the curse of dimensionality problem. But of course, the most important question remaining is how to relate inverse regression to forward regression. To fill up the gap, we shall derive Theorem 2.1 in section 2.6, which is the foundation of the SIR theory.

Generally speaking, inverse regression factorizes the joint density of $\mathbf{x}$ and Y into the condition density $h(\mathbf{x}|y)$ and the marginal density $k(y)$. While only $E(\mathbf{x}|Y)$ is considered in this chapter, other quantities can be utilized as well. For example, in later chapters, we shall also discuss how to use conditional covariance $cov(\mathbf{x}|Y = y)$ for extending the basic SIR algorithm.

2.2 An algorithm of SIR.

Let $(y_1, \mathbf{x}_1), \cdots, (y_n, \mathbf{x}_n)$ be the original data set with $(p+1)$ variables and n cases. Imagine that they have been stored as illustrated in Table 2.1. The algorithm of SIR consists of the following steps.

Table 2.1: ORIGINAL DATA SET

Y_1	$\mathbf{x}_1(=(x_{11}, x_{12}, \cdots, x_{1p})')$
Y_2	$\mathbf{x}_2(=(x_{21}, x_{22}, \cdots, x_{2p})')$
Y_3	$\mathbf{x}_3(=(x_{31}, x_{32}, \cdots, x_{3p})')$
Y_4	$\mathbf{x}_4(=(x_{41}, x_{42}, \cdots, x_{4p})')$
Y_5	$\cdots\cdots\cdots$
.	$\cdots\cdots\cdots$
.	$\cdots\cdots\cdots$
.	$\cdots\cdots\cdots$
Y_n	$\mathbf{x}_n(=(x_{n1}, x_{n2}, \cdots, x_{np})')$

Table 2.2: SORTING by Y and SLICING.

$Y_{(1)}$	$\mathbf{x}_{(1)}(=(x_{(1)1}, x_{(1)2}, \cdots, x_{(1)p})')$
$Y_{(2)}$	$\mathbf{x}_{(2)}(=(x_{(2)1}, x_{(2)2}, \cdots, x_{()2p})')$
$Y_{(3)}$	$\mathbf{x}_{(3)}(=(x_{(3)1}, x_{(3)2}, \cdots, x_{(3)p})')$
$Y_{(4)}$	$\mathbf{x}_{(4)}(=(x_{(4)1}, x_{(4)2}, \cdots, x_{(4)p})')$
$Y_{(5)}$	$\cdots\cdots\cdots$
.	$\cdots\cdots\cdots$
.	$\cdots\cdots\cdots$
.	$\cdots\cdots\cdots$
$Y_{(n)}$	$\mathbf{x}_{(n)}(=(x_{(n)1}, x_{(n)2}, \cdots, x_{(n)p})')$

Step 1. Sort the data by Y. This is illustrated by Table 2.2.

Step 2. Divide the data set into H **slices as equally as possible**. Let n_h be the number of cases in slice h. In Table 2.2, slices are separated by bold lines. The number of slices H is a user-specified parameter. For example, we find between 10 to 20 slices to be reasonable for a sample of size $n = 300$. As to be discussed later, there are theoretical results indicating that SIR outputs do not change much for a wide range of H.

Step 3. Within each slice, compute the sample mean of x, $\bar{\mathbf{x}}_h = n_h^{-1} \sum_{(i)\in \text{slice}\, h} \mathbf{x}_{(i)}$. Table 2.3 shows the slice means for both Y and $\mathbf{x}$. Note that SIR uses Y values only to create slices. Once slices are formed, they can be discarded. Thus although the slice means of Y are shown in Table 2.3, they need not be computed.

Table 2.3: Slice means.

$\bar{Y}_1$	$\bar{\mathbf{x}}_1 (= (\bar{x}_{11}, \bar{x}_{12}, \cdots, \bar{x}_{1p})')$
$\bar{Y}_2$	$\bar{\mathbf{x}}_2 (= (\bar{x}_{21}, \bar{x}_{22}, \cdots, \bar{x}_{2p})')$
.	.....
.	.....
.	.....
$\bar{Y}_H$	$\bar{\mathbf{x}}_H (= (\bar{x}_{H1}, x_{H2}, \cdots, x_{Hp})')$

Step 4. Compute the covariance matrix for the slice means of x, weighted by the slice sizes:

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

Here $\bar{\mathbf{x}}$ denotes sample mean of $\bar{\mathbf{x}} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i$.

Step 5. Compute the sample covariance for $\mathbf{x}_i$**'s,** $\hat{\Sigma}_{\mathbf{x}} = n^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$.

Step 6. Find the SIR directions by conducting the eigenvalue decomposition of $\hat{\Sigma}_\eta$ **with respect to** $\hat{\Sigma}_{\mathbf{x}}$:

$$\begin{aligned} &\hat{\Sigma}_\eta \hat{\beta}_i = \hat{\lambda}_i \hat{\Sigma}_{\mathbf{x}} \hat{\beta}_i \\ &\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \end{aligned} \tag{2.1}$$

The i-th eigenvector $\hat{\beta}_i$ is called the i-th SIR direction. The first few SIR directions can be used for dimension reduction. They serve as the coefficients linking the input nodes to the intermediate nodes in Figure 1.7. For further analysis, the following additional steps are helpful.

Step 7. Project x along the SIR directions; that is, use each SIR direction to form a linear combination of x. We shall call $\hat{\beta}_1' \mathbf{x}$ the first SIR variate, $\hat{\beta}_2' \mathbf{x}$ the second SIR variate, and so on. Table 2.4 shows the reconstructed data after projection. Compared to Table 2.1, this amounts to only a change in the coordinate system of the regressor.

Table 2.4: Y and SIR variates

Y_1	$\hat{\beta}_1'\mathbf{x}_1, \hat{\beta}_2'\mathbf{x}_1, \cdots$
Y_2	$\hat{\beta}_1'\mathbf{x}_2, \hat{\beta}_2'\mathbf{x}_2, \cdots$
Y_3	$\hat{\beta}_1'\mathbf{x}_2, \hat{\beta}_2'\mathbf{x}_3, \cdots$
Y_4	$\hat{\beta}_1'\mathbf{x}_4, \hat{\beta}_2'\mathbf{x}_4, \cdots$
Y_5	$\cdots\cdots\cdots$
.	$\cdots\cdots\cdots$
.	$\cdots\cdots\cdots$
.	$\cdots\cdots\cdots$
Y_n	$\hat{\beta}_1'\mathbf{x}_n, \hat{\beta}_2'\mathbf{x}_n, \cdots$

Step 8. Plot Y against the SIR variates. These 2-D or 3-D plots offer a graphical summary useful for revealing the regression structure. We shall argue that under fairly general conditions, these plots are more informative than other scatterplots of Y against any projections of $\mathbf{x}$.

2.3 SIR and principal component analysis.

It is easier to remember the eigenvalue decomposition step of SIR by standarizing $\mathbf{x}$ before analysis. For now, suppose that the covariance of $\mathbf{x} = (x_1, \cdots, x_p)'$ is an identity matrix I. In other words, all regressor variables x_i, $i = 1, \cdots, p$ have the same variance (=1) and are uncorrelated with each other. Then on the rightside of the equality in (2.1), the matrix $\hat{\Sigma}_{\mathbf{x}}$ can be removed. Thus Step 6 is merely the principal component analysis applied to the slice means of $\mathbf{x}$ in Table 2.3. We can summarize SIR as follows: (1) partitioning the cases into H groups according to the Y values; (2) finding the H slice means of $\mathbf{x}$; (3) applying a principal component analysis on slice means of $\mathbf{x}$.

It is important to remember that our use of principal component analysis differs from the conventional way. We use Y to form slices while the conventional way did not use any information from Y at all.

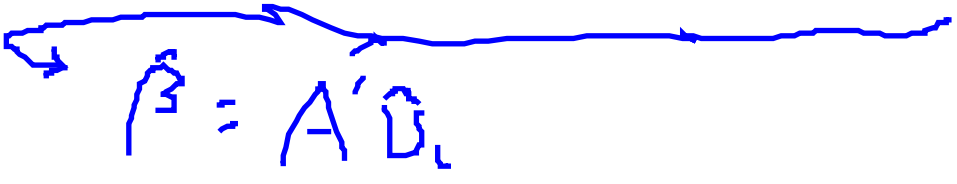
SIR is invariant under affine transformation of $\mathbf{x}$. We can always find a new coordinate to standardize $\mathbf{x}$ first. Suppose A is an invertible matrix so that

$$\mathbf{z} = A\mathbf{x}, \; cov(\mathbf{z}) = I$$

An example is to take A as $\hat{\Sigma}_{\mathbf{x}}^{-1/2}$; but there are better ones for saving time in computing. The covariance matrix for the slice means of $\mathbf{z}$ is equal to $A\hat{\Sigma}_\eta A'$. Let $\hat{v}_i$ be the i-th eignevalue :

$$A\hat{\Sigma}_\eta A'\hat{v}_i = \hat{\lambda}_i \hat{v}_i$$

Multiplying both sides by A^{-1} and using the relationship that $A\hat{\Sigma}_{\mathbf{x}}A' = I$, we can rewrite the above equation as $\hat{\Sigma}_\eta(A'\hat{v}_i) = \hat{\lambda}_i \hat{\Sigma}_{\mathbf{x}}(A'\hat{v}_i)$. Now comparing with (2.1), we see that



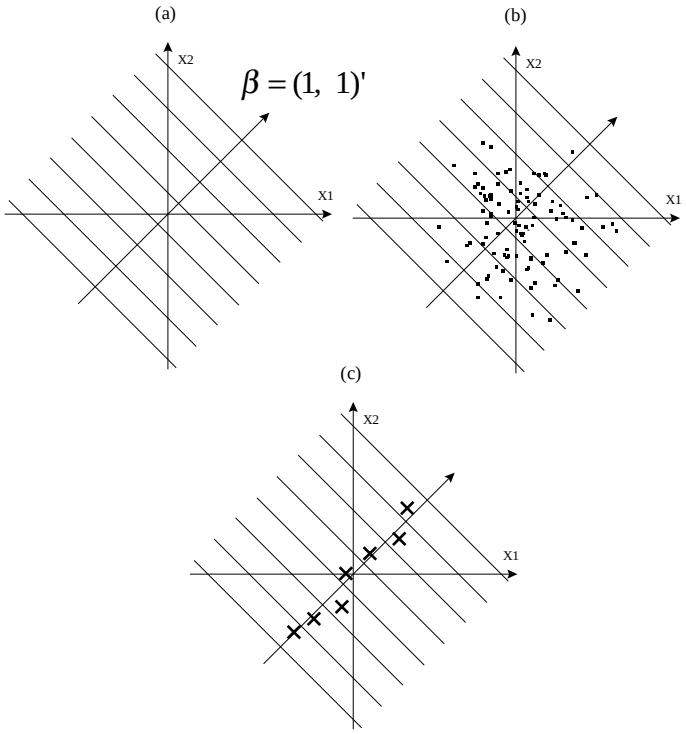


Figure 2.6: Contour Plot of $y = g(\beta_{\mathbf{x}}')$ and SIR.

dimensional affine subspace which can be related to the e.d.r. space.

Theorem 2.1. *Under Condition (1.1) of Chapter 1 and the Linear design Condition (2.2) to be given next, the centered inverse regression curve* $E(\mathbf{x}|y) - E\mathbf{x}$ *is contained in the linear subspace spanned by* $\Sigma_{\mathbf{x}}\beta_k$, $k = 1, ..., K$, *where* $\Sigma_{\mathbf{x}}$ *denotes the covariance matrix of* $\mathbf{x}$.

(L.D.C.) Linear Design Condition.

$$\text{For any } b \text{ in } R^p, \text{ the conditional expectation } E(b'\mathbf{x}|\beta_1'\mathbf{x}, ..., \beta_K'\mathbf{x}) \text{ is linear in } \beta_1\mathbf{x}, \cdots, \beta_K'\mathbf{x}; \text{ that is, for some constants } c_o, c_1, ..., c_K,$$
$$E(b'\mathbf{x}|\beta_1'\mathbf{x}, ..., \beta_K'\mathbf{x}) = c_o + c_1\beta_1'\mathbf{x} + ... + c_K\beta_K'\mathbf{x}. \tag{2.2}$$

(L.D.C.) is satisfied when the distribution of $\mathbf{x}$ is elliptically symmetric; for example, the normal distribution. But elliptic symmetry is **NOT** is necessary condition for **(L.D.C)** to hold. As to be discussed in a later chapter, this condition is not as restrictive as it may appear. We shall argue that the violation of this condition is often mild and the bias of SIR is not large.

A proof of Theorem 2.1 will be given in Section 2.7.

It is easier to remember Theorem 2.1 for the special case that $\mathbf{x}$ has been standardized to $\mathbf{z} = A\mathbf{x}$ by some invertible matrix A so that $E\mathbf{z} = 0, cov(\mathbf{z}) = I$. We can rewrite (1.1) of chapter 1 as

$$Y = g(\theta_1'\mathbf{z}, \cdots, \theta_K'\mathbf{z}, \epsilon) \tag{2.3}$$

where $\theta_i' = \beta_i' A^{-1}$.

Corollary 2.1. *Assume that* **(L.D.C.)** *holds. Then for model (2.3), the standardized inverse regression curve $E(\mathbf{z}|Y = y)$ is contained in the space spanned by the standarized e.d.r. directions $\theta_i, i = 1, \cdots, K$.*

As a random vector, $E(\mathbf{z}|Y)$ has a covariance matrix $cov(E(\mathbf{z}|Y))$. By Corrollary 2.1, this matrix is seen to be degenerate in any direction orthogonal to the θ_k's. Therefore, the eigenvalue decomposition

$$\begin{aligned} cov(E(\mathbf{z}|Y))v_i &= \lambda_i v_i, \qquad i = 1, \cdots, p \\ \lambda_1 &\geq \cdots, \geq \lambda_p \end{aligned} \tag{2.4}$$

must give no more than K nonzero eignevalues. All eigenvectors v_i with nonzero eigenvalues must fall into the standardized e.d.r. space.

Denote the random vector $E(\mathbf{x}|Y)$ by η and the covariance matrix of η by Σ_η;

$$\Sigma_\eta = Cov(E(\mathbf{x}|Y))$$

Since $Cov(E(\mathbf{z}|Y)) = Cov(E(A\mathbf{x}|Y)) = A\,Cov(E(\mathbf{x}|Y))A' = A\Sigma_\eta A'$, (2.4) can be written as

$$A\Sigma_\eta A' v_i = \lambda_i v_i$$

Multiplying both sides by A^{-1}, this gives

$$\Sigma_\eta(A'v_i) = \lambda_i A^{-1} v_i = \lambda_i (A^{-1}(A')^{-1})(A'v_i)$$

Denote $A'v_i$ by $\mathbf{b}_i$. We have derived the following eigenvalue decomposition :

$$\begin{aligned} \Sigma_\eta \mathbf{b}_i &= \lambda_i \Sigma_{\mathbf{x}} \mathbf{b}_i \\ \lambda_1 &\geq \cdots \geq \lambda_p \end{aligned} \tag{2.5}$$

We shall refer to the eigenvalue decomposition (2.5) as the population version of SIR. There are no more than K non-zero eigenvalues. We shall call eigenvector $\mathbf{b}_i$ the population version of a SIR direction (for $\lambda_i \neq 0$only). Since v_i falls into the standardized e.d.r. space, it is a linear combination of $\theta_k = A^{-1}\beta_k, k = 1, \cdots, K$. Therefore, $\mathbf{b}_i$ can be written as linear combination of $\beta_k, k = 1, \cdots, K$. The following corollary is a summary of this conclusion. It establishes the Fisher consistency for the population version of SIR.

Corollary 2.2. *Assume that* **(L.D.C)** *holds. Then for model (1.1) of Chapter 1, the population version of the SIR direction $\mathbf{b}_i$ falls into the e.d.r. space.*

It is easy to compare the population version of SIR (2.5) with the sample version (2.1). We can interpret the slice mean $\mathbf{x}_h$ obtained at Step 2 of the SIR algorithm in Section 2.1

as an estimate of $E(\mathbf{x}|Y = y)$ for y falling within the interval associated with slice h. The matrix $\hat{\Sigma}_\eta$ given in Step 3 is a natural estimate for the covariance matrix Σ_η.

Remark. We estimate $E(\mathbf{x}|Y)$ by a step function consisting of $\mathbf{x}_h, h = 1, \cdots, H$. It is feasible to use more sophisticated nonparametric regression methods such as kernel, nearest neighbor, or smoothing splines to yield a better estimate of the inverse regression curve. This is especially attractive for relatively small samples. However, intuitively speaking, since we only need the main orientation (but not any other detailed aspects) of the estimated curve, possible gains due to smoothing are not likely to be substantial for large samples.

2.7 Proof of Theorem 2.1

We shall give the proof for the case the $K = 1$ first. Assume that $E\mathbf{x} = 0$ without loss of generality. We want to show that the vector $E(\mathbf{x}|y)$ is proportional to $\Sigma_\mathbf{x}\beta$. The key argument is by conditioning :

$$E(\mathbf{x}|y) = E(E(\mathbf{x}|\beta'\mathbf{x}, \epsilon)|y) = E(E(\mathbf{x}|\beta'\mathbf{x})|y) \tag{2.6}$$

Now the **(L.D.C)** together with the assumption $E\mathbf{x} = 0$, implies that

$$\begin{aligned} E(\mathbf{x}|\beta'\mathbf{x}) &= [(var(\beta'\mathbf{x}))^{-1}cov(\mathbf{x}, \beta'\mathbf{x})]\beta'\mathbf{x} \\ &= [(var(\beta'\mathbf{x}))^{-1}\Sigma_\mathbf{x}\beta]\beta'\mathbf{x} \end{aligned} \tag{2.7}$$

Here the term inside the brackets following the first equality is a simple application of the formula for the slope of the simple linear regression of each component of $\mathbf{x}$ against the variable $\beta'\mathbf{x}$. Let $k(y) = (var(\beta'\mathbf{x}))^{-1}E(\beta'\mathbf{x}|y)$. It follows that

$$E(\mathbf{x}|y) = k(\beta')|y)\Sigma_\mathbf{x}\beta$$

which is proportional to $\Sigma_\mathbf{x}\beta$ as desired.

For the case that K is larger than 1, a formula for $E(\mathbf{x}|\beta_1'\mathbf{x}, \cdots, \beta_K'\mathbf{x})$ is also not hard to find. First let $B = (\beta_1, \cdots, \beta_K)$, which is a p by K matrix.

$$\begin{aligned} E(\mathbf{x}|\beta_1'\mathbf{x}, \cdots, \beta_K'\mathbf{x}) &= E(\mathbf{x}|B'\mathbf{x}) \\ &= [cov(B'\mathbf{x})^{-1}cov(B'\mathbf{x}, \mathbf{x})]'B'\mathbf{x} \\ &= [(B'\Sigma_\mathbf{x}B)^{-1}B'\Sigma_\mathbf{x}]'B'\mathbf{x} \end{aligned} \tag{2.8}$$

Here the first bracket term is due to the multiple linear regression of each component of $\mathbf{x}$ separately against the K-dimensional variable $B'\mathbf{x}$. Let $\mathbf{k}(y) = (B'\Sigma_\mathbf{x}B)^{-1}E(B'\mathbf{x}|y)$, a K dimensional vector for each fixed y. Then from (2.6) we see that

$$E(\mathbf{x}|y) = (\Sigma_\mathbf{x}B)\mathbf{k}(y)$$

which shows that $E(\mathbf{x}|y)$ falls into the linear space generated by $\Sigma_\mathbf{x}\beta_k$'s, as desired.