Chapter 3

Sliced Inverse Regression

3.1 Sliced Inverse Regression: Population-Level Development

Sliced Inverse Regression (SIR), introduced by Li (1991), is the first and most commonly known Sufficient Dimension Reduction estimator. The term "inverse regression" refers to the conditional expectation $E(X|Y)$. The word "inverse" is used because, in usual regression analysis, what is of interest is the conditional mean $E(Y|X)$. The word "slice" refers to the fact that we estimate the conditional mean $E(X|Y)$ by taking an interval of Y.

Similar to the development in Section 1.10, here we also require linearity of conditional mean of the form $E(X|\beta^{\mathsf{T}}X)$, more general in form than Assumption 1.1.

Assumption 3.1 *Let $\beta \in \mathbb{R}^{p\times d}$ be a matrix such that* $\text{span}(\beta) = \mathscr{S}_{Y|X}$. *We assume that $E(X|\beta^{\mathsf{T}}X)$ is a linear function of the d-dimensional random vector $\beta^{\mathsf{T}}X$.*

The next lemma is a generalization of Lemma 1 in Section 1.10; its proof is omitted.

Lemma 3.1 *Suppose $\beta^{\mathsf{T}}\Sigma\beta$ is positive definite. Then*

$$E[X - E(X)|\beta^{\mathsf{T}}X] = P_{\beta}^{\mathsf{T}}(\Sigma)[X - E(X)].$$

A random vector is said to have an elliptical distribution if there is a positive definite matrix A such that the density of X depends on x only through $x^{\mathsf{T}}Ax$; that is,

$$f_X(x) = h(x^{\mathsf{T}}Ax)$$

for some function $h : \mathbb{R} \to \mathbb{R}$. The following result was proved in Eaton (1986).

Lemma 3.2 *If X is integrable and has an elliptical distribution then $E(X|B^\top X)$ is linear in B for any matrix B. If $E(X|v^\top X)$ is linear in $v^\top X$ for each $v \in \mathbb{R}^p$, then X has an elliptical distribution.*

We now prove that SIR is unbiased.

Theorem 3.1 *Suppose X is square-integrable and $\Sigma = \mathrm{var}(X)$ is nonsingular. Then, under Assumption 3.1,*

$$\Sigma^{-1}[E(X|Y) - E(X)] \in \mathscr{S}_{Y|X}.$$

PROOF. First, assume $E(X) = 0$. Then

$$\begin{aligned} E(X|Y) &= E(E(X|\beta^\top X, Y)|Y) \\ &= E(E(X|\beta^\top X)|Y) \\ &= E(P_\beta^\top(\Sigma)X|Y) \\ &= P_\beta^\top(\Sigma)E(X|Y), \end{aligned} \tag{3.1}$$

where the second equality holds because $Y \perp\!\!\!\perp X|\beta^\top X$; the third equality follows from Assumption 3.1 and Lemma 3.1. Because $P_\beta^\top(\Sigma) = \Sigma P_\beta(\Sigma)\Sigma^{-1}$, the right-hand side of (3.1) can be rewritten as $\Sigma P_\beta(\Sigma)\Sigma^{-1}E(X|Y)$, and consequently,

$$E(X|Y) = \Sigma P_\beta(\Sigma)\Sigma^{-1}E(X|Y).$$

Hence $\Sigma^{-1}E(X|Y) \in \mathrm{span}(P_\beta(\Sigma)) = \mathscr{S}_{Y|X}$. In the case where $E(X) \neq 0$, we simply apply the above result to $X - E(X)$. □

The above result shows that, under the linear conditional mean assumption, we can recover the gradient direction of the regression function $E(Y|X)$, or more generally the conditional distribution $F_{Y|X}$, by estimating the inverse conditional expectation $E(X|Y)$. This is significant because, to estimate the forward conditional moment $E(Y|X)$ nonparametrically, we need to smooth over a p-dimensional space; but to estimate $E(X|Y)$ nonparametrically, we only need to smooth over 1-dimensional space – the space of Y, which is much more accurate than high-dimensional smoothing. This is similar to the phenomenon observed in Section 1.10, where the symmetry in the distribution of X allows us to use linear regression to recover the gradient direction of $E(Y|X)$, thus avoiding fitting a high-dimensional surface nonparametrically.

Since we do not know the true β, Assumption 3.1 cannot be verified. Hence, in practice, we replace Assumption 3.1 by the stronger assumption that X has an elliptical distribution. We can intuitively verify the elliptical distribution condition by looking at the scatter plot matrix of the components of X, as we do in Section 3.4, Figure 3.3. Even though elliptical shape of the distribution of (X_i, X_j) for all $i \neq j$ does not imply the ellipticity of the joint distribution of $X = (X_1, \ldots, X_p)$, the scatter plot matrix is often a simple and effective way of detecting non-ellipticity.

Figure 3.1 illustrates how the vector $E(X|Y) - E(X)$ recovers the central subspace in the special case where $\Sigma = I_p$. When $E(Y|X) = E(Y|\beta^\top X)$, the vector $E(X|Y) - E(X)$ is aligned with the direction of β as long as the distribution of X is symmetric about the direction of β, which is guaranteed by the linear conditional mean condition.

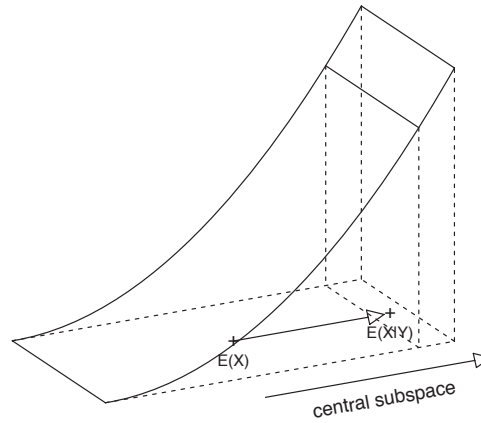

Figure 3.1 *Illustration of unbiasedness of SIR.*

Corollary 3.1 *Under the assumptions of Theorem 3.1,*

$$\operatorname{span}(\Sigma^{-1}\operatorname{cov}[E(X|Y)]\Sigma^{-1}) \subseteq \mathcal{S}_{Y|X}.$$

PROOF. Let $U = E(X|Y) - E(X)$. We need to $\operatorname{span}[\Sigma^{-1}E(UU^{\mathsf{T}})\Sigma^{-1}] = \mathcal{S}_{Y|X}$. Since Σ is nonsingular, it is equivalent to

$$\begin{aligned}\operatorname{span}[\Sigma^{-1}E(UU^{\mathsf{T}})] \subseteq \mathcal{S}_{Y|X} &\Rightarrow \operatorname{span}[E(UU^{\mathsf{T}})] \subseteq \Sigma\mathcal{S}_{Y|X} \\ &\Rightarrow (\Sigma\mathcal{S}_{Y|X})^{\perp} \subseteq \operatorname{span}[E(UU^{\mathsf{T}})]^{\perp}.\end{aligned}$$

Now let $v \in (\Sigma\mathcal{S}_{Y|X})^{\perp}$. Because, by Theorem 3.1, $U \in \Sigma\mathcal{S}_{Y|X}$, we have $(\Sigma\mathcal{S}_{Y|X})^{\perp} \subseteq \operatorname{span}(U)^{\perp}$. Hence $v \perp U$, which implies $E(v^{\mathsf{T}}U)^2 = 0$, or $v^{\mathsf{T}}E(UU^{\mathsf{T}})v = 0$. Hence $v \in \operatorname{span}[E(UU^{\mathsf{T}})]^{\perp}$. □

Let Λ_{SIR} denote the matrix $\operatorname{cov}[E(X|Y)]$. This corollary implies that we can use the column space of

$$\mathcal{S}_{\text{SIR}} = \operatorname{span}(\Sigma^{-1}\Lambda_{\text{SIR}}\Sigma^{-1})$$

to recover at least a part of the central subspace. This can be formulated as solving a generalized eigenvalue problem, as described in Section 1.3. That is, $\mathcal{S}_{\text{SIR}}$ is spanned by the set of eigenvectors $\{v : \Lambda_{\text{SIR}}v = \lambda\Sigma v, \lambda > 0\}$ in $\text{GEV}(\Lambda_{\text{SIR}}, \Sigma)$. This means we

first solve the standard eigenvalue problem $\mathrm{GEV}(\Sigma^{-1/2}\Lambda_{\mathrm{SIR}}\Sigma^{-1/2}, I_p)$:

$$A = \{u : \Sigma^{-1/2}\Lambda_{\mathrm{SIR}}\Sigma^{-1/2} = \lambda v, \lambda > 0\},$$

and then recover $\mathscr{S}_{\mathrm{SIR}}$ by the set of transformed eigenvectors $\{\Sigma^{-1/2}u : u \in A\}$.

We can easily extend this result as follows using the relation $\mathscr{S}_{g(Y)|X} \subseteq \mathscr{S}_{Y|X}$.

Corollary 3.2 *If $E(X|\beta^{\mathsf{T}}X)$ is linear in $\beta^{\mathsf{T}}X$ then for any measurable function $g(Y)$,*

$$\Sigma^{-1}[E(X|g(Y)) - E(X)] \in \mathscr{S}_{Y|X}.$$

3.2 Limitation of SIR

While the theory of the above section guarantees that $\mathscr{S}_{\mathrm{SIR}}$ is always a subspace of $\mathscr{S}_{Y|X}$, it says nothing about whether it can recover the entire central subspace or merely a proper subspace thereof. In this section we use an example to demonstrate the cases where SIR can fail to recover $\mathscr{S}_{Y|X}$ fully.

Suppose X is a p-dimensional random vector with $E(X) = 0$ and

$$Y = f(X_1) + \varepsilon,$$

where $\varepsilon \perp\!\!\!\perp X$, $\Sigma = I_p$, X has a spherical distribution, and f is symmetric about 0. In this case the central subspace is spanned by $(1,0,\ldots,0)$. We will show that $E(X|Y) = 0$. First, because f is symmetric, $f(-X_1) = f(X_1)$; so (Y, X_1) and $(Y, -X_1)$ have the same distribution. Hence

$$E(X_1|Y) = E(-X_1|Y), \quad \text{which implies} \quad E(X_1|Y) = 0.$$

For $i \neq 1$, because X has a spherical distribution and $X \perp\!\!\!\perp \varepsilon$, we have

$$(X_1, \ldots, X_p, \varepsilon) \stackrel{\mathscr{D}}{=} (X_1, \ldots, -X_i, \ldots, X_p, \varepsilon)$$

where $\stackrel{\mathscr{D}}{=}$ means the two sides of the equality have the same distribution. Hence

$$E(X_i|f(X_1) + \varepsilon) = E(-X_i|f(X_1) + \varepsilon), \quad \text{which implies} \quad E(X_i|Y) = 0.$$

Thus we conclude that $E(X|Y) = 0$.

We see that, in this special case, $\mathscr{S}_{\mathrm{SIR}} = \{0\}$. Even though the assertion $\mathscr{S}_{\mathrm{SIR}} \subseteq \mathscr{S}_{Y|X}$ is not violated, $\mathscr{S}_{\mathrm{SIR}}$ is useless as it does not provide any information about the central subspace.

The situation is illustrated by Figure 3.2, where the U-shaped surface represents $f(X_1)$. The conditional expectation $E(X|Y)$ is at the center of the set $A \cup B$, which is at the center of distribution of X, as represented by "+" in the figure. At the same time unconditional mean $E(X)$ is also at the center of the distribution of X, located at the same point represented by "+". Hence $E(X) - E(X|Y) = 0$. This limitation is one of the motivations of developing other SDR methods, such as the Sliced Average Variance Estimator (SAVE, Cook and Weisberg (1991)), Contour Regression (Li et al. (2005)), and Directional Regression (Li and Wang (2007)).

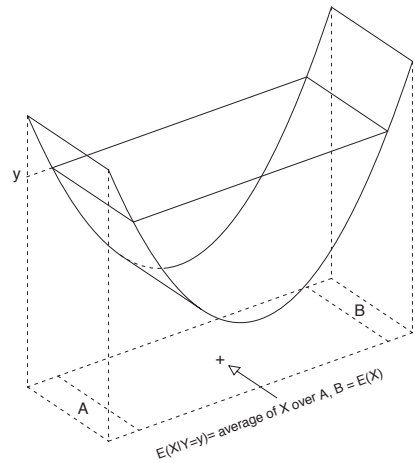

Figure 3.2 *Illustration of limitation of Sliced Inverse Regression.*

3.3 Estimation, Algorithm, and R-codes

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of independent observations on (X, Y). We first describe the estimation procedure for SIR at the population level. Let $J_1, \ldots, J_h$ be intervals in Ω_Y. Let

$$g(Y) = \sum_{\ell=1}^{h} \ell I(Y \in J_\ell).$$

That is, $g(Y)$ takes the value ℓ if Y falls in the ℓ interval. By Corollary 3.2, we have

$$E[X - E(X)|g(Y)] \in \Sigma \mathscr{S}_{Y|X}.$$

Let $Z = \Sigma^{-1/2}(X - EX)$. Let

$$\Lambda = \text{var}[E(Z|g(Y))].$$

Suppose Λ has rank r, which is at most d, the dimension of $\mathscr{S}_{Y|X}$. Let $v_1, \ldots, v_r$ be the eigenvectors of Λ corresponding to its nonzero eigenvalues. Then, by Theorem 2.2 of Chapter 2, $u_k = \Sigma^{-1/2} v_i$, $k = 1, \ldots, r$, belong the the central subspace $\mathscr{S}_{Y|X}$. The random variables

$$u_1^{\mathsf{T}}(X - E(X)), \ldots, u_r^{\mathsf{T}}(X - E(X))$$

are called sufficient predictors, and are the result of the SDR.

For estimation, we mimic the above process at the sample level. We summarize the estimation procedure as the following algorithm.

Algorithm 3.1 Sliced Inverse Regression

1. Compute the sample mean and sample variance:

$$\hat{\mu} = E_n(X), \quad \hat{\Sigma} = \mathrm{var}_n(X).$$

and compute the standardized random vectors

$$Z_i = \hat{\Sigma}^{-1/2}(X_i - \hat{\mu}), \quad i = 1, \ldots, n.$$

2. Approximate $E[Z|g(Y) \in J_\ell]$ or $E(Z|Y \in J_\ell)$ by

$$E_n(Z|Y \in J_\ell) = \frac{E_n[ZI(Y \in J_\ell)]}{E_n[I(Y \in J_\ell)]}, \quad \ell = 1, \ldots, h.$$

3. Approximate $\mathrm{var}[E(Z|g(Y))]$ by

$$\hat{\Lambda} = \sum_{i=\ell}^{h} E[I(Y \in J_\ell)] E_n(Z|Y \in J_\ell) E_n(Z^T|Y \in J_\ell). \tag{3.2}$$

4. Let $\hat{v}_1, \ldots, \hat{v}_r$ be the first r eigenvectors of $\hat{\Lambda}$, and let $\hat{\beta}_k = \hat{\Sigma}^{-1/2}\hat{v}_k$, $k = 1, \ldots, r$. The sufficient predictors are

$$\hat{\beta}_k^{\mathsf{T}}(X_1 - \hat{\mu}), \ldots, \hat{\beta}_k^{\mathsf{T}}(X_n - \hat{\mu}), \quad k = 1, \ldots, r.$$

Let S_{ik} represent the random variable $\hat{u}_k^{\mathsf{T}}(X_i - \hat{\mu})$, and let $S_i = (S_{i1}, \ldots, S_{ir})^{\mathsf{T}}$. We now have a sample of the lower dimensional predictor $S_1, \ldots, S_n$, which serves as a compressed version of the high-dimensional predictor $X_1, \ldots, X_n$. Under the premise of $r = d$, we can perform statistical analysis of Y versus S without losing information about the relation between Y and X. For example, we can perform regression analysis or classification based on the sample $(S_1, Y_1), \ldots (S_n, Y_n)$.

There are several issues that we will resolve in later chapters. For example, in the above algorithm r is assumed known, but in practice it must be estimated. Also, in practice, we often choose J_ℓ so that each interval has roughly an equal number of observations. The above algorithm is implemented by the following R-codes.

1. Function to compute power of a matrix This function computes the `alpha` power of a matrix `a`, which must be a symmetric matrix.

```
matpower = function(a,alpha){
a = round((a + t(a))/2,7); tmp = eigen(a)
return(tmp$vectors%*%diag((tmp$values)^alpha)%*%t(tmp$vectors))}
```

2. Function to discretize Y This function computes $g(Y)$ from Y. The input `y` is the original sample of response; `h` is the number of slices. The code divides the sample of Y roughly evenly.

```
discretize=function(y,h){
```

```
n=length(y);m=floor(n/h)
y=y+.00001*mean(y)*rnorm(n)
yord = y[order(y)]
divpt=numeric();for(i in 1:(h-1)) divpt = c(divpt,yord[i*m+1])
y1=rep(0,n);y1[y<divpt[1]]=1;y1[y>=divpt[h-1]]=h
for(i in 2:(h-1)) y1[(y>=divpt[i-1])&(y<divpt[i])]=i
return(y1)}
```

3. Function to compute $\hat{\beta}$ This function computes the vectors $\hat{\beta}_1, \ldots, \hat{\beta}_r$. The input `x` is a matrix of dimension $n \times p$, whose rows are X_i. `r` is the dimension of $\mathscr{S}_{\text{SIR}}$. The choice of `r` will be discussed in a later chapter.

```
sir=function(x,y,h,r,ytype){
p=ncol(x);n=nrow(x)
signrt=matpower(var(x),-1/2)
xc=t(t(x)-apply(x,2,mean))
xst=xc%*%signrt
if(ytype=="continuous") ydis=discretize(y,h)
if(ytype=="categorical") ydis=y
yless=ydis;ylabel=numeric()
for(i in 1:n) {if(var(yless)!=0) {ylabel=
               c(ylabel,yless[1]);yless=yless[yless!=yless[1]]}}
ylabel=c(ylabel,yless[1])
prob=numeric();exy=numeric()
for(i in 1:h) prob=c(prob,length(ydis[ydis==ylabel[i]])/n)
for(i in 1:h) exy=rbind(exy,apply(xst[ydis==ylabel[i],],2,mean))
sirmat=t(exy)%*%diag(prob)%*%exy
return(signrt%*%eigen(sirmat)$vectors[,1:r])}
```

3.4 Application: The Big Mac Index

In this section we illustrate SIR using a data set involving 10 economic variables from 45 countries. The data set is taken from the *Arc Software* of the University of Minnesota, which can be found at the website

http://www.stat.umn.edu/arc/software.html

The detailed description of the data set can be found at the above website. The 10 variables are

```
X1.  Min labor to buy 1 kg bread
X2.  Lowest cost of 10k public transit
X3.  Electrical engineers' annual salary
X4.  Tax rate paid by engineer
X5.  Annual cost of 19 services
X6.  Primary teacher salary
X7.  Tax rate paid by primary teacher
X8.  Average days vacation per year
X9.  Average hours worked per year
Y.   Min labor to buy a Big Mac and fries
```

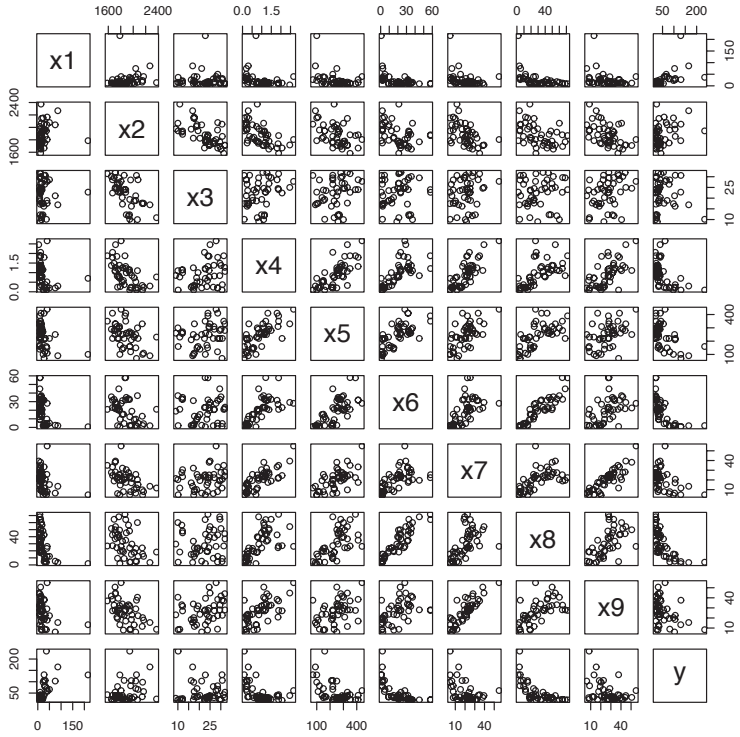

Figure 3.3 *Scatter plot matrix of ten economic variables in the Big Mac data.*

The first 9 variables are taken as predictors and the last one is taken as the response. The Big Mac index is sometimes used by economists as a informal measure of a country's purchasing-power parity (PPP). The goal of our study is to find the linear combinations of the above variables that best predict the Big Mac index. To explore the basic shape of the multivariate data, we present in Figure 3.3 the scatter plot matrix. We see that the random vector $(X_2, \ldots, X_9)$ roughly follows an elliptical distribution, but the joint distributions of the first variable with the other variables are skewed. In practice, we often make a transformation before dimension reduction to make the predictor distribution roughly elliptical. This will be discussed in a later chapter. For now, we carry out the dimension reduction without transformation. Also, we can see that Y clearly depends on X, especially X_6 and X_8.

We apply SIR to this data set, using 8 slices of roughly equal sizes. The scatter plot of Y versus the first and the second SIR predictor are presented in Figure 3.4, which demonstrates a strong nonlinear relation between Y and $\hat{\beta}_1^{\top} X$, and hardly any relation between Y and X_2. This indicates the dimension of the central subspace is 1.

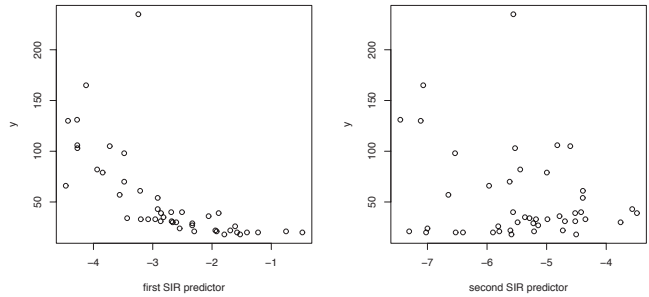

Figure 3.4 *Scatter plot of the response versus the first two SIR predictors.*

Also observe that there appears to be a stronger relation between Y and $\hat{\beta}_1^\top X$ than between Y and other individual predictors. To confirm this, in Table 3.1 we present the Spearman's correlation between Y versus $X_1, \ldots, X_9$ and $\hat{\beta}^\top X$. We use Spearman's correlation rather than Pearson's correlation because the dependence of Y on these variables is clearly nonlinear.

Table 3.1 *Spearman's correlation with the response*

X_1	0.607	X_7	−0.426
X_2	0.238	X_8	−0.797
X_3	0.159	X_9	−0.348
X_4	−0.607	$\hat{\beta}_1^\top X$	−0.894
X_5	−0.457	$\hat{\beta}_2^\top X$	0.036
X_6	−0.827		

For further development of Sliced Inverse Regression, see, for example, Fang and Zhu (1996) and Chen and Li (1998). In particular, Fang and Zhu (1996) extends SIR by replacing slice averages with kernel regression estimate (of X versus Y), and studied its asymptotic behavior when the kernel bandwidth decreases to 0 as $n \to \infty$. Chen and Li (1998) recast SIR as a minimization problem, which is influential on the development of related methods, such as the parametric inverse regression Bura and Cook (2001) and the canonical correlation method Fung et al. (2002), which will be developed in the next chapter.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com