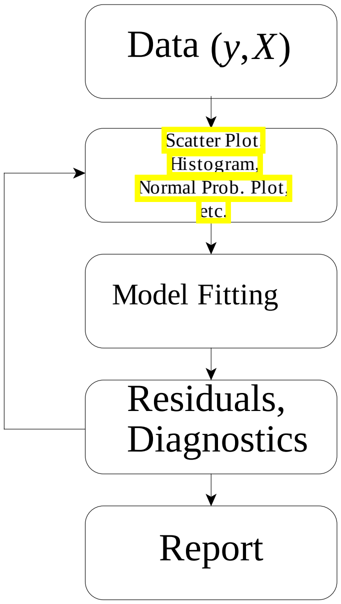

Figure 1.4: A Regression Diagram for Low Dimensional Problems.

analysis.

As we have seen, the role of graphics is quite critical in the routine practice of regression analysis. But it only works when the number of regressors is small, say one, two, or three. With two regressors, we can use 3-D techniques to display the data pattern between Y and $\mathbf{x}$. With three regressors, this is getting harder, although it still can be done with the help of color as the fourth dimension for example. But what to do with three or more regressors? We can certainly plot two of them again Y each time for example. But this task could become extremely laborious. For $p = 10$, we already have 45 such combinations. It is not clear how to effectively put together all information from different plots. Yet the coordinate variables are not the only choices for inspecting relationship among variables. Sometimes linear combinations can be more informative. But this only accelerates the overloading of plot inspection. In summary, the effectiveness of such preliminary graphical inspection is severely impaired as the dimension gets higher.

For larger dimensional problems, empirical model building often begins with multiple linear regression (for continuous Y). It is of course hard to believe that this overly-simplied model can prevail under most situations. But there is no obvious better alternative. What we can count on is no more than the rationale that any nonlinear function may be approximately linear within a suitable domain. How true this assumption is depends on each application and it is extremely hard to tell in advance; see Chapter 10 however.

It appears that in order to maintain the spirit underlying the regression paradigm as illustrated in Figure 1.4, we have to reduce the regressor dimensionality first.

1.3 Principal component analysis.

Each time when the issue of dimension reduction is mentioned, one would normally associate it with Principal component analysis (PCA). No doubt that PCA is perhaps the most popular procedure of dimension reduction. But what does PCA do in regression? How helpful is it? Before discussing such issues, let's take a brief look at the procedure itself first.

PCA projects the high dimensional data to a lower dimensional space with the hope that the essential structure in the original data can be kept as much as possible. The projected space is chosen so that the points can spread out as much as possible.

Let $\mathbf{x}$ be the p-dimensional variable of interest. The first principal component is a linear combination $\mathbf{b}'\mathbf{x}$ of the coordinate variables of $\mathbf{x}$ that has the largest variance among all $\mathbf{b}$ with unitary length:

$$\max_{||\mathbf{b}||=1} \mathbf{b}'\Sigma_{\mathbf{x}}\mathbf{b}$$

Here $\Sigma_{\mathbf{x}}$ denotes the covariance matrix of $\mathbf{x}$.

After finding the first direction, say $\mathbf{b}_1$, we repeat the same procedure by restricting to those "uncorrelated" directions $\mathbf{b}$, namely those that yield projections uncorrelated with $\mathbf{b}_1'\mathbf{x}$: $0 = cov(\mathbf{b}'\mathbf{x}, \mathbf{b}_1'\mathbf{x}) = \mathbf{b}'\Sigma_{\mathbf{x}}\mathbf{b}_1$. This gives the second principal direction $\mathbf{b}_2$. Continue this procecess to get all other directions, $b_3, \cdots, b_p$. Denote the the variance of $\mathbf{b}_i'\mathbf{x}$ by λ_i. It can be shown that

$$\Sigma_{\mathbf{x}}\mathbf{b}_i = \lambda_i \mathbf{b}_i$$

Thus to find the principal directions we need only to conduct the eigenvalue decomposition on the covariance matrix of $\mathbf{x}$. The sample version of PCA is carried out by replacing $\Sigma_{\mathbf{x}}$ with the sample covariance matrix $\hat{\Sigma}_{\mathbf{x}}$.

Eigenvalues of PCA often decrease rapidly. When this happens, it shows that most of the data spread out very well along the first few directions. Thus it seems hopeful that the most interesting structure in the data may show up along these directions. But this is not a guarantee. In spite of the richness in the literature of PCA, not much theoretical work has be done regarding how successful PCA is in finding nonlinear structure. We shall come back to this issue in Chapter 8 later.

In application, one often needs to rescale each coordinate variable appropriately before applying PCA. One frequently-used rescaling factor is the standard deviation. This amounts to use correlation matrix instead of the covarinace matrix for eigenvalue decomposition.

1.4 Effective dimension reduction in regression.

To reduce dimensionality in regression problems, one possibility is to apply PCA on $\mathbf{x}$ first, keeping the first few principal components for modeling the relationship with Y. This is called the principal component regression. The result is sometimes helpful, sometimes not. One simple explanation is that this way of reducing the regressor dimensionality is totally independent of the output variable Y. Thus any two different data sets would always reduce to the same linear combinations, as long as the input variables **x** have the same distributions. This is so, even if the relationship between **x** and Y is not the same for the two data sets.

To address the dimension reduction issue in regression, one must not treat **x** separately from Y. This is what we shall develop in this section. At the center of the scene will be the notion of effective dimension reduction *(e.d.r.)*. This notion conveys the desirable situation in which one can reduce the dimension of **x** without losing any information which is essential in predicting Y.

1.4.1 The model.

Li(1991) introduced the following model

$$Y = g(\beta_1'\mathbf{x}, \beta_2'\mathbf{x}, ..., \beta_K'\mathbf{x}, \epsilon). \tag{1.1}$$

Here we consider Y as a univariate output variable. The case of multivariate output will be treated in Chapter 17. The dimension of **x** is denoted by p. The random error ϵ is independent of **x**, but its probability distribution is unknown. Our primary interest is on the K p-dimensional vectors $\beta_1, \cdots, \beta_K$.

It is easier to see how this model is related to dimension reduction by comparing Figure 1.5 to Figure 1.6. Figure 1.5 shows the most general situation in regressing Y on **x** :

$$Y = f(x_1, \cdots, x_p, \epsilon)$$

On the top of the chart, there are p input nodes together with a node for random error ϵ. The unknown function f is represented by the black box in the middle, leading to the output node Y at the bottom. Compared to this most general situation, Figure 1.6 adds an intermediate layer of nodes. These intermediate nodes combine data from the input nodes linearly using weights indicated along the line segments. From this chart, it is clear that the relationship between **x** and Y is determined only through $\beta_1'\mathbf{x}, \cdots, \beta_K'\mathbf{x}$; $K = 2$ is shown there. The black box represents the unknown function g in (1.1).

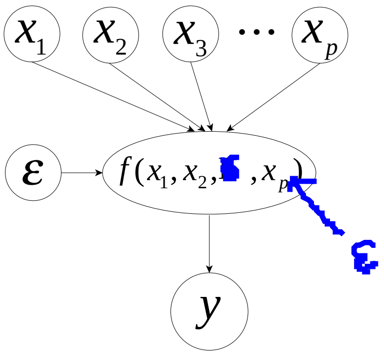

Figure 1.5: General Regression Model.

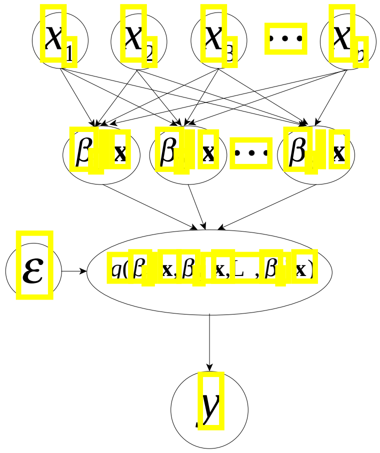

Figure 1.6: Model 1.1.

If g is known, then (1.1) is not much different from a simple neural net model, or a nonlinear regression model. But what makes (1.1) special is that g is unknown and can be completely general. This leads immediately to the question of how to estimate the beta vectors. As it turns out, there are several ways to proceed and this is what we shall study in this book, starting from Chapter 2. For now, let's first take a look at many commonly used models in regression. They all belong to (1.1), each with a different specification about g.

1.4.2 Special cases.

1. Multiple Linear regression.

$$y = \alpha + \beta'\mathbf{x} + \epsilon \tag{1.2}$$

2. Box-Cox transformation..

$$h(y) = \alpha + \beta'\mathbf{x} + \epsilon \tag{1.3}$$

where $h(y)$ is a power transformation :

$$h(y) = (y^\lambda - 1)/\lambda$$

3. Transformation-inside.

$$y = g(\beta'\mathbf{x}) + \epsilon \tag{1.4}$$

4. Transformation-both-sides

$$h(y) = g(\beta'\mathbf{x}) + \epsilon \tag{1.5}$$

5. Heterosedasticity (Taguchi).

$$y = \alpha + \beta_1'\mathbf{x} + \epsilon g(\beta_2'\mathbf{x}) \tag{1.6}$$

There are many more special cases of (1.1). Some of them will be encountered in later chapters. We turn to the discussion on the β vectors.

1.4.3 The e.d.r. directions.

The notion of e.d.r. direction plays a critical role in the methodological development of SIR/PHD.

Definition 1.1 Under (1.1), the space $\mathcal{B}$ generated by $\beta_1, \cdots, \beta_K$ is called the e.d.r. space. Any non-zero vector in the e.d.r. space is called an e.d.r. direction.

Observe that by changing g suitably, (1.1) can be reparametrized by any set of K linearly independent e.d.r. directions. Thus it is the e.d.r. space $\mathcal{B}$ that can be identified; the individual vectors β_1, ..., β_K themselves are not identifiable (unless further structural conditions on g are imposed). Finding the e.d.r. space or a subspace of it will be our primary goal.

This problem is different from the estimation of regression coefficients. The difference can be manifested by reconsidering the multiple linear Model under the following way of reparametrization :

$$y = \alpha + b(\tilde{\beta}'\mathbf{x}) + \epsilon$$

where we restrict that β has the unit length and b is nonnegative.

Note that when b equals 0, $\tilde{\beta}$ is not well-defined. The roles of $\tilde{\beta}$ and b have been mixed up in (1.2). The vector $\tilde{\beta}$ is used to identify the relative contribution or importance from each factor. As we shall see later, the estimation of $\tilde{\beta}$ is less sensitive to the link violation. Estimation of the e.d.r. space can be viewed as equivalent to the estimation of $\tilde{\beta}$ up to a sign. From the visualization point of view, the sactterplot for y against $\tilde{\beta}$ is as informative as that for y against β. Once we identify an e.d.r. direction, we can standardize it to have a unit length. To further decide the sign, we can simple choose the one that yields a positive correlation with Y. If we happen to choose the wrong one, we can usually find that out after a glance at the scatterplot : the regression line is going down. The scalar factor b determines the size of R-squared.

1.4.4 The rationale.

All models are imperfect in some sense. Like others, (1.1) should be interpreted as an approximation to reality. However, the fundamental difference between this and other statistical models is that (1.1) takes the weakest form to reflect our hope that a low dimensional projection of a high dimensional regressor variable contains most of the information that can be gathered from a sample of a modest size. (1.1) does not impose any structures on how the projected regressor variable effects the output variable. In addition, we may vary K to reflect the degree of the anticipated dimension reduction. At $K = p$, (1.1) becomes a redundant assumption. By comparison, most regression models assume $K = 1$ with additional structures on g.

A philosophical point is to be emphasized here : the estimation of the projection directions can be a more important statistical issue than the estimation of the structure of g itself. In fact, the structure of g is impossible to identify unless we have other scientific evidence beyond the data under study. One can obtain two different versions of g to represent the same joint distribution of y and $\mathbf{x}$. Thus what we can estimate at most are statistical quantities such as the conditional mean or quantiles of Y given $\mathbf{x}$. On the other hand, at the beginning stage of data analysis when one does not have a fixed objective in mind, the need for estimating such quantities is not as pressing as that for finding ways to simplify the data. Our formulation of estimating the e.d.r. directions is one way to address such a need in data analysis. After finding a good e.d.r. space, we can project data to this smaller space. Then we are in a better position to identify what should be pursued further : model building, response surface estimation, cluster analysis, heteroscedasticity analysis, variable selection, or inspecting scatterplots (or spinning plots) for interesting features. After dimension reduction, if we want to estimate the response surface for example, then we can apply nonparametric smoothing, Box-Cox transformation, or other techniques. Needless to say, the door is open for further serious work.

1.4.5 An equivalent version.

(1.1) is equivalent to :

the conditional distribution of y given $\mathbf{x}$ depends on $\mathbf{x}$
only through the K dimensional variable $(\beta_1'\mathbf{x}, ..., \beta_K'\mathbf{x})$

or, to put it slightly differently,

conditional on $\beta_1'\mathbf{x}, \cdots, \beta_k'\mathbf{x}$, y and $\mathbf{x}$ are independent.

The reduced variable, $(\beta_1\mathbf{x}, \cdots, \beta_K\mathbf{x})$, is as informative as the original $\mathbf{x}$ in predicting y.

Note that this version does not seperate the role of g from ϵ in (1.1). In fact, there are more than one way to construct g and ϵ. For instance, if we denote the c.d.f. for the conditional distribution of Y given $\beta_1'\mathbf{x} = \theta_1, \beta_K'\mathbf{x} = \theta_K$ by $F_{\theta_1,\cdots,\theta_K}, (\cdot)$, then we can take $g(\beta_1'\mathbf{x}, \cdots, \beta_K'\mathbf{x}, \epsilon)$ to be $F^{-1}(\beta_1'\mathbf{x}, \cdots, \beta_K'\mathbf{x}, \epsilon)$ and assume that ϵ follows the uniform distribution on $[0, 1]$. For the special case 1 of section 1.4.2, the multiple linear model, this leads to the following expression:

$$Y = \Phi^{-1}(a + \beta'\mathbf{x} + \epsilon^*)$$

where Φ is the c.d.f. of the original random variable ϵ. The distribution of ϵ^* is uniform on $[0, 1]$.

Strictly speaking, for any given joint distribution of Y and $\mathbf{x}$, our definition of the e.d.r. space is not mathematically rigorous. This is because if (1.1) holds for a set of vectors $\beta_1, \cdots, \beta_k$, then we can always add another vector to enlarge the e.d.r. space without violation (1.1) or the equivalent form mentioned above. Of course, the key in the notion of e.d.r. space is to find the one with the smallest dimension. Now a question arises : is this space unique ? Cook(1994) studied this question and it lead to a refinement to the defintion 1.1. It turns out that under certain regularity conditions, the e.d.r. space with the smallest dimension is unique. We shall assume this is the case from now on.

$\hat{\beta}_i = A'\hat{v}_i$. Therefore the SIR variates $\hat{v}_i'\mathbf{z}$ obtained from the standardarized regressor $\mathbf{z}$ are the same as the SIR variates obtained from the original regressors, $\hat{v}_i'\mathbf{z} = \hat{v}_i' A\mathbf{x}\hat{\beta}_i'\mathbf{x}$.

2.4 Some simulation examples.

In each of the following examples, each regressor variable, $x_1, \cdots, x_n$, and the error term ϵ are generated independently from the normal distribution with mean 0 and variance 1. The sample size n is 100 and the regressor dimension p is 5. The simulation is carried out in Xlisp-stat. A program for implementing the SIR algorithm can be obtained by *http://www.stat.ucla.edu/ kcli*. We take $H = 10$ slices in each example. A computer output is given below.

```
 ; loading "sir-model.lsp"
; finished loading "sir-model.lsp"
> ; loading "sir-program.lsp"
; finished loading "sir-program.lsp"
> (def x(g-normal-rand 5 100))
X
> (def err(normal-rand 100))
ERR
> (def y (+ 5 (nth 0 x) (nth 1 x) (nth 2 x) err ))
y
> (def out-linear (sir-model x y))
==================================================
        *** Sliced Inverse Regression Model ***
Number of slices:      10
the first direction found by SIR:
(-0.55552 -0.518951 -0.523881 0.0125557 0.00506731
the second direction found by SIR:
(0.163475 -0.393415 0.141448 -0.879026 -0.213294
the third direction found by SIR:
(0.689774 -0.205839 -0.676929 0.191294 -0.278481
the companion output eigenvalues of SIR:
(0.788627 0.126048 0.0821447 0.0683831 0.0160881
the sum of all eigenvalues:
1.08129
================================================
OUT-LINEAR
> def y-trans (** y 2))
Y-TRANS
> (sir-mode x y-trans)
================================================
        *** Sliced Inverse Regression Model ***
Number of slices:      10
the first direction found by SIR:
(-0.55552 -0.518951 -0.523881 0.0125557 0.00506731
the second direction found by SIR:
(0.163475 -0.393415 0.141448 -0.879026 -0.213294
the third direction found by SIR:
(0.689774 -0.205839 -0.676929 0.191294 -0.278481
the companion output eigenvalues of SIR:
(0.788627 0.126048 0.0821447 0.0683831 0.0160881
================================================
#<Object:  5037498, prototype = SIR-MODEL-PROTO>
> (def y (/ (nth 0 x) (+ .5 (** (+ (nth 1 x) 1.5) 2))))
Y
> (sir-model x y)
================================================
```

```
        *** Sliced Inverse Regression Model ***
Number of slices:      10
the first direction found by SIR:
(-0.916301 0.126228 -0.0229527 0.0548278 -0.0519934
the second direction found by SIR:
(0.0451041 0.875115 0.250871 -0.0710619 0.00957446
the third direction found by SIR:
(-0.0178142 0.0438353 -0.0142057 0.637159 0.790221
the companion output eigenvalues of SIR:
(0.779919 0.642207 0.126759 0.0883023 0.0101885
the sum of all eigenvalues:
1.64738
================================================
#<Object:  5314930, prototype = SIR-MODEL-PROTO>
> (spin-plot (list (nth 0 x) y (nth 1 x))))
#<Object:  5297810, prototype = SPIN-PROTO>
>
```

use

(sir-i-ii-model x y)

$$y = \frac{x_1}{(0.5 + (x_2 + 1.5)^2)}$$

$$x = (x_1, \cdots, x_5)$$

Figure 2.0: A computer output

Example 1. Linear model.

$$Y = 5 + x_1 + x_2 + x_3 + 0x_4 + 0x_5 + \epsilon$$

From the computer output, we see that the first eigenvalue (=.788) is quite large and the first SIR direction (=($-$.555, $-$.518, $-$.523, .012, .005)) is nearly proportional to the e.d.r. direction $(1, 1, 1, 0, 0)$. The scatterplot of Y against the first variate shows the linearity relationship very well.

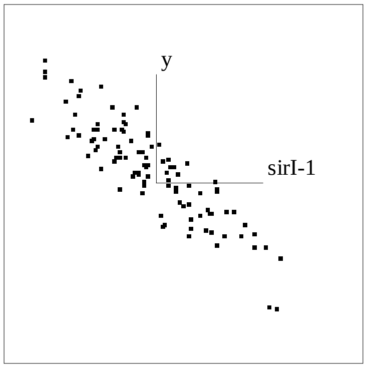

Figure 2.1: SIR View for Example 1. Linear Model.

Example 2. Transformation-outside model.

$$Y = (5 + x_1 + x_2 + x_3 + \epsilon)^2$$

The data are taken from the first example with Y being changed by the square transformation. The output of SIR remains the same as before. This is to be expected because the transformation is monotone (all Y values from Example 1 are positive) and only the ordering of Y is used in SIR. SIR estimates the e.d.r. direction quite well. The plot of Y on the first SIR variate reveals a hetroscedastic pattern - the variance of Y increases along the x-axis direction.

Example 3. Transformation-inside model.

$$Y = (5 + x_1 + x_2 + x_3)^2 + \epsilon$$

In this example, Y is constructed with the **x** values and ϵ borrowed from Example 1. The output of SIR shows that the e.d.r. direction is estimated well by the first SIR direction. The SIR plot, Figure 2.3, reveals a pattern rather different from Figure 2.2.

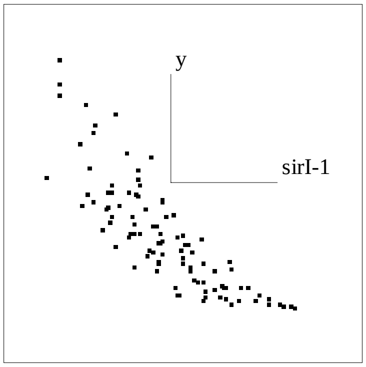

Figure 2.2: SIR View for Example 2. Transformation-Outside Model.

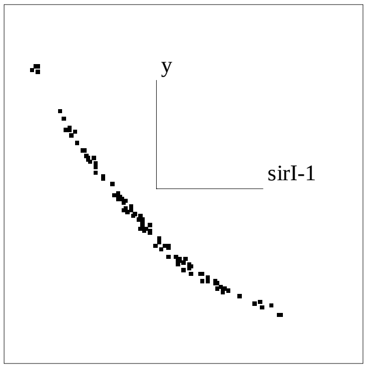

Figure 2.3: SIR view for Example 2. Transformation-Inside Model.

Example 4. Rational function.

$$y = \frac{x_1}{0.5 + (x_2 + 1.5)^2} + \sigma\epsilon$$

For now, we only consider the noise-free case $\sigma = 0$. The dimension K of the e.d.r. space is two. We still take $H = 10$. The output of SIR shows that there are two large eigenvalues. The corresponding eigenvectors, $(-.916, .126, -.022, .055, -.051)$ and $(.045, .875, .251, -.07, .009)$ match e.d.r. directions $(1, 0, 0, \cdots)$ and $(0, 1, 0, \cdots)$ closely. The 3-D plot of Y against the first two SIR variates is shown in Figure 2.4(a)-(d) from a few angles. This is nearly the same as the one given by plotting Y against the true e.d.r. variates, x_1 and x_2, Figure 2.5(a)-(d). Fig 2.5(e) is the true response surface.

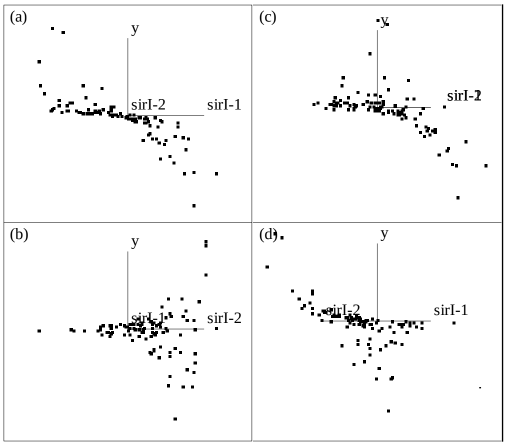

Figure 2.4: SIR View for Example 4. Rational Function.

2.5 Contour plotting and SIR.

Before presenting the theory of SIR in the next section, we can use contour plots to illustrate why SIR works well in finding e.d.r. directions. We begin by assuming $p = 2$, $K = 1$, and there is no error term in (1.1). Thus the model can be expressed as

$$Y = g(b_1x_1 + b_2x_2)$$

The e.d.r. direction is $\beta = (b_1, b_2)'$.

Contour plotting is a popular way of representing a real-valued function with two arguments. In our daily life, we have encountered many contour plots such as weather maps for temperature, atmosphere pressure, and so on. A contour of g consists of points (x_1, x_2) with the same Y value. In our case, contours are straight lines perpendicular to the e.d.r. direction $\beta = (b_1, b_2)'$. Note that this property has nothing to do with the functional form of the univariate function g. As it becomes clear soon, this is an important reason why SIR can find e.d.r. directions without knowing the functional form of g. Figure 2.6(a) illustrates a contour plot for $\beta = (1, 1)'$.

Now if **x** has a spherical distribution, normal with identity covariance for example, then it is easy to find that data points for **x** would be symmetrically scattered between contour lines, as shown in Figure 2.6(b). If we apply the SIR algorithm to data generated from this model, then the slicing step will create parallel slots like those in Figure 2.6(a), and the averaging step would give slice means which should fall near the line along the e.d.r. direction, as marked by stars in Figure 2.6 (c). Now to find this line, we can apply PCA. This is what is done at the eigenvalue decomposition step.

The story is slightly different for the case that the covariance matrix of **x** is not an identity. In general, the shape of the **x** data points should look more like an ellipsoid. Thus points on the opposite side of the e.d.r. direction within each slot is not symmetric. The slice means will not fall on the 45^o line. But we can show that they will fall along another line. Thus a

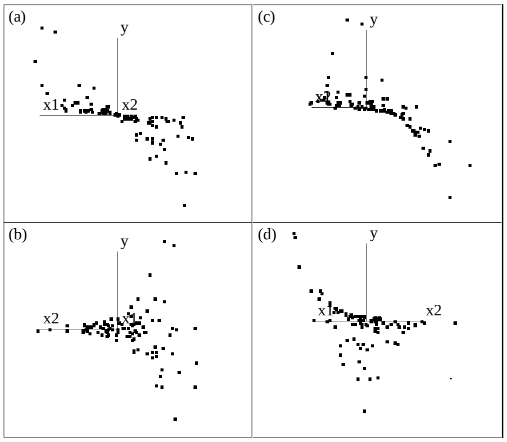

Figure 2.5: Best View for Example 4. Rational Function.

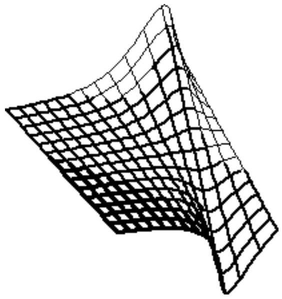

Figure 2.5(e): Response surface of Example 4, Rational function.

straightforward PCA will find the wrong direction. This is why we need the term $\hat{\Sigma}_{\mathbf{x}}$ on the right-side of (2.1) to make an appropriate adjustment.

2.6 Fisher consistency for SIR.

In this section, we shall establish the Fisher consistency property of SIR for finding e.d.r. directions. Imagine either that our sample consists of the entire population or that the sample size is infinitely large. Fisher consistency for a statistical estimation procedure describes the desirable situation that the estimate produced must coincide with what we want to estimate. Fisher consistency is just one way of saying that the procedure has no estimation bias (in theory).

Now consider the trajectory of the inverse regression $E(\mathbf{x}|Y = y)$ as y varies. In general, this draws a curve in R^p. The center of this curve is located at $E(E(\mathbf{x}|Y)) = E\mathbf{x}$. However, the following theorem shows that under suitable conditions, this curve indeed lies on a K-