$

Reversing the ordering of the two "max", this is the same problem that $b_1$ solves. It follows that $b_{T_1}$ is proportional to $b_1$ and $E(b_1'\mathbf{x}|Y) = b_1'\eta(Y)$ can be taken as the optimal transformation $T_1(Y)$.

Next, for any direction $b$ uncorrelated to $b_1$, i.e, $0 = cov(b'\mathbf{x}, b_1'\mathbf{x})$, we also have

$$(cov(T_b(Y), T_1(Y)) = cov(E(b'\mathbf{x}|Y), E(b_1'\mathbf{x}|Y)) = b'cov(\eta(Y))b_1 = \lambda_1 b'\Sigma_{\mathbf{x}} b_1 = 0,$$

where the next to the last identity is due to the definition of eigenvector. This implies that to find $b_2$ , the double maximization problem (1.9) can be restricted to those $b$ that are

uncorrelated with $b_1$ as well as to those $T(\cdot)$ that are orthogonal to $T_1(\cdot)$ . On the other hand, we shall show that the same restriction applies when finding $T_2(\cdot)$. To do this, it is enough to check that for any $T(y)$ that is orthogonal to $T_1(y)$ , we have $b_1' \Sigma_{\mathbf{xx}} b_h = 0$. Since the regression coefficient $b_h$ is equal to $\Sigma_{\mathbf{x}}^{-1} cov(T(Y), \mathbf{x})$, it suffices to verify that $cov(T(Y), b_1'\mathbf{x}) = 0$. But by the same conditional expectation argument used in (1.4), we see that $cov(T(Y), b_1'\mathbf{x}) = cov(T(Y), E(b_1'\mathbf{x}|y)) = cov(T(Y), h_1(Y)) = 0$, proving the claim.

It follows that $b(T_2)$ is proportional to $b_2$, and that $E(b_2\mathbf{x}|Y)$ can be taken as $T_2(Y)$. For $p$ larger than 2, We can repeat the same argument to complete the proof.

These theorems offer an interpretation for the eigenvalues in the output of SIR:

the $i$th eigenvalue of SIR is equal to the $R$-squared value of the linear regression when $Y$ is transformed to $T_i(Y)$.

## 6.2 Some Remarks.

**Remark 2.1.** As mentioned before, in the search of an "optimal" transformation we do not restrict to the monotone ones. Monotone transformations are reasonable only if we believe in the adequacy of transformation models. While there may be many good reasons to require monotonicity for the first transformation (see Ramsay 1988), they are less compelling for the second transformation. Indeed, Theorem 1.2 implies that *the second one can not be monotone if the first one is so*.

**Remark 2.2.** No single index can reflect all interesting aspects in a scatterdiagram; otherwise we may need only the index, not graphics. Our transformation-based index $R^2(b)$ is no exception. It performs poorly when the scatterdiagram of $Y$ against $b'\mathbf{x}$ contains a pattern of symmetry about some vertical line. The correlation coefficient is zero and we cannot increase it by transforming $Y$. Thus $R^2(b)$ is always zero no matter how interesting the pattern of symmetry is. This offers an explanation for why SIR cannot recover the e.d.r. direction in a simple quadratic function $Y = (\beta'\mathbf{x})^2$; see Cook and Weisberg (1991) and the Rejoinder of Li(1992) for more discussion. One remedy is to consider double transformation (Carroll and Ruper 1988); namely to allow the transformation on $b'\mathbf{x}$ as well. We may use the maximum correlation between $y$ and $b'\mathbf{x}$ to quantify interestingness in the scatterplot :

$$\max_{T,g} \rho(T(y), g(b'\mathbf{x}))$$

where $T$, $g$ are any square integrable functions. How to maximize this index over all possible directions efficiently is still to be explored.

Nonlinear multivariate analysis techniques such as correspondence analysis, optimal scaling, and others ( Gifi1991), and ACE (Breiman and Friedman 1985, Koyak 1987) use maximum correlation in statistics in a rather different manner. For example, ACE proposes the model

$$T(Y) = \sum_{i=1}^{p} g_i(x_i) + \epsilon$$

where $\mathbf{x} = (x_1, \cdots, x_p)'$, and $g_i$'s. Only one transformation on $Y$ is allowed for the purpose of rescaling. Each regressor is allowed to make transformation, a feature that SIR does not have. However, the additivity assumption can be too strong; a remedy to this is given by MARS (Freedman 1991), but without allowing the transformation on $Y$. These tools aim at finding a good approximation of the regression function $E(Y|\mathbf{x})$ without graphical guidance.

**Remark 2.3.** Although transformation has been used frequently in Statistics, there is one major difference between ours and others. We use transformations of $y$ to suggest interesting patterns in the data only; while others use transformations for functional approximation. Consequently, our transformations are disposable. After finding the directions, we are no longer obligated to these transformations for modeling. The subsequent analysis should be based on what is seen.

**Remark 2.4**. The duality relationship displayed in Theorem 1.2 can be put into a more general context in terms of Hilbert spaces. To simplify the notations, assume that $E\mathbf{x}$ is 0. Consider an infinite dimensional Hilbert space, $\mathcal{H}_1$, consisted of all squared integrable random variables $T(Y)$ that are transformed from $Y$ and have mean 0. Let $\mathcal{H}_2$ be the $p-$dimensional Hilbert space, consisted of the linear combinations of $\mathbf{x}$, $b'\mathbf{x}$. These two Hilbert spaces generate a larger Hilbert space, denoted by $\mathcal{H}$. Measure the distance between two elements, $v_1, v_2$, in $\mathcal{H}$, by the standard deviation of $v_1 - v_2$. Then for any projected variable $b'\mathbf{x}$, the closest element in $\mathcal{H}_1$ is $E(b'\mathbf{x}|Y) - EY$, which is a version of $T_b(y)$. Likewise, for any transformation $T(Y)$ the closest element in $\mathcal{H}_2$ is the best linear fit, $b(T)'\mathbf{x}$. Consider the $p$ dimensional Hilbert subspace, $\mathcal{H}_3$, of $\mathcal{H}_1$, generated by $T_b(Y)$, $b \in R^p$. The duality relationship in Theorem 2.2 simply says that one can find orthogonal basis vectors, say $e_i, i = 1, ..., p$, in $\mathcal{H}_2$ and orthogonal basis vectors, $say$, $v_i, i = 1, ..., p$, in $\mathcal{H}_3$ such that the closest element in $\mathcal{H}_1$ to $e_i$ is a multiple of $v_i$, and conversely the closest element in $\mathcal{H}_2$ to $v_i$ is a multiple of $e_i$. This is the canonical analysis between $\mathcal{H}_1$ and $\mathcal{H}_2$, a special form of singular value decomposition problems prevalent in nonlinear multivariate analysis. In principal, it is possible to enlarge $H_2$ by including a few second order terms (or B-spline terms ) of $\mathbf{x}$.

## 6.3 Examples.

In this section, we shall explain why SIR works in a variety of situations from the transformation based viewpoint.

### 6.3.1 Curves and clusters.

Consider the model

$$Y = sign(\beta_1'\mathbf{x} + \sigma_1\epsilon_1)log(|\beta_2'\mathbf{x} + \alpha + \sigma_2\epsilon_2|), \tag{3.1}$$

where the function $sign(\cdot)$ takes value 1 or -1 depending on the sign of the argument. All coordinates of $\mathbf{x}$ and $\epsilon_1, \epsilon_2$ are independent standard normal random variables. For a clear

illustration, we first study the noise-free case, $\sigma_1 = \sigma_2 = 0$. Take the dimension of $\mathbf{x}$ to be $p = 15$ and generate $n = 300$ cases with

$$\beta_1' = (1\,1\,1\,1\,1\,1\,1\,1\,0\,0\,0\,0\,0\,0\,0), \beta_2' = (0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,1\,1\,1\,1), \alpha = 5.$$

We run SIR with the number of slices equal to 20. Other numbers, 10 and 30, also show similar results. A rotation plot for $Y$ against the first two projections is shown in Figures 6.2(a)-(d). The first eigenvector (Fig. 6.2(a)) finds two curves spreading out symmetrically about the horizontal axis and the second one (Figure 6.2(c)) shows a pattern of two clusters. Table 6.1 gives the first two output eigenvectors and eigenvalues. They are approximately proportional to $\beta_2$ and $\beta_1$ as desired.

Table 6.1: The first two eigenvectors(with standard deviations and ratios and eigenvalues of SIR for (3.1) without error terms.

| | |
|---|---|
| *first vector* | (-.05, -.03, -.01, -.03, -.01, -.03, .01, -.03, -.01, .39, .41, .44, .45, .42, .43) |
| *S.D.* | (.02, .02, .02, .02, .02, .02, .02, .03, .02, .02, .02, .02, .03, .02, .02) |
| *ratio* | (-2.1, -1.5, -0.5, -1.4, -0.3, -1.6, 0.3, -1.4, -0.3, 18, 18, 19, 18, 20, 18) |
| *second vector* | (.35, .39, .35, .24, .28, .30, .32, .27, .33, -.00, -.01, .03, -.02, .04, .11) |
| *S.D.* | (.05, .05, .04, .05, .05, .05, .05, .05, .05, .05, .04, .05, .05, .05, .05) |
| *ratio* | (7.2, 7.7, 8.0, 4.8, 5.7, 6.5, 6.7, 5.1, 6.6, -0.0, -0.3, 0.6, -0.3, 0.8, 2.2) |
| *eigenvalues* | (0.88, .61, .16, .13, .12, .08, .07, .05, .04, .02, .02, .01, .01, .00, .00) |

Figure 6.2(a) shows approximately the scatterplot of $Y$ against $\beta_2'\mathbf{x}$. The symmetry about the horizontal axis is due to the sign function which acts on $\beta_1'\mathbf{x}$ behind the screen. This symmetry yields a zero correlation coefficient between $Y$ and $\beta_2'\mathbf{x}$. But it can be increased greatly by folding the picture over along the $x-$axis, which amounts to taking the absolute value transformation $|Y|$. This explains why SIR is capable of finding this direction. According to Theorem 6.2, the optimal transformation is $T_1(Y) = E(\beta_2'\mathbf{x}|Y)$, which should give an even higher correlation coefficient, about $\sqrt{.84} \approx .92$ as estimated by the squared root of the first eigenvalue of SIR, than the absolute value transformation.

Figure 6.2(c) shows approximately the scatterplot of $Y$ against $\beta_1'\mathbf{x}$. This is the direction to be found by a linear least squares of $Y$ against $\mathbf{x}$, because $Y$ is uncorrelated with any directions orthogonal to $\beta_1'\mathbf{x}$.

Figures 6.2(b) and 6.2(d) show two views of the rotation plot found by SIR. These static views themselves do not offer much additional information. But when we rotate the plot around the vertical axis on the screen, the two curves in 6.2(a) are then turned into two thin plates, floating in and out.

We also repeat the simulation with the noise level set at $\sigma_1 = \sigma_2 = 1$. The output of SIR is also quite close to the directions of $\beta_2$, $\beta_1$; see Table 6.2 and Figures 6.3(a)-(b). The curves are now blurred.

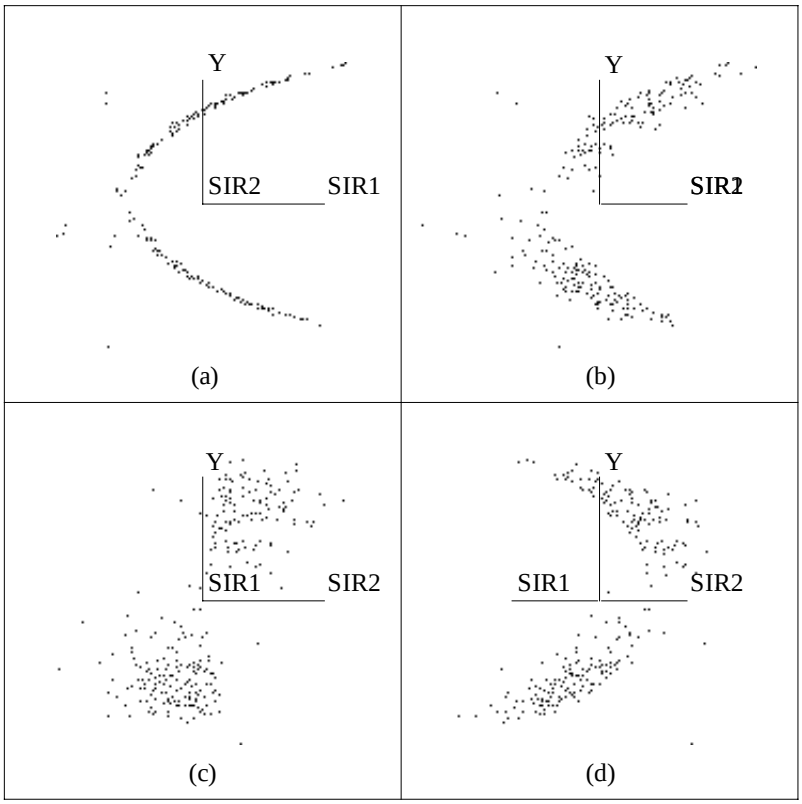


Figure 6.2: SIR's view of Data Generated From (3.1).

**Remark 3.1.** We also simulated the case with $\alpha = 0$. SIR fails in this case because of the symmetry on the $\beta_2$ direction. Second-moment based methods (Cook and Weisberg 1991, the Rejoinder of Li 1991) and variants of principal Hessian direction (Li 1992b) are capable of finding the $\beta_2$ direction.

### 6.3.2 Heteroscedasticity.

A popular model for studying heteroscedasticity is

$$y = \beta_1'\mathbf{x} + \epsilon g(\alpha + \beta_2'\mathbf{x}), \tag{3.2}$$

where $g$ is often conveniently taken to be a power transformation function (c.f. (2.1)); see, e.g., Carroll, Wu, and Ruppert (1988).

To see how SIR helps the residual analysis, we take $g(x) = .2x$ and generate 100 cases for $p = 6$ with

$$\beta_1' = (1, 1, 1, 1, 0, 0), \beta_2' = (0, 0, 0, 0, 1, 1), \alpha = 3, \epsilon \sim N(0, 1)$$

Table 6.2: The first two eigenvectors(with standard deviations and ratios) and eigenvalues of SIR for (3.1) with error terms.

| | |
|---|---|
| *first vector* | (-.01, .06, .01, -.01, .02, .01, -.01, -.05, -.01, -.46, -.46, -.44, -.43, -.39, -.38) |
| *S.D.* | (.03, .03, .03, .03, .03, .03, .03, .04, .03, .03, .03, .03, .04, .03, .03) |
| *ratio* | (-0.4, 1.9, 0.4, -0.4, 0.5, 0.4, -0.4, -1.5, -0.3, -14, -14, -13, -12, -13, -11) |
| *second vector* | (.33, .31, .34, .27, .33, .32, .39, .22, .34, -.02, -.17, .09, .06, -.05, .14) |
| *S.D.* | (.05, .06, .05, .06, .05, .05, .05, .06, .06, .05, .05, .06, .06, .05, .06) |
| *ratio* | (6.1, 5.5, 7.0, 4.7, 6.0, 6.1, 7.2, 3.7, 6.1 -0.3, -3.1, 1.5, 1.0, -0.9, 2.5) |
| *eigenvalues* | (.78, .55, .17, .12, .11, .10, .07, .05, .04, .03, .02, .01, .01, .01, .00) |

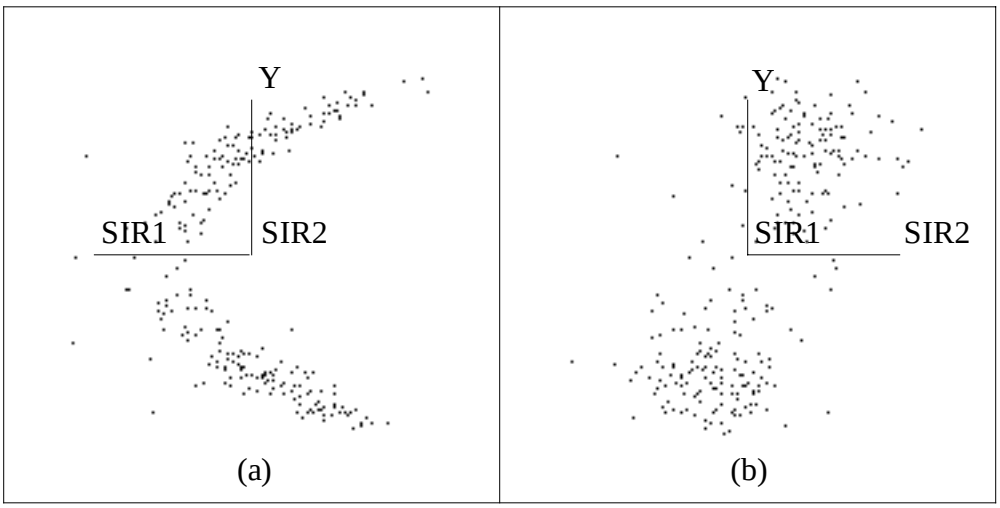


Figure 6.3: SIR's view of Data Generated From (3.1) with $\sigma_1 = \sigma_2 = 1$, $p = 15$

Fit the data by the usual linear least squares and find the residual $r$. Since $\beta_1'\mathbf{x}$ is uncorrelated with $\beta_2'\mathbf{x}$, the heteroscedasticity occurs along a direction orthogonal to the direction of the best linear fit. Thus we do not anticipate to find any pattern by examining the usual residual plot, the plot of $Y$ against predicted values (see Figure 6.4 (a) ).

Now we run SIR on $r$; see Table 6.3. Figure 6.4(b) gives the plot of $r$ against the first direction found by SIR. It does reveal the heteroscedasticity pattern well.

The reason why SIR can help in residual analysis is easy to understand. Although $r$ is, by definition, uncorrelated with $\mathbf{x}$, we can apply transformation on $r$ to increase the correlation and SIR does that in an "optimal" way. There is no need to take the absolute value transformation on $r$ before applying SIR. The flexibility in allowing for non-monotone transformation is the key to the success.

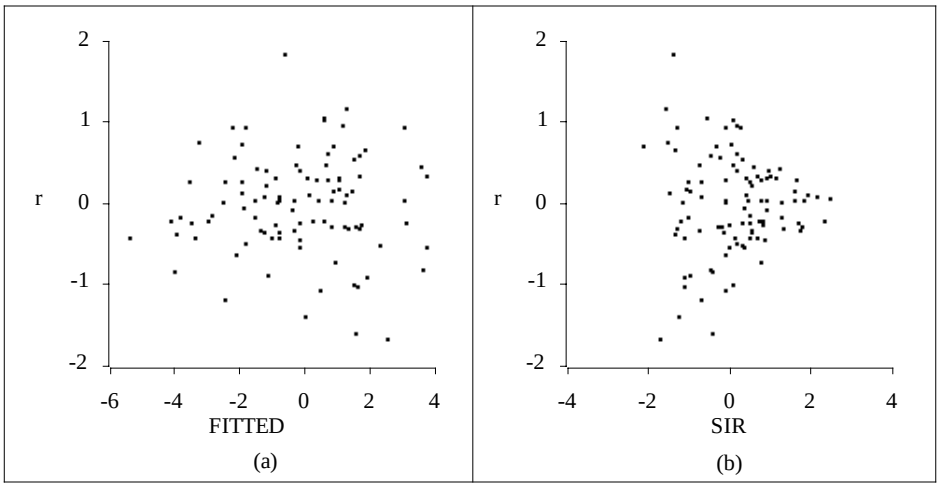


Figure 6.4: Residuals against Linear Squares Fit(a) and Direction(b) of Model(3.2).

Table 6.3: The first eigenvector(with standard deviations and ratios) and eigenvalues of SIR for residuals of (3.2).

| | |
|---|---|
| *first eigenvector* | (-.05, -.04, .18, .06, -.71, -.79) |
| *S.D.* | (.13, .17, .13, .14, .13, .14) |
| *ratio* | (-0.4, -0.2, 1.4, 0.5, -5.4, -5.5) |
| *eigenvalues* | (.37, .23, .13, .07, .03, .01) |

### 6.3.3 Horseshoe and helix.

A five-dimensional input variable $\mathbf{x} = (x_1, ..., x_5)'$ is obtained by first generating 1000 cases for $\mathbf{x}$ from the standard normal distribution and then retaining only those cases that satisfy the constraint :

$$x_1^2 - 0.5 < x_2 < x_1^2 + 0.5 \tag{3.3}$$

This reduces the sample size to 296. Now a linear model is used to generate $Y$

$$Y = x_1 + 0.5\epsilon, \ \epsilon \sim N(0, 1) \tag{3.4}$$

The output of SIR shows two large eigenvalues; see Table 6.4. Figures 6.5(a)-(d) are some static pictures of the rotational plot found by SIR. By rotating the plot about the vertical axis, we find data points spinning like a helix or a slide.

The first direction shows a linear pattern (Figure 6.5(a)) and the second direction finds a curve (Figure 6.5(c)). They correspond to $x_1$ and $x_2$ approximately. The scatterdiagram of these two SIR directions, Figure 6.5 (d), shows a horseshoe pattern, exhibiting the quadratic constraint (3.3). In this example, $x_2$ is nonlinearly correlated with $x_1$, a situation where

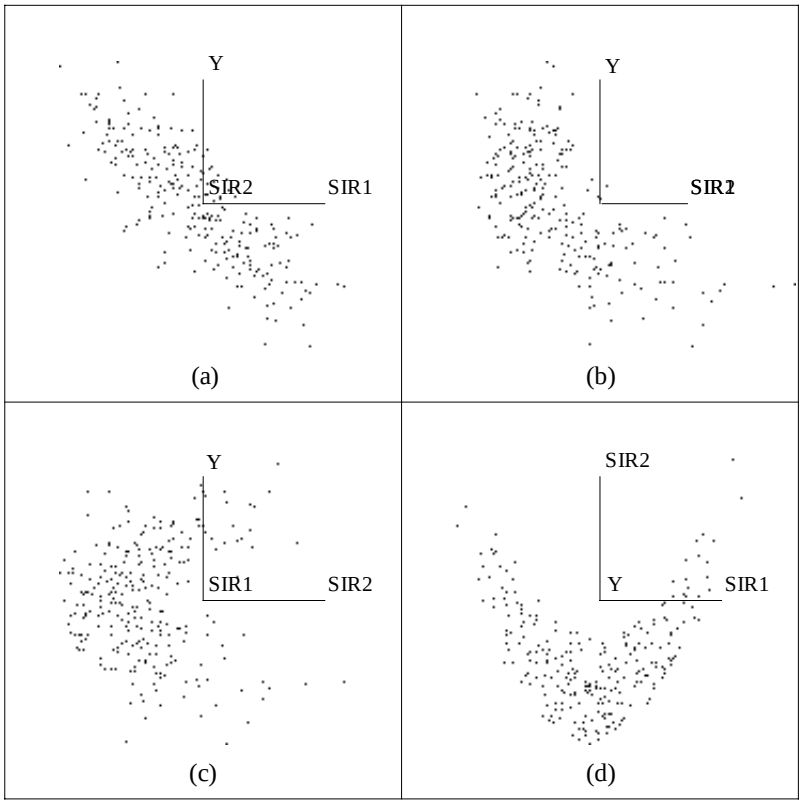


Figure 6.5: SIR's View of Data Generated From (3.3)-(3.4).

*Condition (1.3)* is severely violated. SIR picks up this additional direction because $Y$ can be transformed to retain a significant correlation with $x_2$.

A data set with a pattern like the one just observed here creates some difficulties in modeling which have not received proper attention in the literature. First of all, we may not be able to tell if the number of the components is one or two. For example, a data generated by a two-components model of the form

$$Y = sign(x_1)\sqrt{|x_2|} + .5\epsilon$$

presents little visual difference from the one we just find. In addition, even if a one-component model is assumed, we may not have much information to estimate the correct direction well without knowing the correct functional form.

Perhaps exhibiting this low dimensional nonlinear confounding patterns is scientifically more important than attempting to resolving this issue statistically. Graphics gives scientists something to focus on. It helps stimulate relevant knowledge.

Table 6.4: The first two eigenvectors(with standard deviations and ratios) and eigenvalues of SIR for (3.3), (3.4).

| | |
|---|---|
| *first eigenvector* | (-1.64, .10, .02, -.03, .01) |
| *S.D.* | (.07, .09, 0.04, 0.04, .04) |
| *ratio* | (-23, 1.1, 0.5, -0.6, 0.3) |
| *second eigenvector* | (.16, 2.1, .06, -.01, -.02) |
| *S.D.* | (.15, .19, .09, 0.09, 0.09) |
| *ratio* | (1.0 11 0.6 -0.1 -0.2) |
| *eigenvalues* | (.66, .30, .056, .023, .01) |

## 6.4 Simple estimates for the standard deviations of the SIR directions.

Outputs from multiple linear regression(MLR) software often attach an estimated standard deviation (i.e. standard error) to each regression coefficient. With that, users can easily form the t-ratio (= the ratio of the coefficient estimate to the standard error) for a quick assessment on the (statistical) significance of each regressor variable. It would be desirable if SIR outputs can provide similar information. But the asymptotics for SIR is more difficult than MLR. The formulae for the covariance matrix of each eigenvector $\hat{v}_i$ can be derived by combining some perturbation results for eigenvalue decomposition with large sample probabilistic argument. For general cases, they appear complex and hard to interpret. However, the transformation theory in Section 1 offers a clue for simplification in practical use.

As it turns out, our formula is similar to the familiar one in MLR. For the $i$th SIR direction $\hat{v}_i$, we may attach it with the vector of the squared root of the diagonal elements from the matrix

$$\frac{(1-\hat{\lambda}_i)}{\hat{\lambda}_i} \cdot n^{-1}\Sigma_{\mathbf{x}}^{-1}$$

as the estimated standard deviations. This formula brings out three messages useful to bear in mind:

(m.1) The standard errors of a SIR direction are proportional to those for the standard MLR of $Y$ on $\mathbf{x}$.

(m.2) The inaccuracy of a SIR direction gets greater when the corresponding eigenvalue gets smaller.

(m.3) The ratio $\frac{(1-\hat{\lambda}_i)}{\hat{\lambda}_i}$ plays the role of the average of squared residuals in MLR.

To see how the transformation theory is used for suggesting our formula, first recall from the familiar least squares theory:

$$cov(\hat{\beta}_{ls}) = \sigma^2 \cdot n^{-1}\Sigma_{\mathbf{x}}^{-1}, \tag{4.1}$$

This formula remains popular for practical use even if MLR is conducted after a transformation of $Y$, albeit the controversy regarding whether the effect of transformation can be

ignored or not; Bickel and Doksum(1981), Box and Cox(1964), Hinkley and Runger (1984). Since we can interpret the SIR directions as being proportional to the MLR slope estimate after optimal transformation (Theorem 3.2), (m.1) is well-anticipated. It remains to explain (m.3). Suppose the optimal transformation $T_i(Y)$ were given and we conduct the standard MLR for the transformed $Y$ values. Let $\tilde{b}(T_i)$ be the estimate of the slope vector $b(T_i)$. Recall (3.6) : SIR eigenvector $v_i$ can be obtained from $b(T_i)$ after dividing by the constant $\lambda_i$. This suggests that the covariance matrix of the SIR estimate $\hat{v}_i$ should be equal to the covariance matrix of $\tilde{b}(T_i)$ divided by $\lambda_i^2$. Now apply (4.1) to find out $cov(\tilde{b}(T_i))$. Since the R-squared value of the regression is $\lambda_i$ as stated in Theorem 3.2, the residual variance $\sigma^2$ in (4.1) must be equal to $(1-\lambda_i)var(T_i(y)) = (1-\lambda_i)\lambda_i$. Finally dividing $\sigma^2$ by $\lambda_i^2$, we are led to the ratio $\frac{(1-\hat{\lambda}_i)}{\hat{\lambda}_i}$ given in (m.3).

Like the t-ratios in MLR, the ratios of the SIR estimates over the respective standard errors provide a convenient way to tell if the corresponding coefficients are statistically significant or not. In Appendix B, rigorous asymptotics will be developed for justifying such applications. More specifically, for the $l-$th regressor variable, we may test the null hypothesis $\mathbf{H}_o$ :

$$\mathbf{H}_o : e_l'\beta_i = 0, i = 1, \cdots, k \tag{4.2}$$

where $e_l = (0, \cdots, 0, 1, \cdots, 0)'$ denotes the $l$th basis vector. The standard error we obtained is asymptotically valid under the null hypothesis (4.2).

As a cautionary note, our formula are not valid for constructing confidence intervals. In general, the standard deviations of SIR estimates depend on the true parameters in a rather complex manner. This complexity is largely due to the additional uncertainty caused by approximating the $v_i$ with $\hat{v}_i$ in estimating the transformation $T_i(Y)$; a phenomenon similar to the problem of Bickel and Doksum(1981). Thus it remains unclear how close to the correct ones our simplified standard deviations are.

In deriving the asymptotic distribution, we have also asssumed that the number of slices used in constructing SIR estimate is fixed Although in theory we can use as many as $H = n/2$ slices (Hsing and Carroll 1992), practically we find no obvious advantage in using large $H$.