Chapter 4

Parametric and Kernel Inverse Regression

4.1 Parametric Inverse Regression

As we have seen in the last section, what makes SIR unbiased is the fact that the set of sliced means $\{E(X|Y \in J_\ell) : \ell = 1, \ldots h\}$ are vectors in $\Sigma \mathscr{S}_{Y|X}$. It is then natural to speculate that perhaps performing parametric regression of X versus a set of functions would also produce vectors in $\Sigma \mathscr{S}_{Y|X}$ because, after all, $E(X|Y \in J_\ell)$ is nothing but the projection of X on to the indicator functions $I(Y \in J_h)$. This is the intuition behind the Parametric Inverse Regression (PIR) introduced by Bura and Cook (2001) and the canonical correlation (CANCOR) estimator proposed by Fung et al. (2002). PIR and CANCOR are closely related. However, the former is proposed as an estimator as the dimension d of the central subspace; whereas the latter is an estimator of the central subspace itself. Moreover, PIR uses a different re-scaling matrix than CANCOR. In this section we describe the procedure of regression X on Y parametrically to estimate the central subspace, and refer to it as PIR because the procedure is in the same spirit as Sliced Inverse Regression.

Let $f_1, \ldots, f_m$ be a set of functions of y. For example, these can be

$$\{1, y, y^2, \ldots, y^m\}.$$

We perform regression of X on $\{f_1(y), \ldots, f_m(y)\}$ which, at the population level, means we minimize the objective function

$$E\|X - \alpha_0 - \alpha_1 f_1(Y) - \cdots - \alpha_m f_m(Y)\|^2 \tag{4.1}$$

among all $\alpha_0, \ldots, \alpha_m \in \mathbb{R}^p$. Let

$$F(Y) = (f_1(Y), \ldots, f_m(Y))^\top, \quad B = (\alpha_1, \ldots, \alpha_m).$$

Then the objective function (4.1) can be rewritten as $E\|X - \alpha_0 - BF(Y)\|^2$. By Proposition 1.1 of Chapter 1 we know the solution to this optimization problem is

$$B = \text{cov}[X, F(Y)][\text{var}(F(Y))]^{-1}, \quad \alpha_0 = EX - BE[F(Y)]. \tag{4.2}$$

Before proceeding further, we state without proof a simple generalization of Proposition 2.1.

Lemma 4.1 *Suppose U, V, and W are random variables, random vectors, or random matrices, and the dimensions of U and V are such that the product UV is defined, and suppose the conditional and unconditional expectations involved are defined. Then*

$$E[UE(V|W)] = E[E(U|W)V] = E[E(U|W)E(V|W)].$$

Just as expected, the next theorem shows that the parametric regression of X on $f_0, \ldots, f_m$ does produce vectors in $\Sigma_{XX}\mathcal{S}_{Y|X}$. Henceforth, because several covariance matrices are involved, we will denote Σ by Σ_{XX} to distinguish it from the covariance matrix between, say, X and $F(Y)$.n

Theorem 4.1 *Suppose X and $F(Y)$ are square integrable and Assumption 3.1 of Chapter 3 holds. Then*

$$\text{span}\{\Sigma_{XX}^{-1}\text{cov}(X, F(Y))\} \subseteq \mathcal{S}_{Y|X}$$

PROOF. Without loss of generality, assume $E(X) = 0$ and $E[F(Y)] = 0$ (otherwise we can reset X to be $X - E(X)$ and $F(Y)$ to be $F(Y) - E[F(Y)]$ in the following proof). Then $\text{cov}[X, F(Y)] = E[XF^\top(Y)]$. By Lemma 4.1,

$$E[XF^\top(Y)] = E\{XE[F^\top(Y)|Y]\} = E[E(X|Y)F^\top(Y)].$$

Because $Y \perp\!\!\!\perp X|\beta^\top X$, we have

$$E[E(X|Y)F^\top(Y)] = E\{E[E(X|\beta^\top X, Y)|Y]F^\top(Y)\} = E\{E[E(X|\beta^\top X)|Y]F^\top(Y)\}.$$

By Assumption 3.1 and Lemma 1.1,

$$E\{E[E(X|\beta^\top X)|Y]F^\top(Y)\} = P_\beta^\top(\Sigma_{XX})E[E(X|Y)F^\top(Y)].$$

By Lemma 4.1 again

$$P_\beta^\top(\Sigma_{XX})E[E(X|Y)F^\top(Y)] = P_\beta^\top(\Sigma_{XX})E\{XE[F^\top(Y)|Y]\} = P_\beta^\top(\Sigma_{XX})E[XF^\top(Y)],$$

which implies $\text{span}\{\text{cov}(X, F(Y))\} \subseteq \Sigma\mathcal{S}_{Y|X}$. □

Let Σ_{XF} and Σ_{FF} denote $\text{cov}[X, F(Y)]$ and $\text{var}[F(Y)]$, respectively. The above theorem implies that, for any matrix with full row-rank, $\text{span}(\Sigma_{XX}^{-1}\Sigma_{XF}A) \subseteq \mathcal{S}_{Y|X}$. In

particular, if we take $A = \Sigma_{FF}^{-1}$, then $\text{span}(\Sigma_{XX}^{-1}B) \subseteq \mathscr{S}_{Y|X}$, where B is the regression coefficient matrix B in (4.2). Similarly, if we take $A = \Sigma_{FF}^{-1}\Sigma_{FX}\Sigma_{XX}^{-1}$, then we have

$$\text{span}(\Sigma_{XX}^{-1}\Sigma_{XF}\Sigma_{FF}^{-1}\Sigma_{FX}\Sigma_{XX}^{-1}) \subseteq \mathscr{S}_{Y|X}.$$

In other words, we need to solve the generalized eigenvalue problem

$$\Sigma_{XF}\Sigma_{FF}^{-1}\Sigma_{FX}v = \lambda\Sigma_{XX}v.$$

Or, using the notation of Section 1.3, we solve the problem $\text{GEV}(\Sigma_{XF}\Sigma_{FF}^{-1}\Sigma_{FX}, \Sigma_{XX})$.

4.2 Algorithm, R Codes, and Application

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an i.i.d. sample of (X, Y). As in the case of SIR, to develop the estimation procedure all we need to do is to replace the various covariance matrices by their sample estimate. We summarize the algorithm as follows.

Algorithm 4.1 Parametric Inverse Regression

1. Compute $\hat{\Sigma} = \text{var}_n(X)$, $\hat{\mu} = E_n(X)$. Standardize $X_1, \ldots, X_n$ as

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

2. Select functions $f_1, \ldots, f_m$. For example

$$\tilde{f}_i(Y) = \frac{f_i(Y) - E_n f_i(Y)}{\sqrt{\text{var}_n[f_i(Y)]}}, \quad i = 1, \ldots, m.$$

Form the random vector

$$F(Y) = (\tilde{f}_1(Y), \ldots, \tilde{f}_m(Y))^{\mathsf{T}}.$$

3. Compute $\hat{\Sigma}_{ZF} = \text{cov}_n[Z, F(Y)]$ and $\hat{\Sigma}_{FF} = \text{var}_n[F(Y)]$. Compute $\hat{v}_1, \ldots, \hat{v}_r$, the first r eigenvectors of $\hat{\Sigma}_{ZF}\hat{\Sigma}_{FF}^{-1}\hat{\Sigma}_{FZ}$.
4. Use

$$\hat{\beta}_1 = \hat{\Sigma}_{XX}^{-1/2}\hat{v}_1, \ldots, \hat{\beta}_r = \hat{\Sigma}_{XX}^{-1/2}\hat{v}_r.$$

as the estimates of a set of vectors in the central subspace.

Below is an R-code for calculating $\hat{\beta}_1, \ldots, \beta_r$ for a given r, using the polynomial basis functions $f_i(Y) = Y^i$, $i = 1, , \ldots, 3$.

```
pir=function(x,y,m,r){
xc=t(t(x)-apply(x,2,mean))
signrt=matpower(var(x),-1/2)
xstand=xc%*%signrt
f=numeric();ystand=(y-mean(y))/sd(y)
```

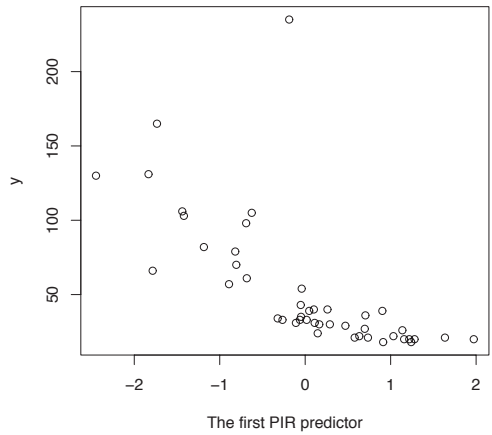

Figure 4.1 *Big Mac index versus the first PIR predictor.*

```
for(i in 1:m) f=cbind(f, ystand^i)
sigxf=cov(xstand,f);sigff=var(f)
cand=sigxf%*%solve(sigff)%*%t(sigxf)
return(signrt%*%eigen(symmetry(cand))$vectors[,1:r])}
```

We now apply PIR to the Big Mac data set, using the polynomial basis $f_i(Y) = Y^i$, $i = 1,2,3$. The Spearman's correlation between Y and the first PIR predictor is -0.886. Figure 4.1 is the scatter plot of Y versus the first PIR direction.

4.3 Relation of PIR with SIR

In this section we show that SIR is in fact a special case of PIR. To see this, consider the following basis functions

$$f_i(Y) = I(Y \in J_i), \quad i = 1,\ldots,h. \tag{4.3}$$

Before proving this result, we first prove a lemma concerning the Moore-Penrose inverse of a matrix, which is denoted by $A^\dagger$.

Lemma 4.2 *Suppose* $\pi = (\pi_1,\ldots,\pi_h)^\mathsf{T}$, *where* $\pi_i \geq 0$ *for* $i = 1,\ldots,h$ *and* $\sum_{i=1}^h = 1$. *Then*

$$[\mathrm{diag}(\pi) - \pi\pi^\mathsf{T}]^\dagger = \mathrm{diag}(\pi)^{-1} - 1_p 1_p^\mathsf{T}.$$

PROOF. Because

$$\mathrm{diag}(\pi) - \pi\pi^\mathsf{T} = \mathrm{diag}(\pi)^{1/2}[I_p - \mathrm{diag}(\pi)^{-1/2}\pi\pi^\mathsf{T}\mathrm{diag}(\pi)^{-1/2}]\mathrm{diag}(\pi)^{1/2},$$

we have

$$(\operatorname{diag}(\pi)-\pi\pi^{\mathsf{T}})^{\dagger}=\operatorname{diag}(\pi)^{-1/2}(I_p-vv^{\mathsf{T}})^{\dagger}\operatorname{diag}(\pi)^{-1/2},$$

where $v=\operatorname{diag}(\pi)^{-1/2}\pi$. Since

$$v^{\mathsf{T}}v=\pi^{\mathsf{T}}\operatorname{diag}(\pi)^{-1}\pi=\pi^{\mathsf{T}}1_p=1,$$

the matrix $(I_p-vv^{\mathsf{T}})^{-1}$ is the projection on to $\operatorname{span}(v)^{\perp}$. It follows that $(I_p-vv^{\mathsf{T}})^{\dagger}=I_p-vv^{\mathsf{T}}$. Hence

$$\begin{aligned}(\operatorname{diag}(\pi)-\pi\pi^{\mathsf{T}})^{\dagger}&=\operatorname{diag}(\pi)^{-1/2}(I_p-vv^{\mathsf{T}})\operatorname{diag}(\pi)^{-1/2}\\&=\operatorname{diag}(\pi)^{-1}-\operatorname{diag}(\pi)^{-1}\pi\pi^{\mathsf{T}}\operatorname{diag}(\pi)^{-1}\\&=\operatorname{diag}(\pi)^{-1}-1_p1_p^{\mathsf{T}},\end{aligned}$$

as desired. □

We now prove the main theorem of this section. In the following, for a matrix A, $A_{\cdot j}$ stands for its jth column.

Theorem 4.2 *Suppose f_i, $i=1,\dots,m$ are chosen as the basis (4.3), then PIR reduces to solving the generalized eigenvalue problem*

$$\operatorname{GEV}(\operatorname{var}[E(X|g(Y)],\operatorname{var}(X)),$$

where $g(Y)=\sum_{i=1}^{h} iI(Y\in J_i)$ is the discretized version of Y defined in Section 3.3.

PROOF. Let $\tilde{f}_i(Y)=I(Y\in J_i)-P(Y\in J_i)$. Then

$$\begin{aligned}(\Sigma_{XF})_{\cdot j}&=E[X\tilde{f}_i(Y)]\\&=E\{X[I(Y\in J_j)-P(Y\in J_j)]\}\\&=P(Y\in J_i)[E(X|Y\in J_j)-E(X)].\end{aligned}$$

Let π denote the vector $\{P(Y\in J_i)\}_{i=1}^{p}$, and let G denote the matrix

$$(E(X|Y\in J_1),\dots,E(X|Y\in J_h)).$$

Then Σ_{XF} can be written as $G\operatorname{diag}(\pi)$. In the meantime,

$$(\Sigma_{FF})_{ij}=\operatorname{cov}(I(Y\in J_i),I(Y\in J_j))=P(Y\in J_i\cap J_j)-P(Y\in J_i)P(Y\in J_j).$$

Since $J_i\cap J_j=\varnothing$, the first term on the right-hand side is $\delta_{ij}P(Y\in J_i)$ where δ_{ij} is the Kronecker δ function. Consequently,

$$\Sigma_{FF}=\operatorname{diag}(\pi)-\pi\pi^{\mathsf{T}}.$$

By the discussion at the end of Section 4.1, PIR is determined by the eigenvalue problem

$$\operatorname{GEV}(\Sigma_{XF}\Sigma_{FF}^{\dagger}\Sigma_{FX},\Sigma_{XX}).$$

However, by Lemma 4.2,

$$\begin{aligned}\operatorname{diag}(\pi)\Sigma_{FF}^{\dagger}\operatorname{diag}(\pi) &= \operatorname{diag}(\pi)[\operatorname{diag}(\pi)^{-1} - 1_p 1_p^{\mathsf{T}}]\operatorname{diag}(\pi)\\ &= \operatorname{diag}(\pi) - \pi\pi^{\mathsf{T}}.\end{aligned}$$

So

$$G\operatorname{diag}(\pi)\Sigma_{FF}^{\dagger}\operatorname{diag}(\pi)\,G^{\mathsf{T}} = G\operatorname{diag}(\pi)G^{\mathsf{T}} - G\pi\pi^{\mathsf{T}}G^{\mathsf{T}}. \tag{4.4}$$

Note that

$$\begin{aligned}G\pi &= \textstyle\sum_{i=1}^{h} E(X|Y\in J_i)P(Y\in J_i) = \sum_{i=1}^{h} E(XI(Y\in J_h)) = E(X)\\ G\operatorname{diag}(\pi)G^{\mathsf{T}} &= \textstyle\sum_{i=1}^{h} P(Y\in J_i)E(X|Y\in J_i)E(X^{\mathsf{T}}|Y\in J_i)\\ &= E[E(X|g(X))E(X^{\mathsf{T}}|g(X))],\end{aligned}$$

where $g(Y) = \sum_{i=1} iI(Y\in J_i)$ is the discretized version of Y as defined in Section 3.3. Hence (4.4) is in fact the matrix $\operatorname{var}[E(X|g(Y)]$. □

4.4 Relation of PIR with Ordinary Least Squares

Ordinary Least Squares (OLS) is another important method for Sufficient Dimension Reduction. Its unbiasedness as a dimension reduction estimator was first established by Li and Duan (1989). We have, in effect, demonstrated this result in Section 1.10 to illustrate how we can borrow the symmetry in the distribution of X to recover the central subspace without fitting an accurate regression model. We now show that this can also be derived from the PIR.

Let the set of basis functions for PIR be the singleton $f_1(y) = y$. Then

$$\Sigma_{XF} = \operatorname{cov}(X,Y), \quad \Sigma_{FF} = \operatorname{var}(Y).$$

Thus $\Sigma_{XF}\Sigma_{FF}^{-1}\Sigma_{FX}$ is rank-1 matrix, and the only eigenvector of the problem $\text{GEV}(\Sigma_{XF}\Sigma_{FF}^{-1}\Sigma_{FX}, \Sigma_{XX})$ corresponding to its nonzero eigenvalue is proportional to the vector $\Sigma_{XX}^{-1}\Sigma_{XY}$, which is precisely the population-level expression of the Ordinary Least Squares estimate. At the sample level, we use the OLS estimate

$$\hat{\beta} = \hat{\Sigma}_{XX}^{-1}\hat{\Sigma}_{XY}$$

to estimate the central subspace $\mathscr{S}_{Y|X}$.

Figure 4.2 shows the result of OLS applied the Big Mac data set. The Spearman's correlation between Y and the first PIR predictor is -0.817.

4.5 Kernel Inverse Regression

Another extension of SIR is to replace slice averages of X by kernel smoothing (Fang and Zhu (1996)). Let $\kappa : \Omega_Y \to \mathbb{R}$ be a kernel function. For example, consider the Gaussian kernel function

$$\kappa_b(u) = b^{-1}\exp(-u^2/b).$$

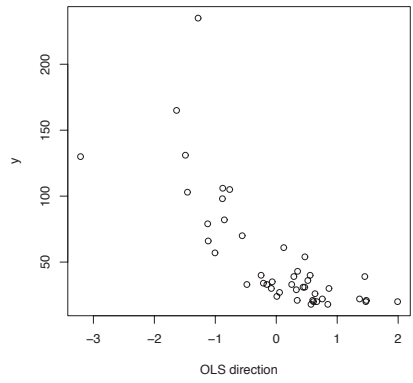

Figure 4.2 *Big Mac index versus the OLS predictor.*

where $b > 0$ is the bandwidth. Instead of approximating $E(X|Y \in J_i)$, we estimate $E(X|Y = y)$ by

$$\hat{E}(X|Y=y) = \frac{E_n[X\kappa_b(Y-y)]}{E_n[\kappa_b(Y-y)]}.$$

As a form of regularization, Fang and Zhu (1996) proposed to replace the denominator by $\max\{E_n[\kappa_b(Y-y)], \varepsilon\}$ where $\varepsilon > 0$ is a tuning constant. That is, we use

$$\tilde{E}(X|Y=y) = \frac{E_n[X\kappa_b(Y-y)]}{\max\{E_n[\kappa_b(Y-y)], \varepsilon\}}$$

to estimate $E(X|Y=y)$. In order to give ε an appropriate scale, we take it to be $\delta E_n\{E_n[\kappa_b(Y-\tilde{Y})]\}$, where

$$E_n\{E_n[\kappa_b(Y-\tilde{Y})]\} = n^{-2}\sum_{i=1}^n\sum_{j=1}^n \kappa_b(Y_i - Y_j). \tag{4.5}$$

We then approximate $\mathrm{var}[E(X|Y)]$ by

$$\mathrm{var}_n[\tilde{E}(X|Y)] = n^{-1}\sum_{i=1}^n [\tilde{E}(X|Y_i) - E_n(X)][\tilde{E}(X|Y_i) - E_n(X)]^{\mathsf{T}}.$$

The central subspace $\mathscr{S}_{Y|X}$ is then estimated by solving the generalized eigenvalue problem

$$\mathrm{GEV}(\mathrm{var}_n[\tilde{E}(X|Y)], \mathrm{var}_n(X)).$$

This method is called the kernel inverse regression (KIR). The kernel estimate can be calculated efficiently by first computing the Gram matrix, as described in the following algorithm.

Algorithm 4.2 Kernel Inverse Regression

1. Standardize X_i and Y_i to Z_i as before and standardize Y_i to be

$$\tilde{Y}_i = \frac{Y_i - E_n(Y)}{\sqrt{\text{var}_n(Y)}}.$$

2. Compute the Gram matrix $K = \{\kappa_b(Y_i - Y_j)\}_{i,j=1}^n$.
3. Compute the vector $V = \{E_n[\kappa_b(Y_i - Y_j)|Y_i]\}_{i=1}^n$ by $K1_n/n$.
4. Compute the number in (4.5) by $C = n^{-2}1_n^\top K1_n$, and compute the vector $W = \{\max(V_i, \delta C)\}_{i=1}^n$.
5. Compute $\text{var}_n[\tilde{E}(X|\tilde{Y})]$ as

$$K\,\text{diag}(W)^{-2}\,K.$$

6. Compute the first r eigenvectors $v_1, \ldots, v_r$ of the matrix in Step 5.
7. Compute $\hat{\beta}_i = \hat{\Sigma}^{-1/2}v_i$, $i = 1, \ldots, r$.

An R code to implement the above algorithm is given below.

```
kir=function(x,y,b,eps,r){
gker=function(b,y){
n=length(y);k1=y%*%t(y);k2=matrix(diag(k1),n,n)
return((1/b)*exp(-(k2+t(k2)-2*k1)/(2*b^2)))}
xc=t(t(x)-apply(x,2,mean))
signrt=matpower(var(x),-1/2)
xst=xc%*%signrt
f=numeric();yst=(y-mean(y))/sd(y)
kern=gker(b,yst)
mea=mean(c(kern%*%rep(1,n)))
den=apply(cbind(kern%*%rep(1,n),rep(eps*mea,n)),1,max)
scale=eigen(kern)$values[1]
exy=(kern%*%xst)*(1/den);mat=t(exy)%*%exy
return(signrt%*%eigen(mat)$vectors[,1:r])}
```

This algorithm requires the tuning constants b and δ. This can be done by Cross-Validation or Generalized Cross-Validation. But for inverse regression we propose a simpler method. Since inverse regression does not fit a forward model of Y versus X, it does not suffer from the issue of over fitting. Thus it is entirely reasonable to use the strength of the dependence between Y and $\hat{\beta}^\top X$ to evaluate the tuning parameters. Since this relation is generally nonlinear, we can use the Spearman's correlation to measure their dependence.

More specifically, recall that the Spearman's correlation between two samples $U_1, \ldots, U_n$ and $V_1, \ldots, V_n$ is defined as the sample correlation of the rank of the two samples. That is, if $R(U_i)$ and $R(V_i)$ are be the ranks of U_i and V_i, respectively, then

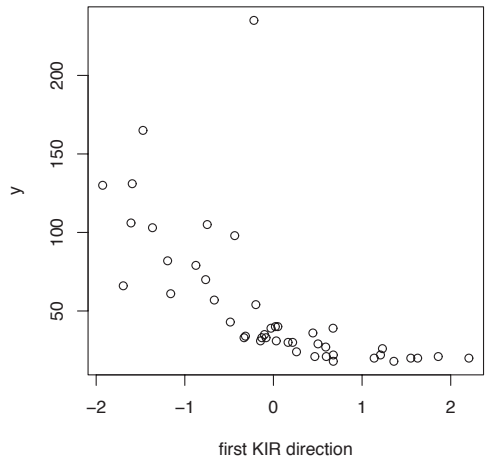

Figure 4.3 *Big Mac index versus the first KIR predictor.*

the Spearman's correlation between the samples is

$$\text{scor}_n(U,V) = \frac{\text{cov}_n[R(U),R(V)]}{\sqrt{\text{var}_n[R(U)]\text{var}_n[R(V)]}}.$$

Let $\hat{\beta}_1(b,\delta)$ be the first KIR direction for a fixed b and δ, and let $\hat{Y}(b,\delta) = [\hat{\beta}(b,\delta)]^{\mathsf{T}}X$ be the prediction of Y, we maximize

$$\text{scor}_n[\hat{Y}(b,\delta),Y]$$

over a grid of (b,δ), say

$$b = 0.1, 0.2, \ldots, 1, \quad \delta = 0.1, 0.2, \ldots, 2.$$

The maximum Spearman's correlation is achieved at $b = 0.2$, $\delta = 1.1$, and the Spearman's correlation is 0.897. Figure 4.3 shows the scatter plot of Y versus the first KIR predictor.

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