High dimensional data analysis via the SIR/PHD approach

April 6, 2000

Preface

Dimensionality is an issue that can arise in every scientific field. Generally speaking, the difficulty lies on how to visualize a high dimensional function or data set. This is an area which has become increasingly more important due to the advent of computer and graphics technology. People often ask : “How do they look?”, “What structures are there?”, “What model should be used?” Aside from the differences that underly the various scientific contexts, such kind of questions do have a common root in Statistics. This should be the driving force for the study of high dimensional data analysis.

Sliced inverse regression(SIR) and principal Hessian direction(PHD) are two basic dimension reduction methods. They are useful for the extraction of geometric information underlying noisy data of several dimensions - a crucial step in empirical model building which has been overlooked in the literature. In this Lecture Notes, I will review the theory of SIR/PHD and describe some ongoing research in various application areas. There are two parts. The first part is based on materials that have already appeared in the literature. The second part is just a collection of some manuscripts which are not yet published. They are included here for completeness.

Needless to say, there are many other high dimensional data analysis techniques (and we may encounter some of them later on) that should deserve more detailed treatment. In this complex field, it would not be wise to anticipate the existence of any single tool that can outperform all others in every practical situation. Real world problems generally require a number of passes to the same data. Different approaches often lead to different structural findings at various stages. Serious readers should try out as many methods as possible on their own.

I started writing a preliminary version of this Lecture Notes in 1991-1992 when I had the chance to teach a seminar course at UCLA on High Dimensional Data Analysis. At that time, SIR/PHD has just begun to appear in official journals. This makes the writing of a book very difficult because most works are yet to be published. Even though such materials have later been used in similar courses and workshops, I hardly have the mood to rewrite it. The real opportunity finally came last year when the colleagues at Institute of Statistical Science, Academia Sinica, initiated the idea of this Lecture Series. I figured that I would have more to write now because there have many new exciting developments along the line. Most noteworthy are the books of Cook(1998), and Cook and Weisberg (1994). I admire the way they presented the ideas, which are not far from I really want to say. It is such a remarkable achievement for them to find a lucid language in dealing with the difficult subject of how to think of graphics in rigorous statistical terms. As expected, with the new language, they have generated many new ideas and useful techniques that go beyond SIR/PHD.

For this book, I am still using the words as I originally thought about the subject of dimension reduction. The basic material is narrowly focused on the development in which I am directly involved. Thus there is no serious attempt to be comprehensive in surveying the whole literature on SIR/PHD. For many researchers, SIR/PHD is still a novel technique and new results are still waiting to be published.

I would like to thank a lot people who one way or the other have helped me in the development of SIR/PHD, including all my co-authors, colleagues, students, friends, and many anonymous referees. I would also like to thank Dr. Chen-Hsing Chen who is in charge of the Lecture Series. My writing of this Lecture Notes would have further been delayed without the persistent request from him. I really appreciate his patience. Finally, many of the computer and graphical outputs are put together by Dr. Chun-Houh Chen, who has been working with me over years. Without his devotion to critical programming works and his many good ideas in implementing SIR/PHD, the progress would be much slowed.

Acknowledgment: Over years, the research of Li has been supported in part by NSF grants.

Contents

Part I

SIR/PHD - THEORY AND PRACTICE

Chapter 1

A Model for Dimension Reduction in Regression

Statistical data analysis is an extremely versatile area. Among the various aspects and activities involved, we shall begin with the discussion about the dual role of dimension reduction and data visualization in this chapter. After a brief account on the graphical facilities that modern computer is equipped with, the stage will be set up in the context of regression analysis. A new statistical model that helps define this dual role is introduced. Our model shall serve as the foundation for theoretical exploration of many methods to be introduced in later chapters, including Sliced Inverse Regression (SIR) and Principal Hessian Directions (PHD).

1.1 Static and dynamic graphics.

Today's desktop/laptop computer has great graphical facility. It allows us not only to draw high quality plots easily but also to interact with graphics. Our discussion on such animated graphical features will be short. We assume that our readers already have some hands-on experience with dynamic graphics.

1.1.1 Graphical tools.

A picture is worth of thousand words. As we all agree, graphics is a powerful way of summarizing the data. Informally, graphical procedures exploit both our talent in recognizing complex patterns and our creative imagination thus inspired (Huber 1987).

Traditional plots like histograms and scatterdiagrams have been routinely used in displaying one or two variables. Normal probability plots and residual plots are popular for model checking. Plots of time series can detect periodicity or other forms of trend.

In principle, any plot consists of just points on a two-dimensional plane. How to add a third dimension to a plot is more challenging. For example, we can display two-D scatterplots for X vs. Y, Y vs. Z, or X vs. Z. But these plots only tell us about two-D marginal distributions. To sense the geometric shape of a three-D structure, animation is effective and one way to do so is to rotate the plot. This idea has been experimented in the pioneering work on *PRIM-9* (Fisherkeller, Friedman, and Tukey 1974), and Huber's versions for *PRIM-ETH* and *PRIM-H* (Cleveland and McGill 1988). After many years of evolution and with substantive efforts by many other developers, graphical environment has become amazingly friendly. In XLISP.STAT(Tierney 1989) for example, we can issue a simple command like

```
    "(spin-plot (list x y z))"
```

and a 3-D scatterplot of variables X, Y, Z will appear instantly on the computer screen, ready for us to "rotate". What we first see on the screen is a static two-D scatterplot of Y (in the vertical axis) against X (in the horizontal axis). The third variable Z is temporarily out of sight. We can imagine that it is kept in the axis perpendicular to the screen. Three buttons for specifying the rotation along the x-axis, the y-axis, or the z-axis are given under the plot. We can point to any button with the cursor and the computer will react instantly by displaying a sequence of projections. This happens so smoothly that the user can feel like a real three dimensional object is rotated. The speed of rotation can also be customized easily. Rotation plot is included in most statistical packages.

Example 1.1

An exercise problem in Rice(1988, page 506) involves a data set whose structure is not easy to reveal via conventional regression analysis. There are three variables in the data. The response variable Y variable is the modules of natural rubber. The two regressors are the amount of decomposed dicumyl peroxide (x_1) , and the temperature (x_2). We fit the data with a multiple linear model first, but the result is not good. One suggestion is to incorporate an interaction term into the model:

$$E(y|x) = c_1x_1x_2 + c_2x_1 + c_3x_2 + c_4$$

However, the residual plot, Figure 1.1(a), again shows that this model is still unsatisfactory. In fact, the line patterns observed in this plot cannot be removed by incorporating higher order polynomials or by other smoothing techniques. For example, Figure 1.1(b) shows the residual pattern after fitting points in Fig 1.1(a) with a parabolic curve. The 3-D plot of Y against x_1 and x_2, Figure 1.2(a)-(d), explains very well why the above model building process is unsuccessful. The response surface shows a clear pattern consisting of straight lines with different slopes and intercepts, each line being parallel to the x_1 axis. This indicates that the pattern of interaction between the two regressors is more complicated. A better model is suggested in Li(1993):

$$E(Y|x) = g_1(x_1)x_2 + g_2(x_1)$$

In this model, Y is linear in x_2 , but with both the slope $g_1(x)$ and intercept $g_2(x_1)$ depending on x_1. Dynamic graphical techniques can perform many tasks that cannot be done with the conventional static graphics. Common dynamic features include highlighting, brushing, linking, scrolling, case removing, rescaling and many more. An introduction to this area can be found in books like Wegman and DePriest(1986), Cleveland and MacGill (1988)). In this book, we shall use Xlisp-Stat as the primary graphical/computing environment.

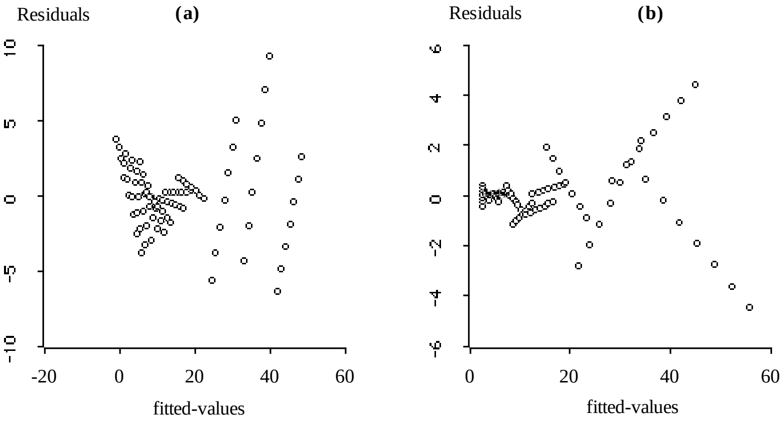


Figure 1.1: Plot of residuals for two models : (a) Linear; (b) Interaction.

1.1.2 Boston housing data.

Let's use the Boston housing data, Harrison and Rubinfeld (1978), to illustrate a few key ideas in the above discussion. This data set has 14 variables and 506 cases. Each case represents a census tract in Boston Standard Metropolitan Statistical Areas. The variable of primary interest is the logarithm of the median value of owner occupied homes. This and other 13 variable names are given in Table 1.1. This data set has many interesting nonlinear patterns.

Table 1.1: Variables for Boston Housing Data.

Y	logarithm of the median value of owner-occupied home
x_1	crime rate by town
x_2	proportion of town' residential land zoned for lots greater than 25,000 square feet
x_3	proportion of nonretail business acres per town
x_4	=1 if tract bounds Charles River, =0 otherwise
x_5	nitrogen oxide concentration in pphm
x_6	average number of rooms in owner units
x_7	proportion of owner units built prior to 1940
x_8	weighted distances to five employment centers in the Boston region
x_9	index of accessibility to radial highways
x_{10}	full property tax rate
x_{11}	pupil-teacher ratio by town school district
x_{12}	black proportion of the population
x_{13}	proportion of population that is in the lower status

First consider the scatterplots between any two variables. There are in total $\binom{14}{2} = 91$

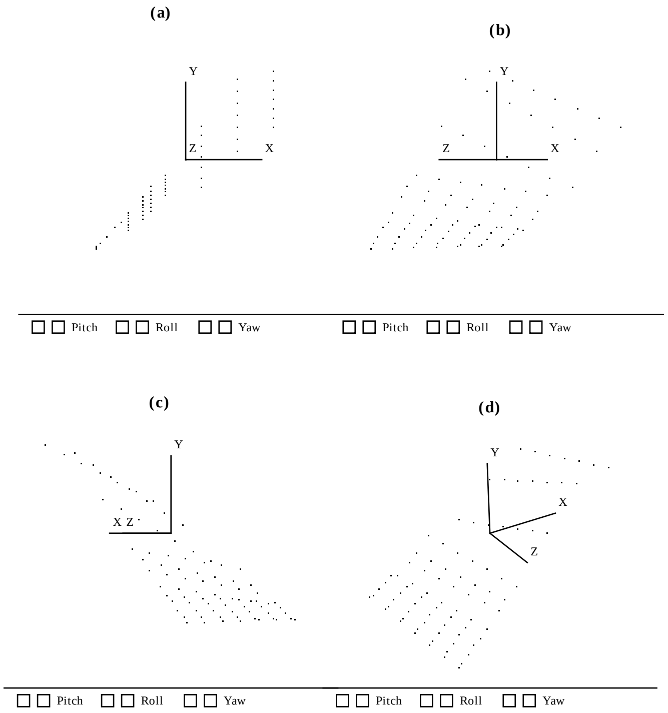


Figure 1.2: 3–D plot of y against x1 (in x-axis) and x2(in z-axis).

such plots. Scatterplot matrix is a useful devise for organizing these plots. In Xlisp-Stat, suppose we already have created four variables, Y, x_1, x_2, x_3. Then if we issue a command

`"(scatterplot-matrix (list Y x1` x_2 x_3`))"`

, we shall see a plot consisting of the scatterplots between any two variables, organized into a matrix of $4 \times 4 = 16$ smaller boxes. The boxes along the diagonal line give the labels of the variables. For example, the top-leftmost box is the scatterplot of $x4$ (vertical axis) against Y (horizontal axis).

For Boston Housing data, displaying a scatterplot matrix with more than 10 variables in a limited space is not practical. It will require a higher resolution that exceeds the current capacity of Xlisp.stat and other packages. Figures 1.3(a), (b) show two scatterplot matrices for some variables selected from the 14 variables.

A quick glance over the two scatterplot matrices, the variable of CRIME RATE, coded as “var 1” appears rather unusual. A group of cases stand out. They Form a vertical line in the scatterplot against x_2 and against x_3. We can highlight this group and examine the scattering patterns in each plot. When highlighted, all points in each box are darkened. We can also remove this group of data and focus the study on all other points.

We may also want to examine 3-D rotation plots. But we will have $\binom{14}{3} = 364$ plots to inspect now. Without further guidance on which ones to see first, this task has become really burdensome.

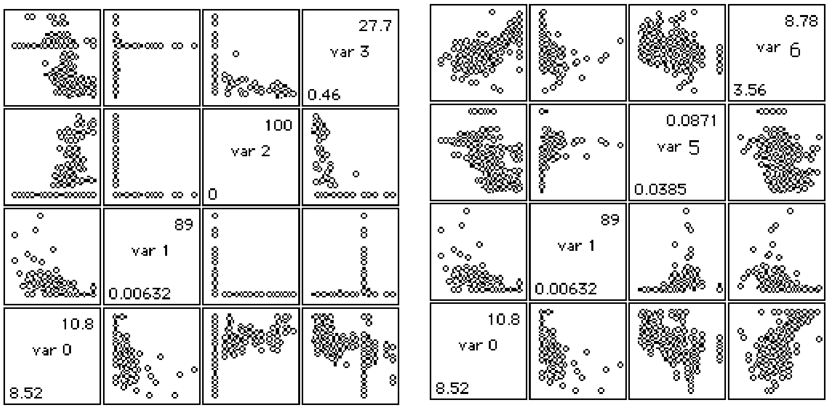


Figure 1.3: Two scatterplot matrices for Boston Housing Data. (a) var0=y, var1=x1, var2=x2, var3=x3; (b) var0=y, var1=x1, var2=x5, var3=x6

1.2 A regression paradigm.

Regression analysis is one area where graphical techniques are indepensible. Figure 1.4 is a paradigm which describes some most common steps in routine practice for small dimensional problems.

Suppose a data set with Y as the response variable and $\mathbf{x}$ as a p dimensional input variable is given. We begin with a preliminary study. Histograms, box plots, or normal probability plots for each variable can be inspected. We then draw the 2-D scatterplot of Y against each component of $\mathbf{x}$, or even turn to the 3-D scatterplots if necessary, hoping to gain some ideas about the regression function. When p is small, say 1, or 2, these plots would normally be quite useful. After such exploration, we can move on and fit the data with a model. The model may be derived from some available scientific theory or it may be taken from the literature or earlier studies. But for the majority of cases, the appropriate models are yet to be discovered. We have to turn to the empirical way of building the model, which begins with a simple and plausible model as suggested by the plots we have examined. Statistical parameter estimation with least squares or mle is conducted for fitting the data to the specified model. After that, we can bring in various methods of residual diagnostics for incorporating more subtle structures. We can detect outliers, find influential points and so on. We can also verify the model assumptions. Here plotting is called for again. Normal probability plots can be used to see if error distributions are standard or not. Residual plots of various kinds can reveal lack-of-fit patterns. This leads to better suggestions on how to refine the existing models or on what alternative models to use.

After the residual analysis, we can go through the same process again with the modified model. It is hoped that after a few rounds of iterations, we may come up with a reasonable

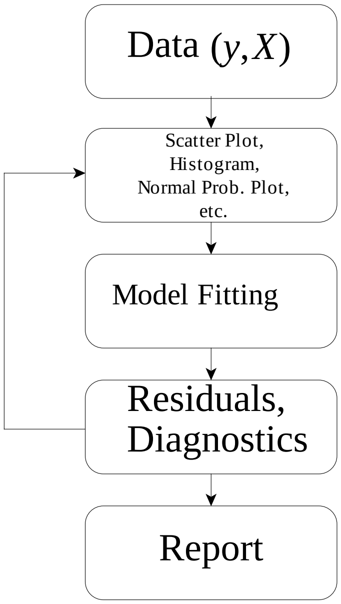


Figure 1.4: A Regression Diagram for Low Dimensional Problems.

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 **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 **x** that has the largest variance among all **b** with unitary length:

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

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

After finding the first direction, say $\mathbf{b}_1$, we repeat the same procedure by restricting to those "uncorrelated" directions **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 **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 **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).

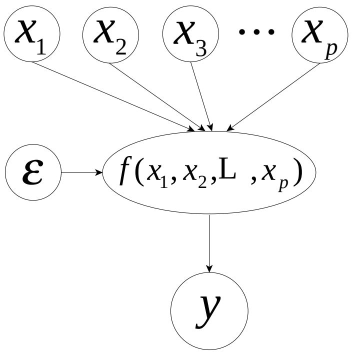


Figure 1.5: General Regression Model.

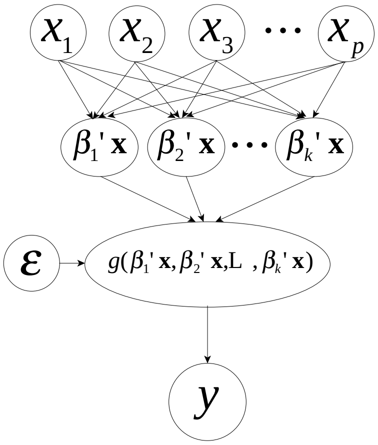


Figure 1.6: Model 1.1.

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.

1.4.6 Discrepancy measure.

We conclude this section by discussing the question of how to evaluate the effectiveness of an estimated e.d.r. direction. An obvious criterion is to evaluate the squared Euclidean distance between the estimated e.d.r. direction b (normalized to have the unitary length) and the true e.d.r. space $\mathcal{B}$. But the result will be sensitive to the scale change in $\mathbf{x}$. To avoid this problem, the following affine-invariant criterion will be considered:

$$R^2(b) = \max_{\beta \in \mathcal{B}} \frac{(b'\Sigma_{\mathbf{x}}\beta)^2}{b'\Sigma_{\mathbf{x}}b \cdot \beta'\Sigma_{\mathbf{x}}\beta} \tag{1.7}$$

, the squared multiple correlation coefficient between the projected variable $b'\mathbf{x}$ and the ideally-reduced variables $\beta_1'\mathbf{x}, \cdots, \beta_K'\mathbf{x}$. For a collection of K estimated directions $b_1, \cdots, b_K$ which generate a linear subspace B, we use the squared trace correlation, denoted by $R^2(B)$, as our criterion; i.e., the average of the squared canonical correlation coefficients between $b_1'\mathbf{x}, \cdots, b_K'\mathbf{x}$ and $\beta_1'\mathbf{x}, \cdots, \beta_K'\mathbf{x}$ (Hooper 1959). It is also reasonable to replace $\Sigma_{\mathbf{x}}$ by the sample covariance matrix in our definition of the criteria.

Chapter 2

Sliced Inverse Regression: Basics

In this chapter, a method for finding e.d.r. directions, sliced inverse regression (SIR), is introduced. An algorithm for implementing SIR is described first. Then a theory for justifying SIR is presented. As it will soon become clear, SIR is based on an unorthodox way of thinking about regression data.

2.1 Forward and inverse regression.

As illustrated in Figure 1.5, the natural way to think about regression is a forward one, going from $\mathbf{x}$ to Y. In general terms, forward regression decomposes the joint distribution of Y and $\mathbf{x}$ into two parts: the marginal density $k(\mathbf{x})$ of $\mathbf{x}$ and the conditional density $h(y|\mathbf{x})$ of Y given $\mathbf{x}$. The primary interest is on $h(y|\mathbf{x})$. Of special importance is the regression function (also called the response surface) $E(Y|\mathbf{x})$ or the second moment $var(Y|\mathbf{x})$.

Unlike other functional-approximation and curve smoothing methods, SIR does not follow the above one-way traffic of going from $\mathbf{x}$ to Y. Instead, SIR reverses the role of $\mathbf{x}$ and Y. We treat Y as if it were the independent variable and treat $\mathbf{x}$ as if it were the dependent variable. This fundamental difference can be amplified when one asks “given $\mathbf{x} = \mathbf{x}_o$, what value will Y take?” One straightforward conventional answer would be “examine the data points close to $\mathbf{x}_o$ and take their average Y values”. But the SIR methodology would respond indirectly by phrasing the question first in a different way: ”given $Y = y$, what values will $\mathbf{x}$ take ?” Instead of just smoothing local information, SIR intends to gain global insight on how Y changes as $\mathbf{x}$ changes by studying the reserve - how the associated x region varies as Y varies.

At first sight, this way of thinking may not appear natural at all. But at least, one advantage is immediate. In forward regression, the response surface $E(Y|\mathbf{x})$ is p-dimensional, which is very difficult to estimate directly. Most smoothing techniques perform poorly because of lack of sufficient data points in each relevant local region if p is large. Huber(1985) gives a nice account on this and related aspects of “curse of dimensionality”. However, for inverse regression, the conditional expectation $E(\mathbf{x}|Y)$ can be taken one coordinate at a time $E(x_i|Y)$, for $i = 1, \cdots, p$. The estimation of $E(x_i|Y)$ should be easy to handle because this is just a one-dimensional curve smoothing problem. This is the reason why we can side-step 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	·········
·	·········
·	·········
·	·········
Y_n	$\mathbf{x}_n(= (x_{n1}, x_{n2}, \cdots, x_{np})')$

Table 2.2: SORTING by Y and SLICING.

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

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.

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

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	$\cdots\cdots\cdots$
$\cdot$	$\cdots\cdots\cdots$
$\cdot$	$\cdots\cdots\cdots$
$\cdot$	$\cdots\cdots\cdots$
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 :

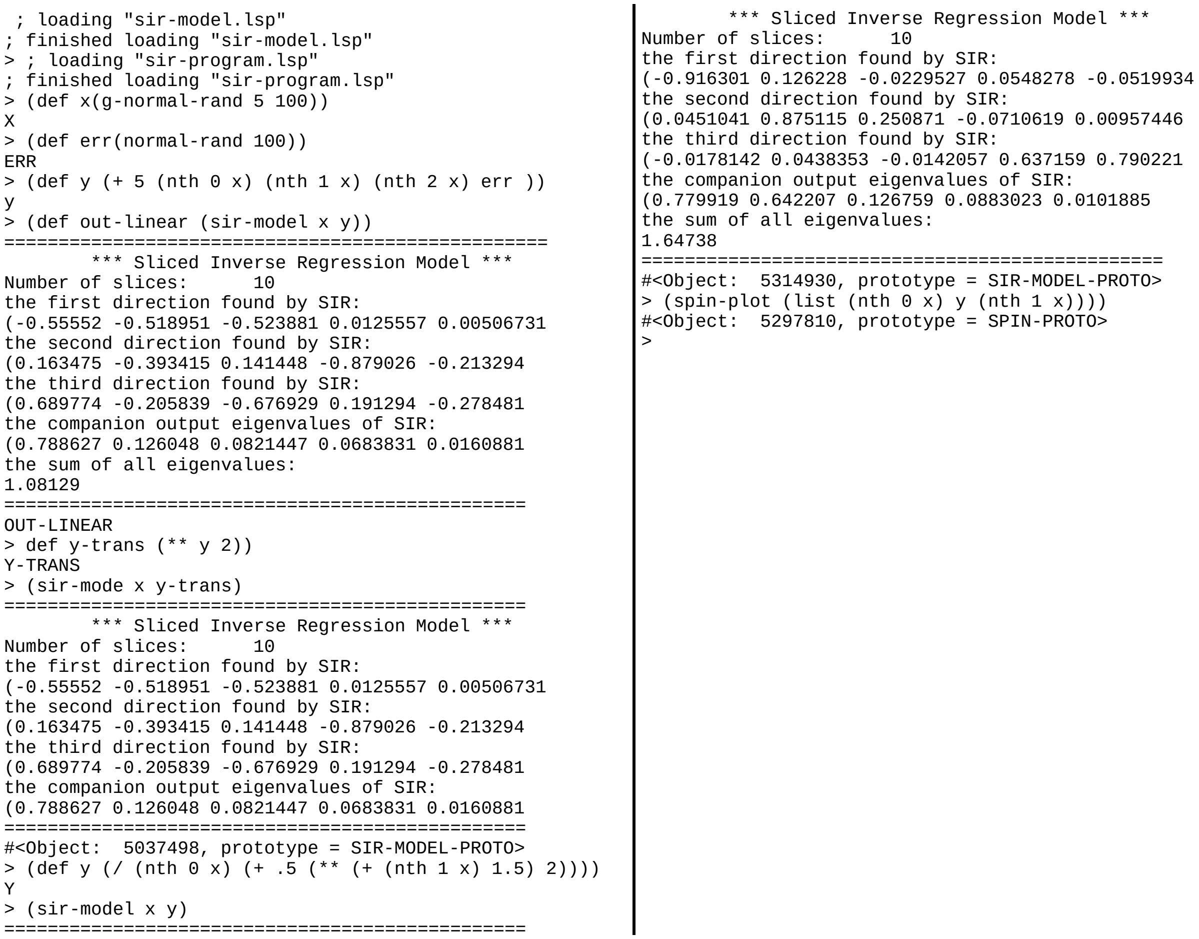
$$A\hat{\Sigma}_\eta A'\hat{v}_i = \hat{\lambda}_i\hat{v}_i$$

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

$\hat{\beta}_i = A'\hat{v}_i$. 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, $x_1, \cdots, x_p$, 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>
>
```

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.

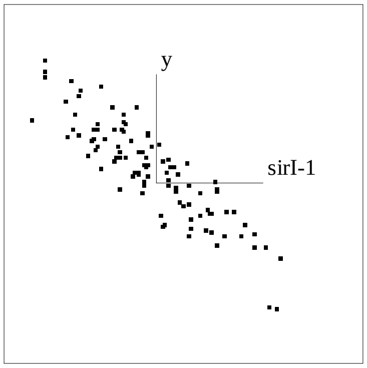


Figure 2.1: SIR View for Example 1. Linear Model.

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.

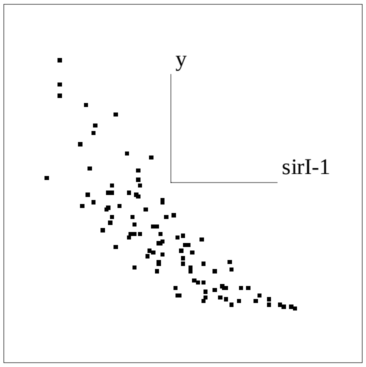


Figure 2.2: SIR View for Example 2. Transformation-Outside Model.

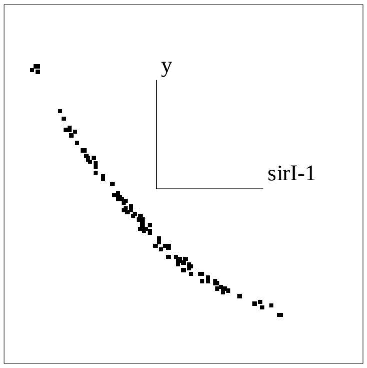


Figure 2.3: SIR view for Example 2. Transformation-Inside Model.

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.

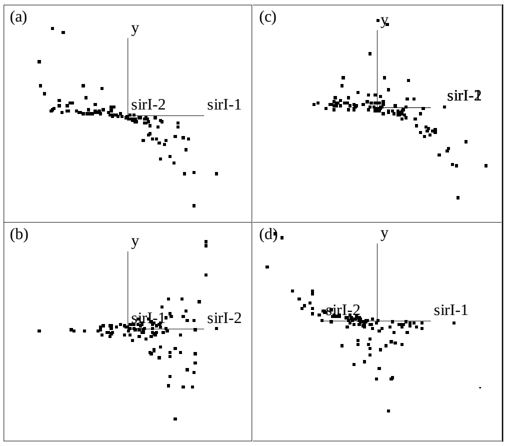


Figure 2.4: SIR View for Example 4. Rational Function.

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

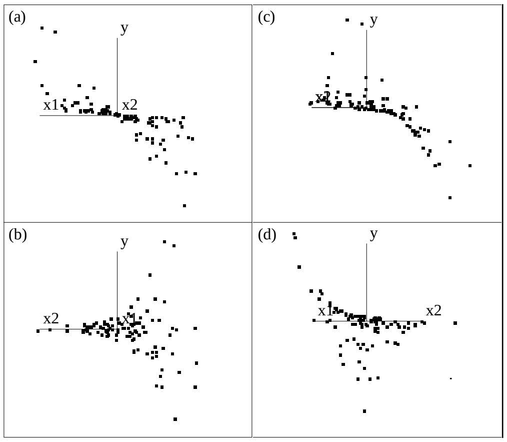


Figure 2.5: Best View for Example 4. Rational Function.

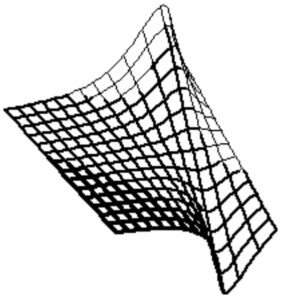


Figure 2.5(e): Response surface of Example 4, Rational function.

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-

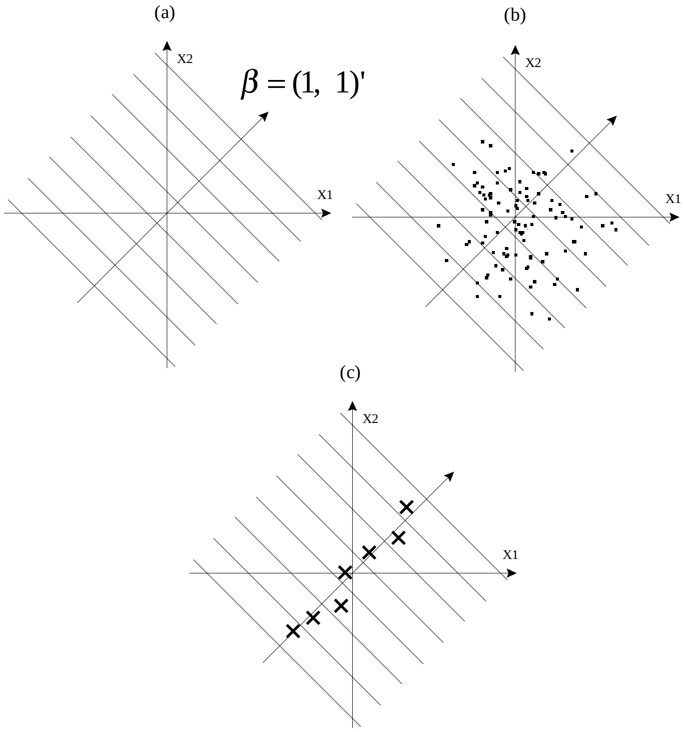


Figure 2.6: Contour Plot of $y = g(\beta'_{\mathbf{x}})$ and SIR.

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* **x***.*

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

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

(L.D.C.) is satisfied when the distribution of **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

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

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.

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

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.

Chapter 3

Sampling Properties of SIR

In this chapter, we discuss the sampling behavior of SIR. First we establish the root-n consistency in Section 3.1. An asymptotic formula which describes how close the SIR directions are to the e.d.r. space is given. Then we derive a chi-square test for determining the number of significant nonzero eigenvalues. This provides an estimate of the reduced dimension K in the dimension reduction model (1.1) of chapter 1. In Section 3.3, we discuss the issue of how many slices should be used. Other sampling aspects of SIR are given in Section 3.4.

3.1 Consistency of SIR.

We shall assume that H is fixed, and the range of Y is partitioned into H intervals, $I_h, h = 1, \cdots$. Slice h consists of cases with $\mathbf{x}_i \in I_h$.

3.1.1 The root n rate.

Let $p_h = P\{y \in I_h\}$, $\mathbf{m}_h = E(\mathbf{x}|y \in I_h)$. Elementary probability theory shows that $\bar{\mathbf{x}}_h$ converges to $\mathbf{m}_h$ at rate $n^{-1/2}$. Let V be the matrix $\Sigma_{h=1}^{H} p_h(\mathbf{m}_h - E\mathbf{x})(\mathbf{m} - E\mathbf{x})'$. It is clear that the $\hat{\Sigma}_\eta$ converges to V at the root n rate. Let $\mathbf{b}_j$ be the jth eigenvector for the eigenvalue decomposition:

$$V\mathbf{b}_j = \lambda_j \Sigma_{\mathbf{x}} \mathbf{b}_j$$

The SIR direction $\hat{\beta}_j$, is seen to converge to the corresponding eigenvector $\mathbf{b}_j$ at the root n rate. Now we use Theorem 2.1 and the simple identity $\mathbf{m}_h = E(E(\mathbf{x}|y)|y \in I_h)$ to see that $\mathbf{b}_j$ will fall in the e.d.r. space.

The case that the range of each slice varies in order to ensure an even distribution of observations is related to the following choice of intervals:

$$I_h = (F_y^{-1}((h-1)/H), F_y^{-1}(h/H)),$$

where $F_y(\cdot)$ is the c.d.f. of y. The root n consistency result still holds.

3.1.2 The descripency measure.

We can approximate the expectation of $R^2(\hat{\mathcal{B}})$, the squared trace correlation between $\beta_k\mathbf{x}$'s and $\hat{\beta}_k\mathbf{x}$'s (see the last section in Chapter 1). For the normal $\mathbf{x}$, we have the following simple approximation :

$$ER^2(\hat{\mathcal{B}}) = 1 - \frac{p-K}{n}(-1 + \frac{1}{K}\sum_{k=1}^{K}\frac{1}{\lambda_k}) + o(\frac{1}{n})$$

where λ_k is the k^{th} eigenvalue of V. A crude estimate of this quantity is given by substituting λ_k with the k^{th} largest eigenvalue of $\hat{V}$. An intuitive interpretation of this result is that the curvature of the inverse regression plays an important role in the effectiveness of using SIR to find the e.d.r. directions. If the curve is too straight, then we might not be able to find more than one directions.

3.1.3 Simulation.

First we use the linear model

$$y = x_1 + x_2 + x_3 + x_4 + 0x_5 + \epsilon \tag{3.1}$$

to generate $n = 100$ data points. The dimension p equals 5 and each component in $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_5)'$ and ϵ are independent and follow the standard normal distribution. There is only one component in this model, $K = 1$, and any vector proportional to $\beta = (1, 1, 1, 1, 0)$ is an e.d.r. direction. Three different estimates are reported here; all of them are based on the first component, $\hat{\beta}_1$, of SIR but with the number of slices $H = 5, 10, 20$ respectively. A summary of the results after 100 replicates is given in Table 3.1. This table provides the mean and the standard deviation (in parentheses) of $\hat{\beta}_1$ after some length and sign adjustments to ensure that each $\hat{\beta}$ has the unitary length and its inner product with (1, 1, 1, 1, 0) is positive. The need for the adjustments is because we are interested in estimating the direction of β only (for now). We may change the sign and the length of $\hat{\beta}_1$ and still get the same direction estimate. Before the adjustment, the mean of our output $\hat{\beta}_1$ is nearly 0 because the output vector of step III in the SIR algorithm determines its sign at random(at step (iii) of the SIR, either $\hat{\eta}_1$ and $-\hat{\eta}_1$ can be the output maximum eigenvector). Thus our adjustment regularizes the output estimate for each of the 100 replicates so that the final average and the standard deviation reported here can meaningfully indicate the performance of $\hat{\beta}_1$ as a direction estimator.

As we see from this table, all estimates are very good. The means are all quite close to the normalized target (.5, .5, .5, .5, 0). How about the standard deviations ? For comparison, observe that the least squares estimate for each coordinate of β, based on the correct linear model, has the standard deviation $\frac{1}{\sqrt{n}} = .1$. Since the target vector is a half of β, the least squares estimate of the target has the standard deviation .05. So in this case, SIR is doing almost as well as if we do know the true model. On the other hand, it is also evident from the table that the sensitivity of the performance of the estimate to the number of slices is rather low.

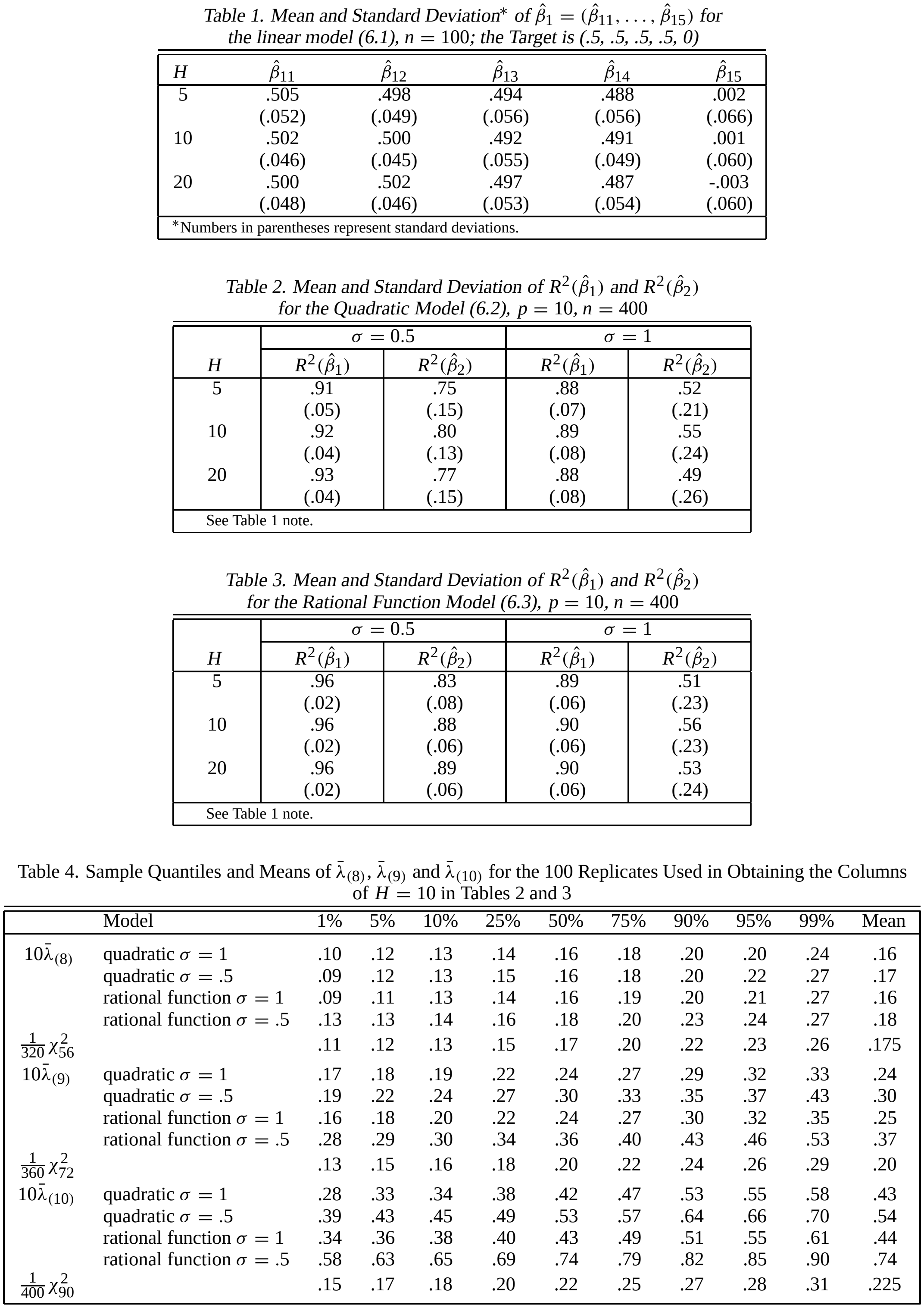


Table 1. Mean and Standard Deviation of $\hat{\beta}_1 = (\hat{\beta}_{11}, \ldots, \hat{\beta}_{15})$ for the linear model (6.1), $n = 100$; the Target is (.5, .5, .5, .5, 0)*

H	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	$\hat{\beta}_{14}$	$\hat{\beta}_{15}$
5	.505	.498	.494	.488	.002
	(.052)	(.049)	(.056)	(.056)	(.066)
10	.502	.500	.492	.491	.001
	(.046)	(.045)	(.055)	(.049)	(.060)
20	.500	.502	.497	.487	-.003
	(.048)	(.046)	(.053)	(.054)	(.060)

*Numbers in parentheses represent standard deviations.

Table 2. Mean and Standard Deviation of $R^2(\hat{\beta}_1)$ and $R^2(\hat{\beta}_2)$ for the Quadratic Model (6.2), $p = 10$, $n = 400$

	$\sigma = 0.5$		$\sigma = 1$	
H	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$
5	.91	.75	.88	.52
	(.05)	(.15)	(.07)	(.21)
10	.92	.80	.89	.55
	(.04)	(.13)	(.08)	(.24)
20	.93	.77	.88	.49
	(.04)	(.15)	(.08)	(.26)

See Table 1 note.

Table 3. Mean and Standard Deviation of $R^2(\hat{\beta}_1)$ and $R^2(\hat{\beta}_2)$ for the Rational Function Model (6.3), $p = 10$, $n = 400$

	$\sigma = 0.5$		$\sigma = 1$	
H	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$
5	.96	.83	.89	.51
	(.02)	(.08)	(.06)	(.23)
10	.96	.88	.90	.56
	(.02)	(.06)	(.06)	(.23)
20	.96	.89	.90	.53
	(.02)	(.06)	(.06)	(.24)

See Table 1 note.

Table 4. Sample Quantiles and Means of $\bar{\lambda}_{(8)}$, $\bar{\lambda}_{(9)}$ and $\bar{\lambda}_{(10)}$ for the 100 Replicates Used in Obtaining the Columns of $H = 10$ in Tables 2 and 3

	Model	1%	5%	10%	25%	50%	75%	90%	95%	99%	Mean
$10\bar{\lambda}_{(8)}$	quadratic $\sigma = 1$	.10	.12	.13	.14	.16	.18	.20	.20	.24	.16
	quadratic $\sigma = .5$	.09	.12	.13	.15	.16	.18	.20	.22	.27	.17
	rational function $\sigma = 1$	.09	.11	.13	.14	.16	.19	.20	.21	.27	.16
	rational function $\sigma = .5$	.13	.13	.14	.16	.18	.20	.23	.24	.27	.18
$\frac{1}{320}\chi^2_{56}$		.11	.12	.13	.15	.17	.20	.22	.23	.26	.175
$10\bar{\lambda}_{(9)}$	quadratic $\sigma = 1$	.17	.18	.19	.22	.24	.27	.29	.32	.33	.24
	quadratic $\sigma = .5$	.19	.22	.24	.27	.30	.33	.35	.37	.43	.30
	rational function $\sigma = 1$	.16	.18	.20	.22	.24	.27	.30	.32	.35	.25
	rational function $\sigma = .5$	.28	.29	.30	.34	.36	.40	.43	.46	.53	.37
$\frac{1}{360}\chi^2_{72}$		.13	.15	.16	.18	.20	.22	.24	.26	.29	.20
$10\bar{\lambda}_{(10)}$	quadratic $\sigma = 1$	.28	.33	.34	.38	.42	.47	.53	.55	.58	.43
	quadratic $\sigma = .5$	.39	.43	.45	.49	.53	.57	.64	.66	.70	.54
	rational function $\sigma = 1$	.34	.36	.38	.40	.43	.49	.51	.55	.61	.44
	rational function $\sigma = .5$	.58	.63	.65	.69	.74	.79	.82	.85	.90	.74
$\frac{1}{400}\chi^2_{90}$		.15	.17	.18	.20	.22	.25	.27	.28	.31	.225

Figure 3.1: which has Table 3.1–3.4

Turning to the multicomponent case, we shall concentrate on the case $K = 2$. Two models will be studied :

$$y = x_1(x_1 + x_2 + 1) + \sigma \cdot \epsilon \tag{3.2}$$

$$y = \frac{x_1}{0.5 + (x_2 + 1.5)^2} + \sigma \cdot \epsilon\ . \tag{3.3}$$

In addition to ϵ, x_1, x_2, we also generate $x_3, ..., x_p$, all variables being independent and following the standard normal distribution. We take $p = 10$ together with $\sigma = .5$ and $\sigma = 1$. The sample size is set at $n = 400$, which is small relative to the dimension $p = 10$. The true e.d.r. directions are the vectors in the plane generated by (1, 0, .., 0) and (0, 1, 0, ..., 0). The first two components of SIR will be used as estimates of e.d.r. directions. Recall the performance measure $R^2(\cdot)$ from section 3.1.2. Since $\Sigma_{\mathbf{xx}} = I$, $R^2(\hat{\beta}_1)$ is simply the square of the cosine of the angle between $\hat{\beta}_1$ and the space generated by the first two coordinates. Writing $\hat{\beta}_1$ as $(\hat{\beta}_{11}, \hat{\beta}_{12}, \cdots, \hat{\beta}_{1p})$, we see that $R^2(\hat{\beta}_1) = (\hat{\beta}_{11}^2 + \hat{\beta}_{12}^2)/\|\hat{\beta}_1\|^2$. The formula for $R^2(\hat{\beta}_2)$ can be given in a similar way. With the number of slices H set at 5,10, and 20 respectively, in Tables 3.2 and 3.3, we report the mean and the standard deviation of $R^2(\hat{\beta}_1)$ and $R^2(\hat{\beta}_2)$ after 100 replicates.

For both models, despite the change in the noise level, the first component is very close to the e.d.r.space as the R^2 values hover in the neighborhood of 90%. The second component is more sensitive to the noise level. But even for the high noise level case, the sample correlation between the projected one dimensional variable $\hat{\beta}_2\mathbf{x}$ and the perfectly-reduced data, the squared root of R^2, is still strong (above .7) on the average. Again, the number of slices has only minor effects on the results. It is interesting to observe that the SIR is doing better for the rational function model (3.3) than the quadratic model (3.2), despite the fact that the strength of the signal as measured by the standard deviation of $E(y|\mathbf{x})$ is weaker for the rational function model (about .8 for (3.3) v.s. about 2.0 for (3.2)). The key to the success of SIR hinges not on the signal-to-noise ratio, but on the eigenvalues of $cov(\mathbf{z}|y)$.

We do not report the average for $R^2(\hat{\mathcal{B}})$, because it is very close to the average of $\frac{1}{2}(R^2(\hat{\beta}_1) + R^2(\hat{\beta}_2))$.

We conclude this subsection by observing that in our models (3.2) and (3.3) , because of the affine invariance, we may change x_1, x_2 to $\beta_1\mathbf{x}$,$\beta_2\mathbf{x}$, where the β's are any two orthonormal vectors, and still obtain the same results as those reported in our tables.

3.2 Eigenvalues.

How many components are there in the data ? Perhaps this question is too ambitious to ask. But the companion SIR eigenvalues do provide us with valuable information for a more practical question:

$$Is\ an\ estimated\ component\ real\ or\ spurious?$$

Table 3.4 gives the empirical quantiles and the mean of $\bar{\lambda}_{(8)}$, $\bar{\lambda}_{(9)}$, and $\bar{\lambda}_{(10)}$, the averages of the smallest 8 , 9, and 10 eigenvalues, for the same 100 replicates used in obtaining the columns of $H = 10$, in Tables 3.2 and 3.3. (the conclusions are similar for other H's). For $\bar{\lambda}_{(8)}$, the numbers are close to the rescaled χ^2 values. Thus guided by the χ^2, not very often we will falsely conclude that the third component is real (or mistakenly claim that there are at least 3 components in the data).

Turning to $\bar{\lambda}_{(9)}$, we expect the numbers to be larger than what are given by using the rescaled χ^2 that falsely assumes only one component in the model. For the rational function model with $\sigma = .5$, this is clearly so, as we see that the 1% quantile of $\bar{\lambda}_{(9)}$ is close to the 99% quantile of the rescaled χ^2. Thus in this case, we correctly infer that there are at least 2 components in the model in each of the 100 replicates. As confirmed by the corresponding $R^2(\hat{\beta}_2)$ reported in Table 3.3, high value of $\bar{\lambda}_{(9)}$ leads to good performance of $\hat{\beta}_2$ as an e.d.r. direction. On the other hand, the distribution of $\bar{\lambda}_{(9)}$ for the quadratic model with $\sigma = 1$ shows a substantial overlap with the rescaled χ^2. This is reflected in the relatively lower average and higher standard deviation of $R^2(\hat{\beta}_2)$ in Table 3.2. But a positive point is that by comparing $\bar{\lambda}_{(9)}$ with the rescaled χ^2, we realize that our data do not strongly support the claim that the second component is real.

Finally, $\bar{\lambda}_{(10)}$ is well above the associated χ^2, assuring the high average and the low standard deviation of $R^2(\hat{\beta}_1)$ in all cases.

3.2.1 Chi-squared test.

As argued before, in order to be really successful in picking up all K dimensions for reduction, the inverse regression curve can not be too straight. In other words the first K eigenvalues for V must be significantly different from zero compared to the sampling error. This can be checked by the companion output eigenvalues.

The asymptotic distribution of the average of the smallest $p - K$ eigenvalues, denoted by $\bar{\lambda}_{(p-K)}$, for $\hat{V}$ can be derived, based on perturbation theory for finite dimensional spaces (Kato 1976, chapter 2). For normal **x**, we have the following result.

Theorem 3.1. *If* **x** *is normally distributed, then* $n(p - K)\bar{\lambda}_{(p-K)}$ *follows a* χ^2 *distribution with* $(p - K)(H - K - 1)$ *degrees of freedom asymptotically.*

3.2.2 Eigenvalues and the assessment of K.

An outstanding dilemma facing all data-analysts is that the more you screen, the more you may find. Good or bad ? While it is desirable to discover as many patterns as possible so one can have a better chance to develop a new theory, this also increases the chance of a false alarm. It is helpful to know whether an observed pattern is spurious or not. Yet this is by no means an easy task, and there is not much discussion on this issue in the literature. For our problem, how many components SIR finds are really there ? The output eigenvalues in the eigenvalue decomposition step of SIR are helpful in answering this question.

First observe that following from Theorem 3.1, we see that theoretically the smallest $p - K$ eigenvalues have to be 0. But in order to be really successful in picking up all K dimensions for reduction, the inverse regression curve can not be either degenerated or close to being degenerated. In other words the first K eigenvalues for the covariance matrix must be significantly different from zero compared to the sampling error. This can be checked by using the companion output eigenvalues of SIR. In the last section, we have derived the asymptotic distribution of the total of the smallest $p-K$ eigenvalues. We may use that result to give a conservative assessment of the number of components in the model.

For $j = 0, 1, 2, ...,$ we define

$$\textit{P-value}_j = P\{\chi^2_{(p-j)(H-j-1)} \geq n(n-p)\bar{\lambda}_{(p-j)}\}$$

This sequence of *P-values* can be used to indicate how many components are found by SIR. A simple forward selection procedure is to start with $j = 0$. If $\textit{P-value}_j$ is less than say .05, then we may claim that there are at least $j+1$ components. Go to the next j till we fail to make the claim. Of course, we may have many other selection procedures to use. Mallows(1973, Technometrics) pointed out the merit of inspecting the plot of the whole sequence of the C_p measures. His point applies to our case too. Thus a sudden jump from a small P-value to a large P-value serves as a better indication of where to stop. We also found the sequence of eigenvalues themselves are often good indicators of how many components found by SIR are worth of close inspection. Typically a value more than .25 is noteworthy. Later on, we shall interprete eigenvalues as R-squared values for multiple linear regression of some suitable transformations of y against $\mathbf{x}$.

Chapter 4

Applying Sliced Inverse Regression

In this chapter, we illustrate how the methodology of sliced inverse regression can provide some crucial visual foresights needed in constructing forward regression models. In application, SIR can be used in concert with other regression methods at various stages of data analysis.

We begin with a small data set in Section 4.1. SIR reveals that a power transformation on the response variable should be helpful.

The next example to be studied is the Boston Housing data as described in Chapter 1. We already know that this data set has many predictors. An immediate question is how to interpret so many coefficients in the SIR directions. This general issue is discussed in Section 4.2. We recommend a way of selecting important regressor variables to relief this burden.

In Section 4.3, we apply our recommendation to the Boston Housing data. After several passes through the data, we reach a very simple conclusion involving only three regressor variables - the crime rate(x_1), the number of rooms(x_6), and the proportion of poor (x_{13}).

In Section 4.4, we show that SIR can be applied in residual analysis. In our example, the main structure found by SIR is a parabolic trend. After removing the trend, the standard residual plot is inspected, which shows no evidence of lack-of-fit. We then apply SIR again to the residuals and find a heteroscedastic pattern.

In addition to analyzing empirically collected data, SIR can be used to visualize the shape of a deterministic function with several arguments. This is illusted by a push-pull circuit example in Section 4.5.

We conclude this chapter with a paradigm of regression analysis, showing where SIR can be applied.

4.1 Worsted yarn.

This is the Example J from Cox and Snell(1981, page 98-102). The data were collected from a 3^3 complete factorial design. This data set was originally analyzed by Box and Cox, using the power transformation model; see also Section 4.2.of Chapter 1. We simply apply SIR for y against $\mathbf{x} = (x_1, x_2, x_3)'$. One component is found in SIR-I. The plot of y against the first SIR projection shows an exponential curve. After taking the log transformation on y,

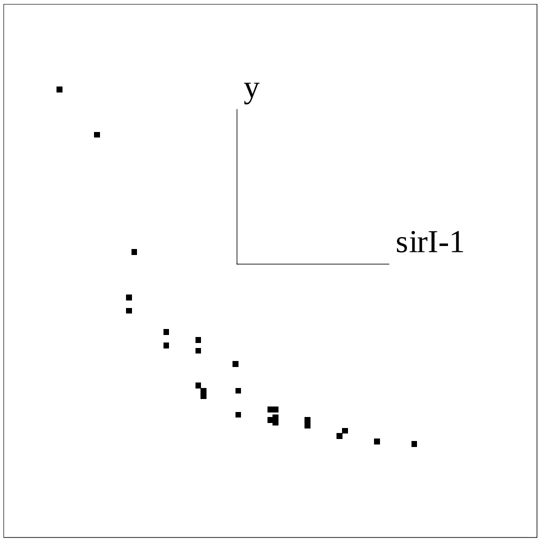


Figure 4.1: SIR view for Worsted yarn Data

the pattern is linear. Note that SIR is invariant under monotone transformation of y. We will come back to the transformation aspect of SIR later on.

4.2 Variable selection.

Boston Housing data has thirteen regressors. This means that each SIR direction would have 13 coefficients. An immediate concern how to interpret so many coefficients properly ? This general issue is discussed in this section.

First of all, Each variable has its own unit so a small coefficient does not mean that the corresponding variable should be ignored. A quick remedy is to report the result after standardizing each variable to have the same variance(=1). But this may not be enough.

In order to obtain a parsimonious description for the estimated e.d.r. space, it is appropriate to select a small subset of regressors for conducting SIR. Just like the variable selection in multiple linear regression, there are several ways of doing it. The following is one simple recommendation.

(1). Conduct SIR with all regressor variables included. Let $\hat{b}_1$, .., $\hat{b}_k$ be the estimated e.d.r. directions.

(2). Then find a projection from a small subset S_1 of regressors, denoted by $\hat{b}'_{s1}\mathbf{x}$, which is still reasonably close to the first projection $\hat{b}'_1\mathbf{x}$ with ,say, an R-square value of 90% (which amounts to about 18.5^o difference between the two projection angles) or better. This can be done by either forward or backward selection procedure in multiple linear regression by treating $\hat{b}'_1\mathbf{x}$ as y.

(3). After S_1 is selected, we then check if using variables from S_1 is good enough to approximate the second projection $\hat{b}'_2\mathbf{x}$ or not. If not, we should enlarge it and continue to the next projection. Let S be the final set of variables selected.

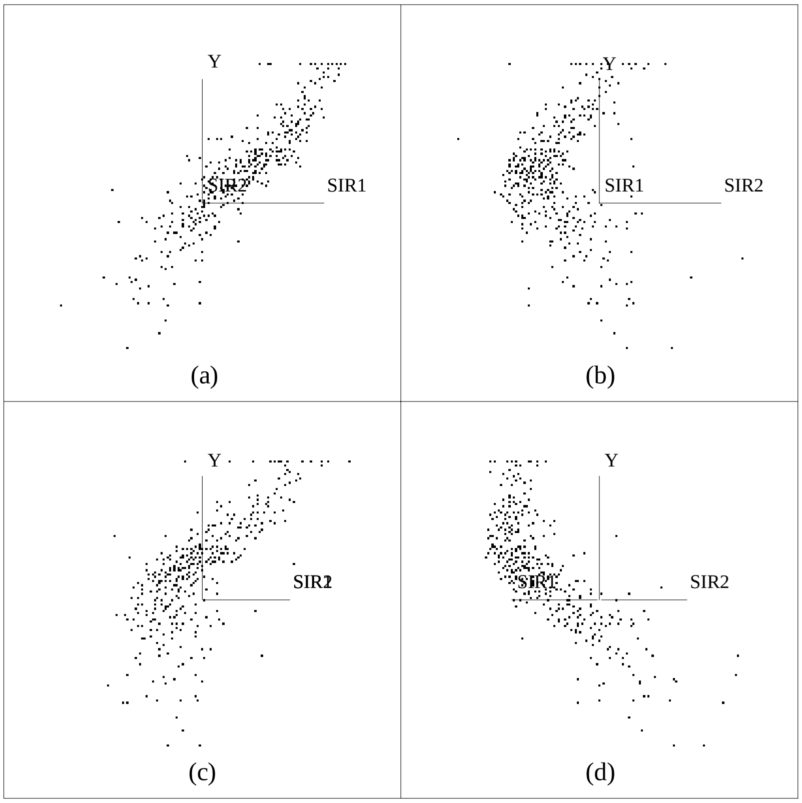
(4). Apply SIR again, this time using only the variables in S.

(5). If necessary, go through the variable selection procedure (2)-(4) again.

Note it is a good practice to compare the plot found from the reduced variables with the original one. If substantial difference is found, then some caution should be taken.

4.3 Boston housing data.

We first apply SIR to the Boston Housing Data (described in Chapter 1). with H, the number of slices, ranging from 10 to 30. The result, based on $H = 15$, is reported in Figures 4.2(a)-(d). As we rotate the cloud along the y-axis, it looks like a helix or slide. A further inspection of the eigenvalues , .82, .48, .20, .08, .05, $\cdots$, reveals that there are three significant components. Figure(4.3) provides the scatter-plot matrix for y and these three projection variables.



(a) (b) (c) (d)

Figure 4.2: SIR view for Worsted yarn Data

4.3.1 Crime rate.

When we carry out the variable selection procedure as described above (forward variable selection is used), the result is not illuminating. For the first projection, we can identify the main contributor to be x_{13}, proportion of poor, with x_6, average number of rooms in houses, as a close runner-up. Primary contributors of the second projection variable are harder to identify. The top candidate x_1, crime rate, leads seven other competing variables only marginally.

We take a closer look at the relationship of crime rate with other variables by inspecting scatterplots. As observed in Chapter 1, a special group of cases with high crime rate stand out from the others. All cases in this group share the same value in each of the following 5 variables : $x_2, x_3, x_9, x_{10}, x_{11}$.

4.3.2 The low crime rate group.

Excluding this high crime rate group, there are 374 cases remaining. We run SIR on these and now there are only two components significant. The pictures are similar to but sharper than the ones obtained from the whole sample. We are able to identify x_6 as the primary contributor of the first component. For the second component, x_1 and x_{13} are the top candidates. We then run SIR again, with x_1, x_6, x_{13} as the regressor variables this time (Figures (4.3(a)-(d)). The first component $\hat{b}_1'\mathbf{x}$ is clearly due to x_6, which has a correlation higher than .99 with $\hat{b}_1'\mathbf{x}$. The second component, can be described roughly as $x_1 + 30x_{13}$ adjusted by the first component for orthogonalization. These two SIR variates are nonlinearly correlated; see Figure 4.4.

Other values of H, ranging from 10 to 30, have provided essentially the same view. The logarithm transformation used to obtain Y is borrowed from Harrison and Rubinfeld (1978). This is not necessary because SIR is invariant under the monotone transformation of Y. SIR would still find the same projections if the original scale were used. In Figure 4.5 the original scale of house price is used. This can be compared with Figures 4.3. It appears that the logarithm transformation is unnecessary.

4.3.3 Intrepretation.

In this study, SIR identifies two key factors of different nature and provides a graphical summary. The variable x_6, average number of rooms, is a physical factor, which may reflect the construction cost and the practical utility of a house to some degree. It affects the physical condition of a house. The other variable $x_1 + 30x_{13}$, the crime rate and percentage of the poor, is a socio-economic factor. It reflects the desirability of the house's neighborhood which in turn affects the area's land value. SIR reveals the nonlinear association between these two factors. The importance of the physical factor is also confirmed by other methods; for example, the straightforward linear regression, the more complicated model fitting of Harrison and Rubinfeld, and ACE of Breiman and Friedman. In fact, in each of these studies, the physical variable is always the leading factor which accounts for the highest percentage of variation in the prediction equation. Because of this consistency from different studies,

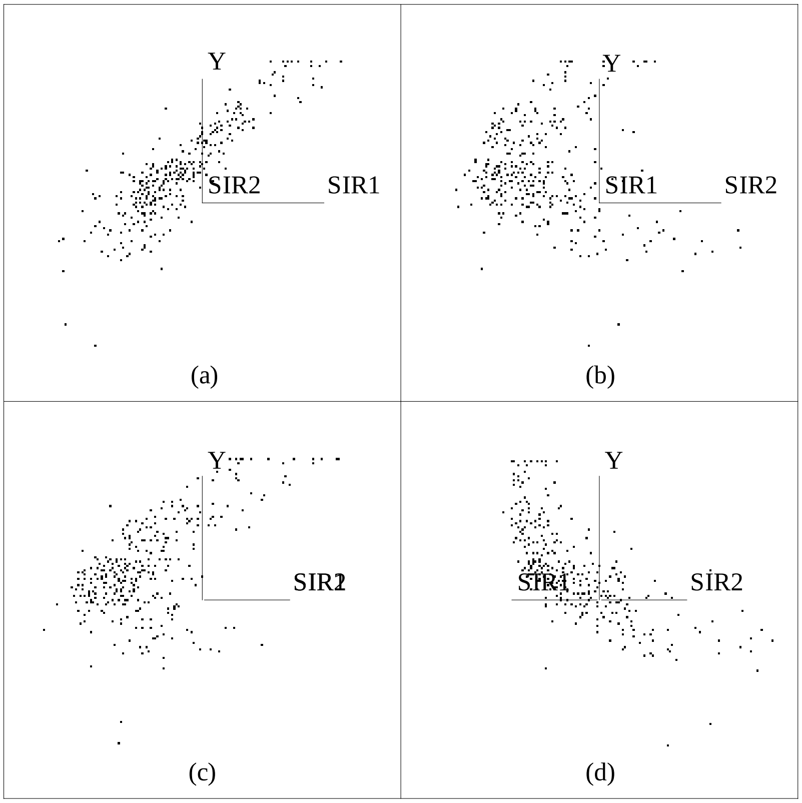


Figure 4.3: SIR's view on the subsample that excludes a high crime rate group with only x_1, x_6, x_{13} as the regressor variables. Y is the logrithm of the median value of owner-occupied house.

one might naively be forced to conclude that x_6 is the dominant factor. To challenge this simple-minded statement, we should resort to the helix type of nonlinear confounding pattern exhibited by the three-dimensional SIR plot. The second factor which appears equally important from the SIR plots, cannot be found from the other studies because their models have precluded structure like the one we found here a priori.

It is usually hard to draw any decisive conclusion from a single study. If the same helix shape of distribution also exists in data from other cities, for example, then the finding would be much more noteworthy. The graphics found here, however, is not available from linear regression or other methods. The exposure of the helix type data cloud offers an alarming diagnosis for methods aiming at the approximation the regression surface, which are sensitive to nonlinear confounding. We shall turn to this point again in Chapter 10.

Finally, we have run SIR with x_1, x_6, x_{13} and x_5 as the regressors. It turns out that the two components found are essentially the same as those without x_5, and that the correlation coefficient for the corresponding components is higher than .99 for each of the two directions found by SIR. This suggests that x_5, nitrogen oxide concentration, does not show a significant

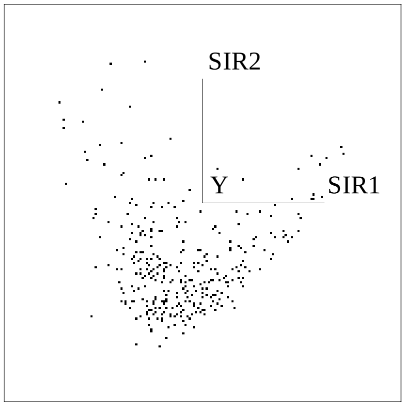


Figure 4.4: The scatterplot of the first two components in Figures 4.3 (a)-(d)

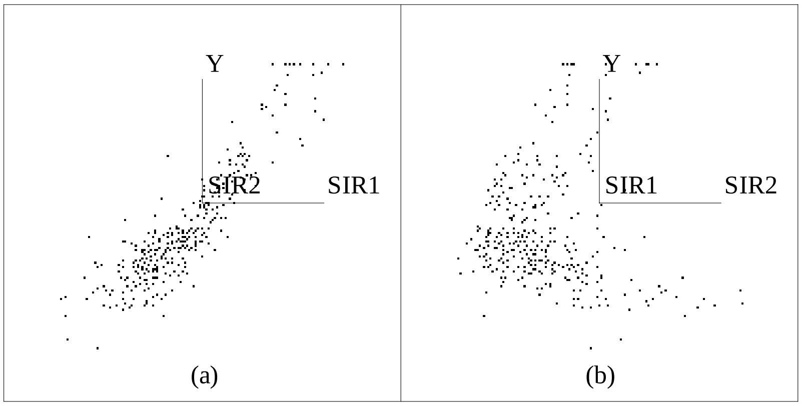


Figure 4.5: SIR's view on the subsample that excludes a high crime rate group with only x_1, x_6, x_{13} as the regressor variables. Y is the original scale median value of owner-occupied house.

role in affecting the relative housing prices in the low crime rate areas.

4.4 Structure removal.

Quite often, finer structure in the data can only be detected after the main structure is removed.

Example 4.4.1 Ozone data. We take a data set from Breiman and Friedman (1985), the data for studying the atmospheric ozone concentration in the Los Angeles basin. We use the daily measurement of ozone concentration in Upland as the output variable y and want to find its relationship with eight meteorological variables (see Table 8.4). There are $n = 330$ observations in the study. First, we apply SIR to the data and find one significant component.

This component is almost identical to the $\hat{b}'_{ls}\mathbf{x}$, the component found by the linear least squares fitting. For certain slice sizes, we can find a marginally significant second component as well, but we decide to ignore that. We then use a forward selection method to find out the important variables contributing to the first component. Three variables x_1, x_2, x_6 are found that explain more than 99% of the total variation of the first component. We run SIR again using only x_1, x_2, x_6 as the input variables. The scatterplot of y against $\hat{b}'_s\mathbf{x}$, the first component found. We use 30 slices here for SIR, but other choices yield almost identical scenes. The correlation between $\hat{b}'_{ls}\mathbf{x}$ and $\hat{b}'_s\mathbf{x}$ is above .99 as well.

A quadratic trend is visible from the SIR plot. After fitting a quadratic polynomial :

$$y = c_0 + c_1u_1 + c_2u_1^2 + \epsilon,$$

where u_1 denotes the variable $\hat{b}'_s\mathbf{x}$.

We finally apply SIR again to the residual and found one significant component, which gives the view Figure 4.6. An interesting triangle pattern of heterogeneity is seen.

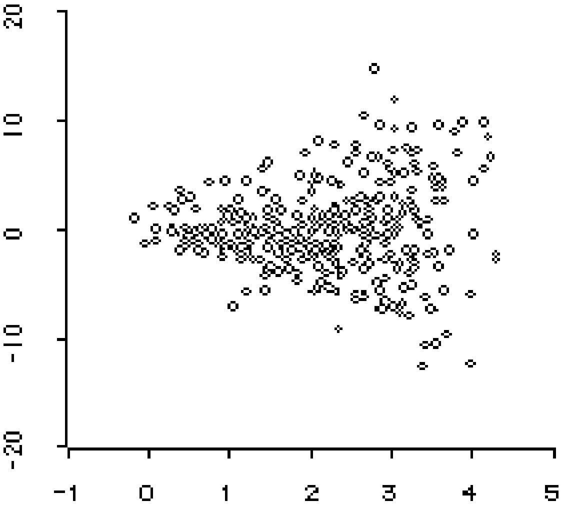


Figure 4.6: Ozone data. Residuals against the direction found by SIR

4.5 OTL push-pull circuit.

Most regression analysis techniques deal with data which are empirically collected. But there are many other cases in which the relationship between the input and output variables can be derived from physical/mathematical laws. The response is deterministic and is indeed already given. There are no data to analyze.

Figure 4.* depicts an OTL push-pull circuit used by a TV manufacturing company in Hanzou City of China(Chen et. al. 1985).

It is desired to have the voltage, V, as close to the nominal value, 6.0volts, as possible. The following Formula gives the relationship between V and six input variables, $\mathbf{x} = (R_{b1}, R_{b2}, R_f, R_{c1}, R_{c2}, log\ m_2)^t$:

$$V = (V_{b1} + .65)\frac{m_2 R_0}{m_2 R_0 + R_f} + (E_c - .65)\frac{R_f}{m_2 R_0 + R_f} + \frac{.74 R_f m_2 R_0}{(m_2 R_0 + R_f) R_{c1}} \tag{4.1}$$

where

$$V_{b1} = E_c R_{b2}/(R_{b1} + R_{b2}),\ R_0 = R_{c2} + 9,\ E_c = 12.$$

The first 5 input variables are resistances in kilo-ohms and the last one is an amplification factor. Due to diversity in quality for the components assembled, as well as the temperature changes or other environmental uncertainty, the actual values for these variables often deviate from their nominal values, causing the fluctuation in the actual output V. Attempts have been made by Chen et. al. to increase the output stability. They found appropriate quality levels for the components and the optimal nominal values for these variables by modifying Taguchi's method for parameter design and tolerance design.

We would like to "see" the shape of this function in a region of interest. To do this, we need to find directions that change the output variable the most. We shall use SIR to help find such directions. The data can be simulated by specifying a distribution for the input variables and apply the equation (4.1) to obtain the output values. For illustration, we specify the means and standard deviations for the input variables as :

$$\text{mean of } \mathbf{x} = (100, 60, 1.0, 1.2, 1.0, 3.9)$$
$$\text{standard deviation of } \mathbf{x} = (1.0, 1.0, 0.10, 0.12, 0.10, 1.1)$$

We simulate 1000 independent observations from a normal distribution. We apply SIR with $H = 15$. A few angles for the rotation plot of y against $\hat{u}_1 = \hat{b}_1\mathbf{x}$ and $\hat{u}_2 = \hat{b}_2\mathbf{x}$ is given by Figure 4.7. As we rotate the plot, a surface that approximates well the data points can be clearly seen.

The main shape of the function does vary as the domain of input changes. For another input specification, a linear surface is found.

In general, visualization help decide whether a linear approximation based on Taylor's expansion is adequate or not.

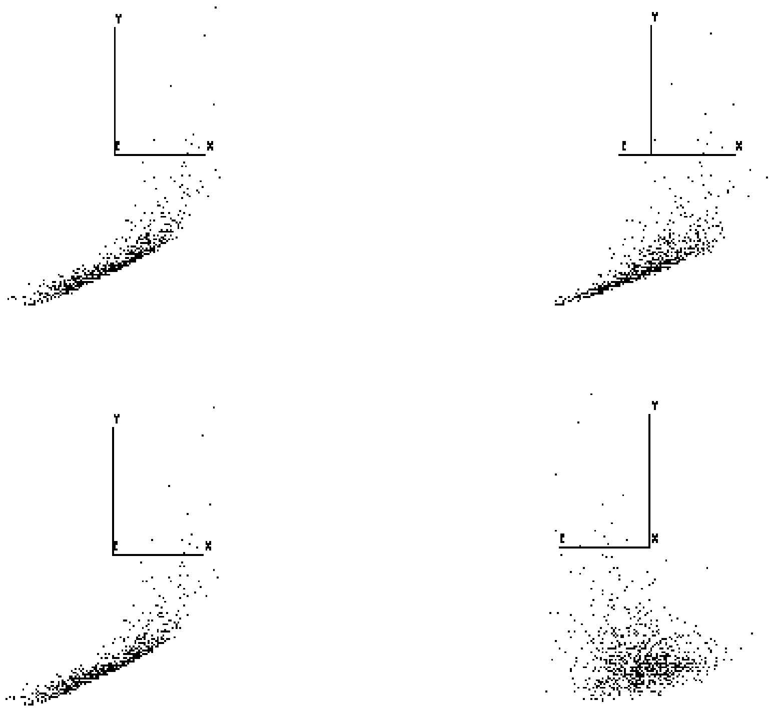


Figure 4.7: SIR’s view for OTL. x and z are the first two directions of SIR.

Chapter 5

Generalization of SIR : Second Moment Based Methods

The basic principle in sliced inverse regression is to reverse the role of Y and $\mathbf{x}$. After partitioning the sample into several slices according to the Y values, summary statistics from each slice can be studied. The sliced inverse regression algorithm discussed earlier combines information from slice means. The larger eigenvectors of SIR are the directions for which the slice means spread out more than other directions, relative to the spread of the original $\mathbf{x}$ along the same directions. We have shown how these eigenvectors can be used to estimate e.d.r. directions.

The covariance matrix is perhaps the next simplest statistic from each slice that deserves our attention. Similar to the first moment method (called SIR-I hereafter), we shall be interested in those directions where the slice covariances vary the most. However unlike SIR-I, there are more than one ways to implement this idea; for example, SAVE (Cook and Weisburg 1991) and SIR-II (Li 1991). Another method, principal Hessian direction (pHd), to be discussed much later, is also closely related.

5.1 A simple symmetric response curve.

Consider the contour plot, Figure 2.6(a)-(c), of $Y = g(x_1+x_2)$ again. We have argued before that the slice means must fall along the e.d.r. direction because points within each slot are symmetrically distributed above and the below the 45^o line. However, suppose $g(t) = t^2$. Then each contour consists of a pair of lines, $x_1 + x_2 = c$ and $x_1 + x_2 = -c$. Thus each slice must have two slots , one on each side of the origin. This additional symmetry forces the average of $\mathbf{x}$ to cancel out eventually within each slice. Thus the slice means are expected to be crowded near the origin and it would be difficult to use principal component analysis to find the right e.d.r. direction.

This argument holds for any symmetric g function. The inverse regression curve is degenerated to a point. This explains why SIR fails to find the e.d.r. direction $\beta = (1, 1)$.

Now let's look at the conditional variance. Along the direction $(1, -1)$ (which is perpendicular to the e.d.r. direction), the variance is the same for each slice. This is easy to see because $(x_1 - x_2)$ is independent of $(x_1 + x_2)$, implying $var(x_1 - x_2|y) = var(x_1 - x_2) = 2$. But along the e.d.r. direction, the conditional variance can vary from slice to slice. For example, suppose $g(t) = t^2$, "conditional on $Y = y$" implies $x_1 + x_2 = \sqrt{y}$ and $-\sqrt{y}$ with probability .5 each. Thus $var(x_1 + x_2|Y = y) = y$, which is see to vary as y changes.

5.2 Slice covariances.

It is helpful to begin with some more examples.

Example 5.2.1. Consider the function $Y = x_1^2 + x_2^2$, whose contours form concentric ellipses about the origin. Suppose the regressor has a spherical distribution. Due to symmetry, it is easy to see that the mean for the regressor between any two contours is always equal to zero. This is another case where the first moment based SIR fails to estimate the e.d.r. space. Strictly speaking, the key theorem for SIR ,Theorem 2.1 in chapter 2, is still true. But by a more careful reading on the statement, one should find that the theorem does not rule out the case that the inverse regression curve may be degenerated. Thus it may happen that the first moment based SIR can only recover a subspace of e.d.r. space.

In this example, the slice covariance matrix is proportional to the identity matrix, I_2. But the proportionality constant varies from slice to slice.

Now suppose we are given one more regressor x_3 which is independent of x_1, x_2. Then since x_3 has nothing to do with Y, the 3-D contours $\{(x_1, x_2, x_3)|Y = c\}$, $c > 0$, are concentric cylinders with a common axis. Again, the slice means of slice remain at the origin. But the slice covariance will take the block form of

$$\begin{pmatrix} c_h I_2 & 0 \\ 0 & d \end{pmatrix}$$

The constant c_h depends on h but $d = var\mathbf{x}_2$ remains the same. This indicates that the directions where the slice covariance changes the most (in this case the first two coordinates) can be good candidates for e.d.r. directions.

Quite often we need both the slice means and slice covariances. This is illustrated in the following example.

Example 5.2.2 Consider $y = x_1^2 + x_2 + 0x_3$. The contours forms parallell parabolic curves on the x_1, x_2 plane. The slice means spread out along the x_2 axis. Thus only the direction $(0, 1, 0)'$ can be found by SIR-I. But the slice covariances take the form

$$\begin{pmatrix} c_h & 0 & 0 \\ 0 & d_h & 0 \\ 0 & 0 & d \end{pmatrix}$$

The direction $(1, 0, 0)'$ can be recovered from the slice covariances.

5.3 Basic properties of slice covariances.

As done before, we shall use capital letter of Y to differentiate from a value y that it may take. For the slice I_h, the covariance of $\mathbf{x}$ will be denoted by $cov(\mathbf{x}|Y \in I_h)$. Note that

$$cov(\mathbf{x}|Y \in I_h) = E(cov(\mathbf{x}|Y)|Y \in I_h) + Cov(E(\mathbf{x}|Y)|Y \in I_h)$$

We shall discuss some properties of $cov(\mathbf{x}|Y = y)$ first. Depending on various constraints on the distribution of $\mathbf{x}$, we have various claims about the relationship between $cov(\mathbf{x}|y)$ and the e.d.r. space. To simplify the discussion, we shall assume that

$$E\mathbf{x} = 0, cov(\mathbf{x}) = I \tag{5.1}$$

As was done before, we can always apply an affine transformation to $\mathbf{x}$ to obtain the standardized version. The interpretation and the application of the results obtained here will be discussed later.

First we assume the **(L.D.C)**.

Theorem 5.1. Assume that the dimension reduction model (1.1) of chapter 1 holds. Under (5.1) and **(L.D.C.)**, for any b in the e.d.r. space and any v in the orthogonal complement of the e.d.r. sapce (that is $b'v = 0$, we have

$$b'cov(\mathbf{x}|y)v = 0 \tag{5.2}$$

The proof of this theorem is almost trivial. Recall the definition of $B = (\beta_1, \cdots, \beta_K)$ from chapter 3. The left side term is equal to

$$\begin{aligned} cov(b'\mathbf{x}, v'\mathbf{x}|y) &= E(b'\mathbf{x}v'\mathbf{x}|y) - E(b'\mathbf{x}|y)E(v'\mathbf{x}|y) \\ &= E(b'\mathbf{x}v'\mathbf{x}|y) \\ &= E(E(b'\mathbf{x}v'\mathbf{x}|B'\mathbf{x}, \epsilon)|y) \\ &= E(b'\mathbf{x}E(v'\mathbf{x}|B'\mathbf{x})|y) \\ &= 0 \end{aligned} \tag{5.3}$$

Here the second equality is due to the fact that $E(v'\mathbf{x}|y) = v'E(\mathbf{x}|y) = 0$, an application of Corollary 2.1 in Chapter 2. The last equality is due to the fact that $E(v'\mathbf{x}|B'\mathbf{x}) = 0$, which is in turn implied by **(L.D.C)** and (5.1).

Next, we assume that

$$\mathbf{x} \textit{ is elliptically symmetric} \tag{5.4}$$

Theorem 5.2. Assume that (5.4) holds. Then under the same setting as in Theorem 5.1, for any v in the orthogonal complement of the e.d.r. space, we have

$$b'cov(\mathbf{x}|y)v = 0, \textit{ for any b orthogonal to v} \tag{5.5}$$

To prove this theorem, we need only to consider those b which is in the orthogonal complement of the e.d.r. space and is orthogonal to v. We still have (5.3). To finish the proof, we observe that

$$E(b'\mathbf{x}v'\mathbf{x}|B'\mathbf{x},\epsilon) = E(b'\mathbf{x}v'\mathbf{x}|B'\mathbf{x}) = 0$$

where the last equality is a simple fact of spherical symmetry of $\mathbf{x}$ (due to (5.1) and (5.4)).

The result of Theorem 5.2. implies that any v in the orthogonal complement of the e.d.r. space must be an eigenvector of $cov(\mathbf{x}|y)$. Consequently, the eigenvalue has to be the same for all v orthognal to The e.d.r. space :

$$cov(\mathbf{x}|y)v = \lambda(y)v$$

where $\lambda(y)$ denotes the common eigenvalue.

Remark 5.1. With b being set to v in (5.3) and the length of v being set to 1, we obtain

$$\lambda(y) = v'cov(\mathbf{x}|y)v = E(E((v'\mathbf{x})^2|B'\mathbf{x})|y)$$

Due to the spherical symmetry, the term $E((v'\mathbf{x})^2|B'\mathbf{x})$ depends only on the length of $B'\mathbf{x}$ and the density of $\mathbf{x}$.

Finally, assume that

$$\mathbf{x} \textit{ is normal.} \tag{5.6}$$

Theorem 5.3. Assume (5.6) holds. Then under the same setting as in Theorem 5.1, for any v in the orthogonal complement of the e.d.r. space, we have

$$cov(\mathbf{x}|y)v = v. \tag{5.7}$$

When $\mathbf{x}$ is normal, $v'\mathbf{x}$ will be independent of $B'\mathbf{x}$. Therefore $E((v'\mathbf{x})^2|B'\mathbf{x})$ is equal to $E(v'\mathbf{x})^2 = 1$. The result of Theorem 5.3 follows from the discussion of Remark 5.1.

5.4 An iterative procedure.

.

From the examples in 5.2, we see that one way to find more e.d.r. directions is to look for directions $\mathbf{b}$ where the conditional variance of $\mathbf{b}'\mathbf{x}$ given Y varies as much as possible. We can translate this into an optimization procedure :

$$\max_{\mathbf{b}} \frac{var(var(b'\mathbf{x}|Y))}{var(b'\mathbf{x}))}$$

This can be carried out iteratively. SAVE as well as the following procedure discussed in the next section can be regarded as some non-iterative variants of this optimization problem.

5.5 SIR II algorithm.

This aims at comparing the slice covriance with the mean slice covariance. Suppose that we have an i.i.d. sample, $(y_i, \mathbf{x}_i), i = 1, \cdots, n$.

(0). Standardize the data as in SIR. That is, standardize **x** by an affine transformation to get $\tilde{\mathbf{x}}_i = \hat{\Sigma}_{\mathbf{x}}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}),\ i = 1, ..., n$, where $\hat{\Sigma}_{\mathbf{x}}$ and $\bar{\mathbf{x}}$ are the sample variance-covariance matrix and sample mean of **x** respectively.

(I) Slice the data in the same way as Step (I) of SIR :

Divide the the range of y into H slices, $I_1, \cdots, I_H$; let the proportion of y_i that falls in slice h be $\hat{p}_h$; that is $\hat{p}_h = \frac{1}{n}\sum_{i=1}^{n} \delta_h(y_i)$, where $\delta_h(y_i)$ takes the values 0 or 1 depending on whether y_i falls into the h^{th} slice I_h or not.

(II) For each slice, compute the sample covariance of **x** :

$$\hat{\Sigma}_h = \frac{1}{n\hat{p}_h}\sum_{i\in I_h}(\tilde{\mathbf{x}}_i - \hat{m}_h)(\tilde{\mathbf{x}}_i - \hat{m}_h)', h = 1, \cdots, H$$

where $\hat{m}_h$ is the sample mean of $\tilde{\mathbf{x}}$ for slice h.

(III) Find the matrix $\hat{\Sigma}_{II}$ defined by

$$\hat{\Sigma}_{II} = \sum_{h=1}^{H} \hat{p}_h(\hat{\Sigma}_h - \bar{\Sigma}.)^2$$

where $\bar{\Sigma}. = \sum_{h=1}^{H} \hat{p}_h \hat{\Sigma}_h$. Then conduct an eigenvalue decomposition of $\hat{\Sigma}_{II}$.

(IV). Let the K largest eigenvectors be $\hat{\eta}_k,\ k = 1, \cdots, K$. Output $\hat{\beta}_k = \hat{\Sigma}_{\mathbf{x}}^{-1/2}\hat{\eta}_k, k = 1, ..., K$.

Examples.

Successful simulation studies have been conducted for several models; for example, $y = sign(\epsilon)(log(|\beta_1\mathbf{x}|) - .75)$ with 300 cases and $p = 10$ (Figures 5.1 and 5.2). For a more complicated model, $y = sign(\beta_2\mathbf{x})(\Phi(|\beta_1\mathbf{x}|) - .5)$, where Φ is normal c.d.f., our second moment based method can only recover the direction of β_1. The second direction, however, is recovered well by further applying the double slicing method to be discussed later.

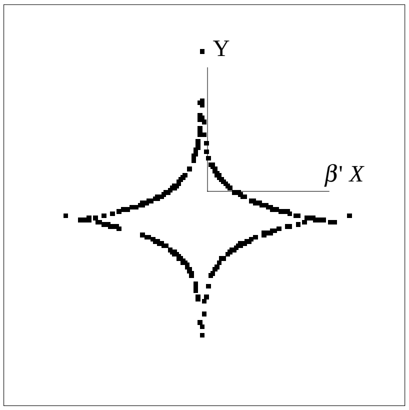


Figure 5.1: Best View for Example 5.1

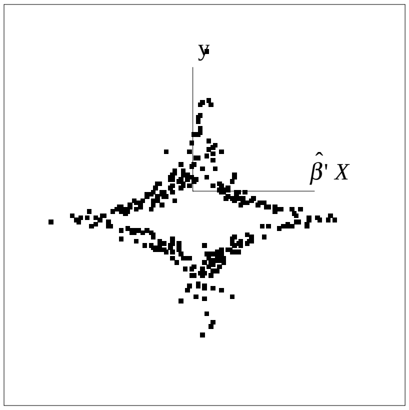


Figure 5.2: SIRII’s View for Example 5.1

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 λ 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 β and project $\mathbf{x}$ on the estimated β 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.

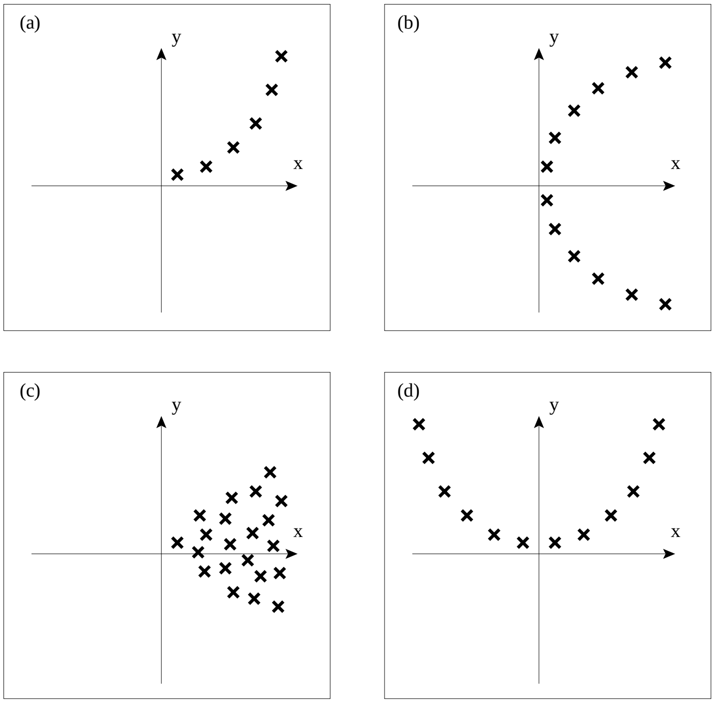


Figure 6.1: Transformation Helps Linearize the Regression for (a), (b), (c), but not (d)

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

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

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

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

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 Σ_η with respect to $\Sigma_{\mathbf{x}}$; i.e.,*

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

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 Σ_η with respect to Σ_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 **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 ith 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 ϵ_1, ϵ_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

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

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 β_2 and β_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 β_2, β_1; see Table 6.2 and Figures 6.3(a)-(b). The curves are now blurred.

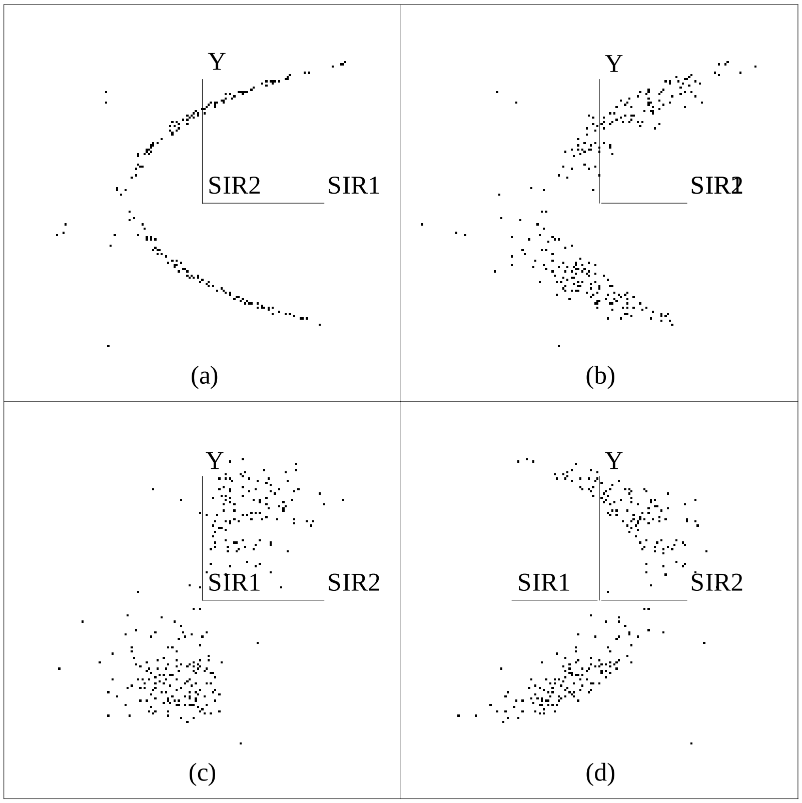


Figure 6.2: SIR's view of Data Generated From (3.1).

Remark 3.1. We also simulated the case with $\alpha = 0$. SIR fails in this case because of the symmetry on the β_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 β_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)

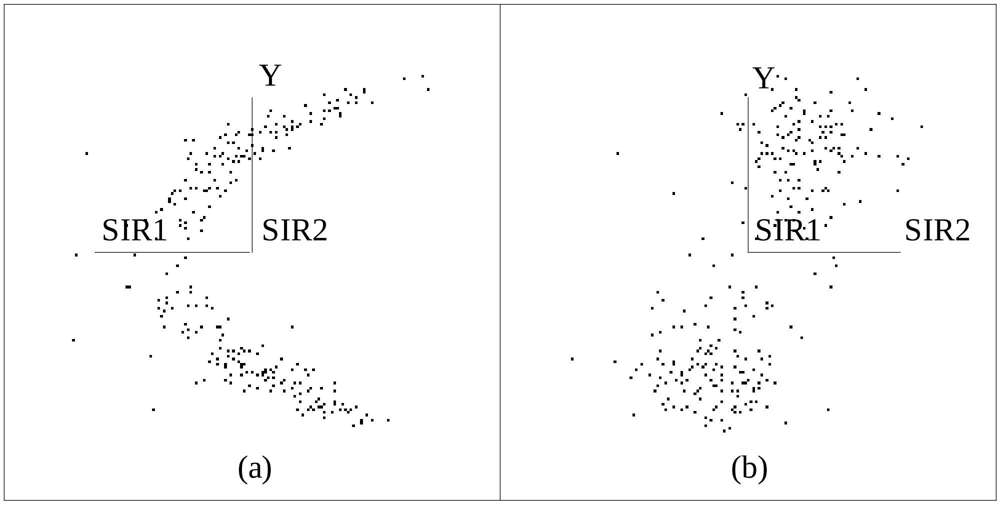


Figure 6.3: SIR's view of Data Generated From (3.1) with $\sigma_1 = \sigma_2 = 1,\ p = 15$

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.

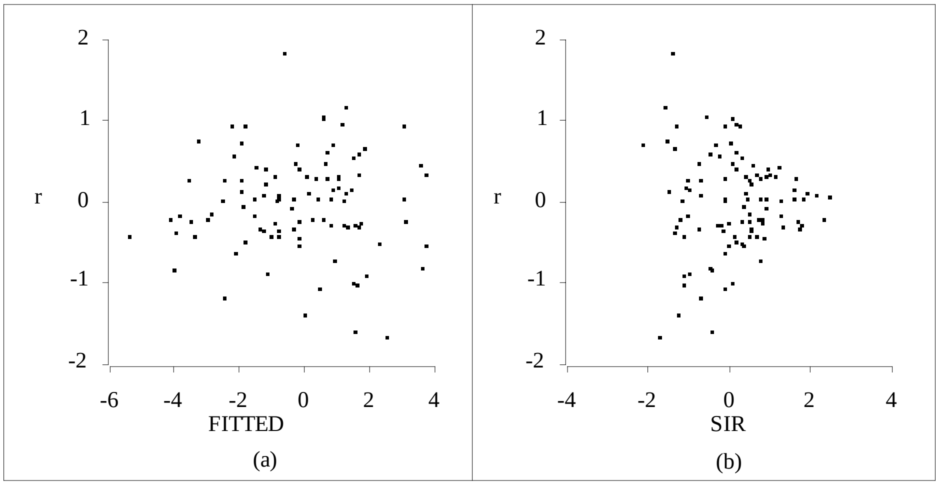


Figure 6.4: Residuals against Linear Squares Fit(a) and Direction(b) of Model(3.2).

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

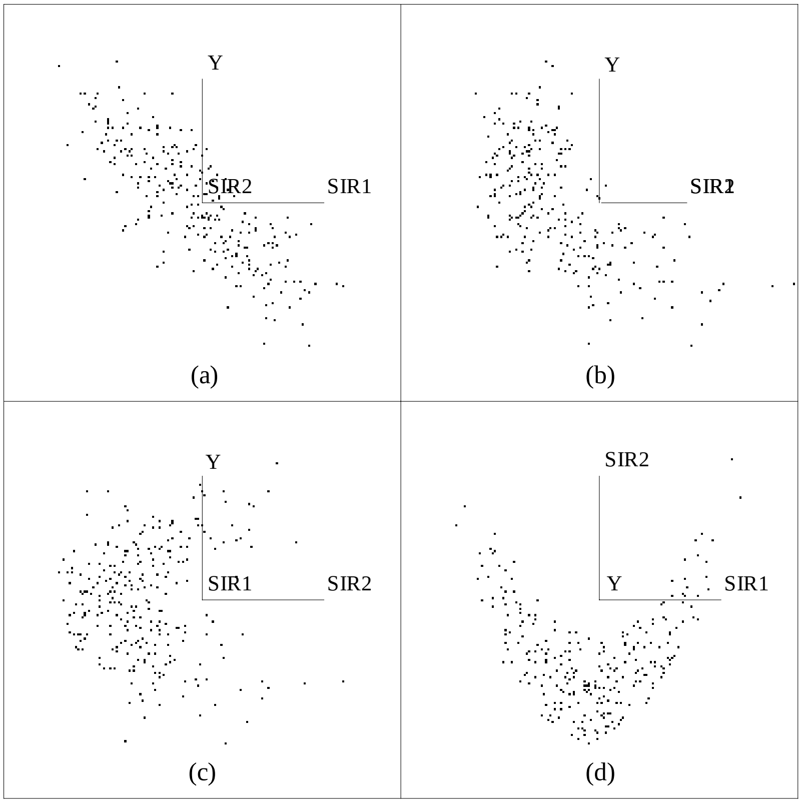


Figure 6.5: SIR's View of Data Generated From (3.3)-(3.4).

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 ith 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 λ_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 λ_i^2. Now apply (4.1) to find out $cov(\tilde{b}(T_i))$. Since the R-squared value of the regression is λ_i as stated in Theorem 3.2, the residual variance σ^2 in (4.1) must be equal to $(1-\lambda_i)var(T_i(y)) = (1-\lambda_i)\lambda_i$. Finally dividing σ^2 by λ_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 lth 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.

Chapter 7

Principal Hessian Directions

The dimension reduction and visualization techniques introduced so far are based on the inverse regression point of view. The roles of Y and $\mathbf{x}$ are interchanged. In this chapter, a forward method, principal Hessian Direction(pHd) (Li 1992, JASA) will be introduced. Let $f(\mathbf{x})$ be the regression function $E(Y|\mathbf{x})$, which is a p dimensional function. Consider the Hessian matrix $H(\mathbf{x})$ of $f(\mathbf{x})$,

$$H(\mathbf{x}) = \text{ the p by p matrix with the } ij^{th} \text{ entry equal to } \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x})$$

Hessian matrices are important in studying multivariate nonlinear functions. The methodology of pHd focuses on the ultilization of the properties of Hessian matrices for dimension reduction and visualization. Similar to SIR, there are a few variants in the approach of pHd. For more recent development on PHD, see Cook(1998).

7.1 Principal Hessian directions.

The Hessian matrix typically varies as $\mathbf{x}$ changes unless the surface is quadratic. Difficulties associated with the curse of dimensionality arise quickly if we were to estimate it for each location. Instead, we turn to the average Hessian,

$$\bar{H} = EH(\mathbf{x})$$

We define the principal Hessian directions to be the eigenvectors $b_1, \cdots, b_p$ of the matrix $\bar{H}\Sigma_{\mathbf{x}}$, where $\Sigma_{\mathbf{x}}$ denotes the covariance matrix of $\mathbf{x}$:

$$\begin{aligned} \bar{H}\Sigma_{\mathbf{x}} b_j = \lambda_j b_j, \qquad & j = 1, \cdots, p \\ & |\lambda_1| \geq \cdots \geq |\lambda_p| \end{aligned} \tag{1.1}$$

Why not defining the principal Hessian directions by the eigenvalue decomposition of the average Hessian $\bar{H}$? One reason is that with right-multiplication of $\Sigma_{\mathbf{x}}$, the procedure becomes invariant under affine transformation of $\mathbf{x}$. This is an important property to have for our purpose of visualization and dimension reduction.

Because of the affine invariance, we may assume that the covariance matrix of $\mathbf{x}$ is I. This often simplifies the discussion.

7.2 Dimension reduction.

Recall the dimension reduction model (1.1) of chapter 1. The regression function takes the form

$$E(Y|\mathbf{x}) = f(\mathbf{x}) = h(\beta_1'\mathbf{x}, \cdots, \beta_K'\mathbf{x}), \tag{2.1}$$

for some function h. Assume that h is twice differentiable.

Lemma 7.3.1 *Under (2.1), the rank of the average Hessian matrix, $\bar{H}$, is at most K. Moreover, the p.h.d.'s with nonzero eigenvalues are in the e.d.r. space $\mathcal{B}$(namely, the space spanned by the β vectors.)*

Proof. Let $\mathbf{B} = (\beta_1, \cdots, \beta_K)$ and $\mathbf{t} = (\beta_1'\mathbf{x}, \cdots, \beta_K'\mathbf{x})' = \mathbf{B}'\mathbf{x}$. Then $f(\mathbf{x}) = h(\mathbf{B}'\mathbf{x})$. To differentiate the Hessian matrix for the function h from that for f, we use the subscripts conveniently. Thus by the chain rule, $H_f(\mathbf{x}) = \mathbf{B}H_h(\mathbf{t})\mathbf{B}'$. Now it is clear that for any direction, v, in the orthogonal complement of $\mathcal{B}$, we have $H_\mathbf{x}(\mathbf{x})v = 0$. Hence the rank of $\bar{H}$ is at most K. In addition, for any p.h.d. b_j with $\lambda_j \neq 0$, we have $0 = (v'\bar{H})\Sigma_\mathbf{x} b_j = v'\lambda_j b_j$, implying that b_j is orthogonal to v. Therefore b_j falls into the e.d.r. space $\mathcal{B}$. The proof is complete. □

This lemma indicates that if we can estimate the average Hessian matrix well, then the associated p.h.d.'s with significant nonzero eigenvalues can be used to find e.d.r. directions. We shall use Stein's lemma to suggest an estimate of the average Hessian matrix in section 7.3.

7.3 Stein's lemma and estimates of the PHD's.

We shall show how to use Stein's lemma to estimate the PHD's when the distribution of **x** is normal.

7.3.1 Stein's lemma.

Recall Stein's lemma from Stein (1981).

Lemma 7.3.1.(Stein 1981, Lemma 4) *If the random variable z is normal, with mean ξ and variance 1, then*

$$E(z-\xi)l(z) = E\dot{l}(z)$$
$$E(z-\xi)^2 l(Z) = El(z) + E\ddot{l}(z)$$

where, in each case, all derivatives involved are assumed to exist in the sense that an indefinite integral of each is the next preceding one, and to have finite expectations.

Proof. The first result is from integration by part. The second ressult follows from the first result. QED. □

Using Stein's lemma, it is easy to derive the following corollary.

Corollary 7. 3.1. Suppose $\mathbf{x}$ is normal with mean $\mu_{\mathbf{x}}$ and the covariance $\Sigma_{\mathbf{x}}$. Let μ_y be the mean of Y. Then the average Hessian matrix $\bar{H}_{\mathbf{x}}$ is related to the weighted covariance

$$\Sigma_{y\mathbf{xx}} = E(Y - \mu_y)(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})'$$

through the identity

$$\bar{H}_{\mathbf{x}} = \Sigma_{\mathbf{x}}^{-1} \Sigma_{y\mathbf{xx}} \Sigma_{\mathbf{x}}^{-1}.$$

Proof. After standardizing $\mathbf{x}$ to have mean 0 and the identity covariance by an affine transformation like $\mathbf{z} = \Sigma_{\mathbf{x}}^{-1/2}(\mathbf{x} - \mu)$, we proceed as if $\mathbf{x}$ is standard normal. Now applying Stein's lemma, we see that

$$\bar{H} = E(Y - \mu_y)\mathbf{xx}'.$$

The proof is complete. □

From this corollary, we can find p.h.d.'s based on the weighted covariance matrix $\Sigma_{y\mathbf{xx}}$ as the following theorem suggests.

Theorem 7.3.1. *When* $\mathbf{x}$ *is normal, the p.h.d.'s,* b_j, $j = 1, \cdots, p$, *can be obtained by the eigenvectors for the eigenvalue decomposition of* $\Sigma_{y\mathbf{xx}}$ *with respect to* $\Sigma_{\mathbf{x}}$ *:*

$$\Sigma_{y\mathbf{xx}} b_j = \lambda_j \Sigma_{\mathbf{x}} b_j, \ \ for \ \ j = 1, \ldots, p.$$

Observe that adding or subtracting a linear function of $\mathbf{x}$ from y does not change the Hessian matrix. Hence instead of using y in Theorem 7.3.1, we may replace it by the residual after the linear least squares fit.

Theorem 7.3.2. Suppose $\mathbf{x}$ is normal. Let $r = y - a - b_{ls}'\mathbf{x}$ be the residual for the linear regression of y on $\mathbf{x}$, where a, b_{ls} are the least squares estimates so that $Er = 0$, and $cov(r, \mathbf{x}) = 0$. Then we have

$$\bar{H}_{\mathbf{x}} = \Sigma_{\mathbf{x}}^{-1} \Sigma_{r\mathbf{xx}} \Sigma_{\mathbf{x}}^{-1},$$

where

$$\Sigma_{r\mathbf{xx}} = Er(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})'$$

Moreover, the p.h.d.'s, b_j, $j = 1, \cdots, p$, can be obtained by the eigenvalue decomposition of $\Sigma_{r\mathbf{xx}}$ with respect to $\Sigma_{\mathbf{x}}$:

$$\Sigma_{r\mathbf{xx}} b_j = \lambda_j \Sigma_{\mathbf{x}} b_j, \ \text{for} \ j = 1, \ldots, p.$$

Corollary 7.3.1 can also be applied to show that polynomial regression can be used to estimate p.h.d.'s, as the following corollary suggests.

Corollary 7.3.2. *Suppose* $\mathbf{x}$ *is normal and consider a polynomial fitting :*

$$\min_{Q(\mathbf{x})} E(y - Q(\mathbf{x}))^2$$

where $Q(\mathbf{x})$ is any polynomial function of $\mathbf{x}$ with degrees no greater than q. Then the average Hessian matrix for the fitted polynomial, is the same as the average Hessian matrix for y, if q is larger than 1.

Proof. Let $\tilde{r}$ be the residual, $y - \tilde{Q}(\mathbf{x})$, where $\tilde{Q}(\mathbf{x})$ is the fitted polynomial. Then $\tilde{r}$ is uncorrelated with any polynomial of $\mathbf{x}$ with degree q or less. In particular, it is uncorrelated with any element in the random matrix $(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})'$. Now we see that

$$\begin{aligned} E(y - \mu_y)(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})' &= E(y - \tilde{r} - \mu_y)(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})' \\ &= E(\tilde{Q}(\mathbf{x}) - E\tilde{Q}(\mathbf{x}))(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})' \end{aligned}$$

Corollary 3.1 implies that the average Hessian matrices for y and $\tilde{Q}(\mathbf{x})$ are the same, completing the proof. □

7.3.2 Estimates for principal Hessian directions.

Theorem 7.3.1 can be used to suggest estimates for p.h.d.'s from an i.i.d. sample, $(y_1, \mathbf{x}_1)$, $\cdots$, $(y_n, \mathbf{x}_n)$. Let $\bar{\mathbf{x}}$ and $\hat{\Sigma}_{\mathbf{x}}$ be the sample mean and the sample covariance of $\mathbf{x}$. Then

(1). Form the matrix $\hat{\Sigma}_{y\mathbf{xx}} = 1/n \sum_{i=1}^{n} (y_i - \bar{y})(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$.

(2). Conduct an eigenvalue decomposition of $\hat{\Sigma}_{y\mathbf{xx}}$ with respect to $\hat{\Sigma}_{\mathbf{x}}$:

$$\begin{aligned} \hat{\Sigma}_{y\mathbf{xx}}\hat{b}_{yj} = \hat{\lambda}_{yj}\hat{\Sigma}_{\mathbf{x}}\hat{b}_{yj}, \qquad & j = 1, \cdots, p \\ & |\hat{\lambda}_{y1}| \geq \cdots \geq |\hat{\lambda}_{yp}|. \end{aligned}$$

Instead of the above $y - based$ method, we may use Theorem 3.2. and suggest the same procedure but with $y_i - \bar{y}$ being replaced by the residual $\hat{r}_i = y_i - \hat{a} - \hat{b}_{ls}'\mathbf{x}_i$, where $\hat{a}, \hat{b}_{ls}$ are the least squares estimates for the linear regression of y against $\mathbf{x}$:

(0). Find the residuals $\hat{r}_i, i = 1, \cdots, n$.

(1). Form the matrix $\hat{\Sigma}_{r\mathbf{xx}} = 1/n \sum_{i=1}^{n} \hat{r}_i(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$.

(2). Conduct the eigenvalue decomposition of $\hat{\Sigma}_{r\mathbf{xx}}$ with respect to $\hat{\Sigma}_{\mathbf{x}}$:

$$\begin{aligned} \hat{\Sigma}_{r\mathbf{xx}}\hat{b}_{rj} = \hat{\lambda}_{rj}\hat{\Sigma}_{\mathbf{x}}\hat{b}_{rj}, \qquad & j = 1, \cdots, p \\ & |\hat{\lambda}_{r1}| \geq \cdots \geq |\hat{\lambda}_{rp}|. \end{aligned}$$

Corollary 7.3.2 suggests yet another way of finding the e.d.r. directions. First fit y by a quadratic polynomial of $\mathbf{x}$. The Hessian Matrix for the fitted quadratic function, say $\hat{B}$, can be easily formed from the estimated quadratic and cross product terms. Then we take the eigenvalue decomposition of the matrix $\hat{B}\hat{\Sigma}_{\mathbf{x}}$ to get the p.h.d.'s. This method (the $q - based$ p.h.d., hereafter) is related with the canonical analysis for exploring and exploiting quadratic response surfaces where the eigenvalue decomposition is taken for the Hessian matrix of the fitted quadratic surface *with respect to the identity matrix*. Box (1954), and Box and Draper (1987), for example, have illustrated well how their techniques have been successfully used to locate stationary points and to obtain a parsimonious description of these points in many designed chemical experiments.

7.4 Sampling properties for normal carriers.

The root-n consistency of the pHd estimates is not hard to establish because our method of moments based estimates of the relevant matrices are clearly root n consistent. We need only to apply standard perturbation formulae for obtaining the asymptotic distributions.

As in the discussion of SIR, the closeness measure between the estimated e.d.r. space, $\hat{\mathcal{B}}_y$ for y based pHd, (respectively, $\hat{\mathcal{B}}_r$), and the true e.d.r. space is given by the squared trace correlation, $R^2(\hat{\mathcal{B}}_y)$ (respectively, $R^2(\hat{\mathcal{B}}_r)$), which is the average of the squared canonical correlation coefficients between $\hat{b}'_{yj}\mathbf{x}$, $j = 1, \cdots, K$, (respectively $\hat{b}'_{rj}\mathbf{x}$, $j = 1, \cdots, K$), and $\beta'_j\mathbf{x}$, $j = 1, \cdots, K$. The closer to one this measure is, the sharper the viewing angle will be. The following theorem gives an approximation for the expected value of this quantity.

Theorem 7.4.1. Assume that $\mathbf{x}$ is normal and that $\Sigma_{y\mathbf{xx}}$ has rank k. Then under the dimension reduction model assumption, we have

$$R^2(\hat{\mathcal{B}}_y) = 1 - (p-k)n^{-1}\sum_{j=1}^{k}(-1 + \lambda_j^{-2} var((y-\mu_y)b'_j(\mathbf{x}-\mu_{\mathbf{x}}))) + o(n^{-1}) \qquad (4.1)$$

and

$$R^2(\hat{\mathcal{B}}_r) = 1 - (p-k)n^{-1}\sum_{j=1}^{k}(-1 + \lambda_j^{-2} var(rb'_j(\mathbf{x}-\mu_{\mathbf{x}}))) + o(n^{-1}) \qquad (4.2)$$

Theorem 7.4.2. *Under the same conditions as in the Theorem 7.4.1, we have*

$$n^{1/2}\textstyle\sum_{j=k+1}^{p}\hat{\lambda}_j \sim N(0, 2(p-k)var(\cdot)) \qquad (4.3)$$

$$n\textstyle\sum_{j=k+1}^{p}\hat{\lambda}_j^2 \sim 2var(\cdot)\chi^2_{(p-k+1)(p-k)/2} \qquad (4.4)$$

where respectively, $\hat{\lambda}_j$ denotes $\hat{\lambda}_{yj}$ or $\hat{\lambda}_{rj}$; $var(\cdot)$ equals $var\ y$ or $var\ r$.

We can use Theorem 4.2 to suggest whether a component found is likely to be real or not, by estimating $var\ y$ (respectively, $var\ r$) with the sample variance of y (respectively, the mean squares for residuals $(n-p)^{-1}\Sigma_{i=1}^n\hat{r}_i^2$).

Remark 4.1. Theorem 4.2 suggests that the residual based estimate is more powerful in detecting a real component because $var\ r$ is typically smaller than $var\ y$. See Cook(1998) for more discussion on some potential inference problems with y-based PHD.

Remark 4.2. For the $q - based$ method, the asymptotic result will be similar. We need only to replace r by the residual of the quadratic fit.

7.5 Linear conditional expectation for x.

The validity of using pHd to estimate e.d.r. directions is justified earlier for the noraml $\mathbf{x}$ via Stein's lemma. Now we like to study how the method behaves under the weaker condition, the **(L.D.C)** used in the theory of SIR : for any $b \in R^p$

$$E(b'\mathbf{x}|\beta_j'\mathbf{x},\ j = 1, \cdots, K) \text{ is linear in } \beta_j'\mathbf{x}\text{'}s \tag{5.1}$$

Theorem 7.5.1. *Under the dimension reduction model assumption and (5.1), the e.d.r. space $\mathcal{B}$ is invariant under the transformation induced by the matrix $\Sigma_{\mathbf{x}}^{-1}\Sigma_{y\mathbf{xx}}$, in the sense that*

$$\{\Sigma_{y\mathbf{xx}}b : b \in \mathcal{B}\} \subseteq \{\Sigma_{\mathbf{x}}b : b \in \mathcal{B}\}$$

Proof. Consider any vector u such that $u'\Sigma_{\mathbf{x}}b = 0$ for any b in $\mathcal{B}$. Then (5.1) implies that $E(u'\mathbf{x}|\beta_j'\mathbf{x},\ j = 1, ..., K) = 0$. It follows that $u'\Sigma_{y\mathbf{xx}}b = E((Y - \mu_y)E(u'\mathbf{x}|\beta_j'\mathbf{x},\ j = 1, ..., K)\mathbf{x}'b) = 0$. This completes the proof. □

Since invariance spaces of a matrix are spanned by its eigenvectors, this theorem suggests that the eigenvectors b_j's can be used to find e.d.r. directions. For instance, if the true e.d.r. space has only one-dimension, $k = 1$, then one of the b_j's must be an e.d.r. direction unless $\Sigma_{y\mathbf{xx}}\beta_1 = 0$, or equivalently,

$$cov(y, (\beta_1'\mathbf{x} - \mu_{\mathbf{x}})^2) = 0. \tag{5.2}$$

Thus although it is not clear which b_j is the right one to use, for the purpose of data visualization we can display all p bivariate plots, y against b_j's, and then choose the one that shows the most interesting structure. But if (5.2) does occur, then we cannot find the e.d.r. direction by this method. Yet we may still hope that some transformation on y might avoid (5.2). Suitably combining second moment SIR estimates is likely to be more productive. Likewise, the case that $k = 2$ leads to viewing $\binom{p}{2}$ sets of three-dimension plots. Some troubles may begin to occur when the dimension of e.d.r. space, k, gets larger because the combination number increases quickly. Note that Theorem 7.5.1 does not promise that large eigenvectors will always be the true e.d.r. directions. But our experience shows that this is indeed very likely to be the case. Pathological cases can exist of course. This is even more transparent for the elliptically symmetric distributions.

Theorem 7.5.2. *Assume that* **x** *follows an elliptically symmetric distribution. Under the dimension reduction model assumption, for the eigenvalues λ_j of the population version of y-based pHd, at least $p - K$ of them take a common value. In addition, all other eigenvectors are e.d.r. directions, if $p - K$ is greater than K.*

Proof. Due to affine invariance, it suffices to consider that case that **x** is spherically symmetric with identity covariance and mean 0. Let P_1 be the projection matrix of rank K with $\mathcal{B}$ as the range space, and $P_2 = I - P_1$. We need only to show that the range of P_2 is a subspace of some eigenspace of $\Sigma_{y\mathbf{xx}}$. First, the result of Theorem 7.5.1 implies that $0 = P_2\Sigma_{y\mathbf{xx}}P_1 = P_1\Sigma_{y\mathbf{xx}}P_2$, or equivalently,

$$\Sigma_{y\mathbf{xx}}P_2 = P_2\Sigma_{y\mathbf{xx}}P_2$$

Fundamental properties from elliptical distributions show that given $P_1\mathbf{x}$ and $\|\mathbf{x}\|^2$, $P_2\mathbf{x}$ is still spherically symmetric with mean 0, and the covariance matrix is $(p-K)^{-1}(\|\mathbf{x}\|^2 - \|P_1\mathbf{x}\|^2)P_2$. From this we see that

$$P_2\Sigma_{y\mathbf{xx}}P_2 = E((y-\mu_y)E(P_2\mathbf{xx}'P_2|P_1\mathbf{x}, \|\mathbf{x}\|^2)) = (p-k)^{-1}[E(y-\mu_y)(\|\mathbf{x}\|^2-\|P_1\mathbf{x}\|^2)]P_2$$

Thus $\Sigma_{y\mathbf{xx}}P_2$ is proportional to P_2, implying that the range space of P_2 is contained in an eigenspace of $\Sigma_{y\mathbf{xx}}$. This proves the theorem. □

This theorem does not say anything about the size of the common eigenvalue,

$$\begin{aligned}(p-k)^{-1}E[(y-\mu_y)(\|\mathbf{x}\|^2 - \|P_1\mathbf{x}\|^2)] &= (p-k)^{-1}E(y-\mu_y)\|P_2\mathbf{x}\|^2 \\ &= cov(E(y|\mathbf{x}), (p-k)^{-1}\|P_2\mathbf{x}\|^2)\end{aligned}$$

But we expect it to be small for most cases. If p is large, $(p-k)^{-1}\|P_2\mathbf{x}\|^2$, becomes nearly independent of $P_1\mathbf{x}$ (unless $||\mathbf{x}||$ is a constant), and hence is expected to be nearly uncorrelated with $E(y|\mathbf{x}) = E(y|P_1\mathbf{x})$. Of course, expections do exist.

It is also clear that our discussion applies to the residual based eigenvectors as defined in Theorem 7.3.2.

7.6 Extension.

Nonlinear transformations of y can be applied before using pHd. For example, we may want to trim out large y values in order to decrease the sensitivity to outliers. We may also use the absolute value of the residual to form the estimate.

We now draw a connection between pHd and second moment based SIR methodology. Let's partition the range of y into H intervals, $I_h, h = 1, ..., H$. Then apply the indicator transformation $\tilde{y} = \delta_h(y) = 1$, or 0, depending on whether y falls into the hth interval or not. Denote $p_h = P\{y \in I_h\}$. Then we have

$$\Sigma_{\tilde{y}\mathbf{xx}} = E(\delta_h(y) - p_h)(\mathbf{x}-\mu_\mathbf{x})(\mathbf{x}-\mu_\mathbf{x})' = p_h[E((\mathbf{x}-\mu_\mathbf{x})(\mathbf{x}-\mu_\mathbf{x})'|y \in I_h) - \Sigma_\mathbf{x}].$$

The y-based pHd theorem can be applied to $\tilde{y}$.

Corollary 7.6.1. *Assume that* $\mathbf{x}$ *is normal. For each slice* h*, conduct the eigenvalue decomposition of the sliced second moment matrix* $E((\mathbf{x}-\mu_\mathbf{x})(\mathbf{x}-\mu_\mathbf{x})'|Y \in I_h)$ *with respect to* $\Sigma_\mathbf{x}$*. Then the eigenvectors with eigenvalues distinct from* 1 *are e.d.r. directions.*

The sample version is easy to obtain. First form the sliced second moment matrix $(n_h - 1)^{-1}\sum_{\mathbf{x}_i \in I_h}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$, where n_h is the number of cases in the hth slice. Then take the eigenvalue decomposition of this matrix with respect to $\hat{\Sigma}_\mathbf{x}$. Let the eigenvalues $\hat{\lambda}_{hj}$'s be arranged to have the order $|\hat{\lambda}_{h1} - 1| \geq \cdots \geq \hat{|}\lambda_{hp} - 1|$.

The sliced second moment matrix $E((\mathbf{x}-\mu_\mathbf{x})(\mathbf{x}-\mu_\mathbf{x})'|y \in I_h)$ discussed above is closely related to the conditional covariance $cov(\mathbf{x}|y \in I_h)$, the core of some specific suggestions for applying second moments in the SIR approach as discussed in Chapter 5. The difference between these two matrices is just a rank-one matrix, $(m_h - \mu_{\mathbf{x}})(m_h - \mu_{\mathbf{x}})'$, where $m_h = E((\mathbf{x} - \mu_{\mathbf{x}})|y \in I_h)$ is the core of the first moment based SIR estimate.

Remark. Limitations. All methods have limitations. SIR and pHd are no exceptions. We shall identify cases that e.d.r. directions cannot be estimated from any transformation version of p.H.d. For simplicity of discussion, take $K = 1$, and concentrate on the case that $E(\mathbf{x}|y) = E\mathbf{x}$, which is the condition to nullify the power of the first moment based SIR. Under this condition, the least squares estimate b_{ls} is equal to 0. Thus the residual-based estimate is the same as the y-based estimate. We are interested in knowing when the weighted covariance matrix $\Sigma_{T(y)\mathbf{xx}} = E(T(y) - ET(y))(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})'$ will be degenerated to 0 for any transformation $T(y)$, in which case no e.d.r. directions can be detected. The following Lemma offers an answer.

Lemma. *Assume* $\mathbf{x}$ *is normal and consider (1.1) of Chapter 1 with* $K = 1$. *Then,*

$$\Sigma_{T(y)\mathbf{xx}} = 0, \textit{for any transformation } T(y),$$

if and only if

$$E[(\beta_1'(\mathbf{x} - \mu_{\mathbf{x}}))^2|Y = y] \textit{ does not depend on } y. \tag{6.3}$$

It is easy to interpret this result from the inverse regression point of view. In general, the conditional distribution of $\beta_1'\mathbf{x}$ given y should depend on y under (1.1). But if this dependence is only through moments of order higher than two, then (6.3) will hold. PHD or any first or second moment based SIR will not find any significant directions. This leaves room for introducing more complicated procedures based on features other than the first two moments of the inverse regression.

7.7 Examples.

Example 7.1. The model used to generate the data is given by

$$y = \cos(2\beta_1'\mathbf{x}) - \cos(\beta_2'\mathbf{x}) + .5\epsilon, \tag{7.1}$$

where $\mathbf{x}$ has $p = 10$ dimensions, $\beta_1 = (1, 0, \cdots)'$, $\beta_2 = (0, 1, 0, \cdots)'$, all coordinates of $\mathbf{x}$ and ϵ are $i.i.d.$ standard normal random variables. For $n = 400$, we study the performance of the residual-based estimate $\hat{b}_{rj}$'s, after 100 simulation runs. A histogram of the closeness measure $R^2(\hat{\mathcal{B}}_r)$ is given in Figure 7.1. The views from the first two directions found in a typical run are given in Figure 7.3, compared with the best views, views from β_1, β_2, given in Figure 7.2. One could better appreciate how they are similar to each other by spinning the two rotation plots and view the data cloud from all angles. Only two directions are found to be significant.

Example 7.2. This example is used to study how violation of the **(L.D.C.)** might affect the estimation. We consider the model

$$y = \beta'\mathbf{x}\sin(2\beta'\mathbf{x}) \tag{7.2}$$

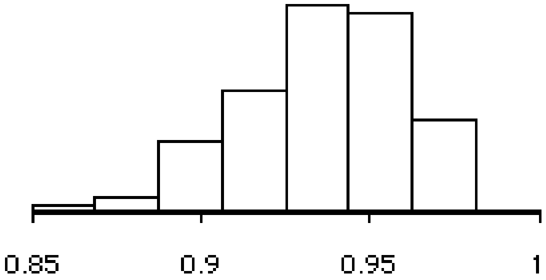


Figure 7.1: Histogram of $R^2(\hat{B})$ for 100 runs.

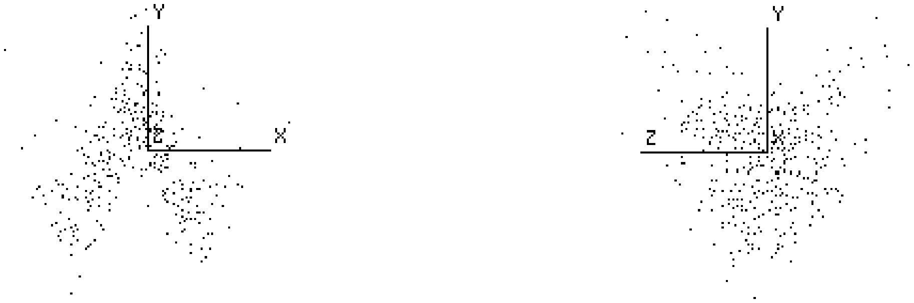


Figure 7.2: Best views of the example 7.1

where $\mathbf{x}$ is uniform on a ten-dimensional cube, $[-1/2, 1/2]^{10}$. First when a direction for β is chosen at random, it is found that the p.H.d. method finds the true direction as well as if $\mathbf{x}$ is indeed normal.

Instead of reporting these favorable cases, we want to study the worst situation. Consider $\beta'\mathbf{x}$ as a sum of p independent random variables and borrow insight from the central limit theorem. We can anticipate the worst case to happen when β is zero on all but two coordinates, the case when $\beta'\mathbf{x}$ is the least normal in a sense. Now for those directions on the plan spanned by first two coordinates, there are 4 good directions for which the linear conditional expectation condition holds; namely the two coordinate axes and the two diagonal lines. Hence we decide to choose $\beta = (1, 2, 0, \cdots)'$ on the ground that this direction is midway between the two good directions $(1, 1, \cdots)'$, and $(0, 1, 0, \cdots)'$. We generate $n = 400$ observations, and use the y-based method to find the e.d.r. direction. From the output given in Table 2, we see some bias in the first direction found. But a close look at the p-values, it is found that the second direction is marginally significant. In fact, a combination of the first two directions, as shown in Figure 7.4 (right), yields a high quality reconstruction of the true curve, shown on the left. By pitching the rotation plots used in producing Figure 7.4 till the y axis is perpendicular to the screen, Figure 7.5 shows how well the distribution for the

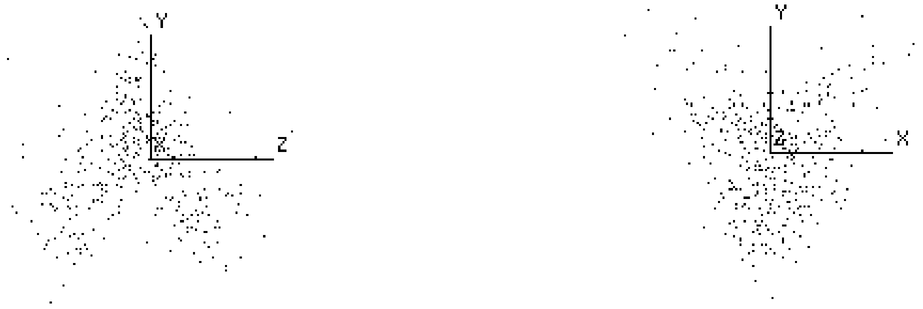


Figure 7.3: Views by the p.h.d. method for Example 7.1

first two projected directions matches the distribution of the first two coordinates of **x**. This demonstrates the potential of our method to find directions b that violate the linear conditional expectation most seriously. One can also argue that under our parameter specification, we can view (4.2) as a two component model with $\beta_1 = (1, 0, \cdots)'$, and $\beta_2 = (0, 1, \cdots)'$. The linear conditional expectation condition is now satisfied, explaining why we can find two directions. Of course, the p-values are only suggestive because of the violation of normality. Judgement based on the pattern of the whole sequence of p-values should be more informative than the individual numbers. We see the drastic increase from .07 to .70 as a strong indication that the third component is not likely to be informative. The residual-based method is also attempted , which yields almost the same result as the one reported here. We conclude this example by reporting that as we enlarge the range of **x** so that the response curve looks more like an M-shape, pHd begins to lose power in detecting the e.d.r. direction. This is because the conditional variance of $\beta'\mathbf{x}$ given y becomes more homogeneous, and Lemma 7.3.1 begins to take effect. It would be interesting to see how well PPR works in such cases.

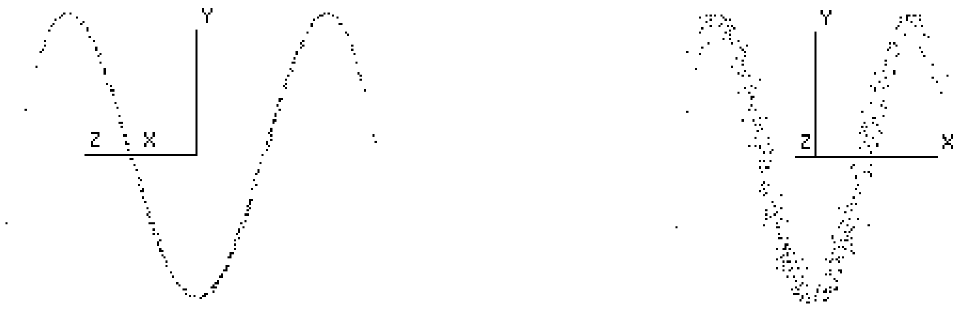


Figure 7.4: Best view(left) and the view by the p.h.d. method for Example 7.2

Example 7.3. This example shows how simple transformations can help p.H.d.. We consider

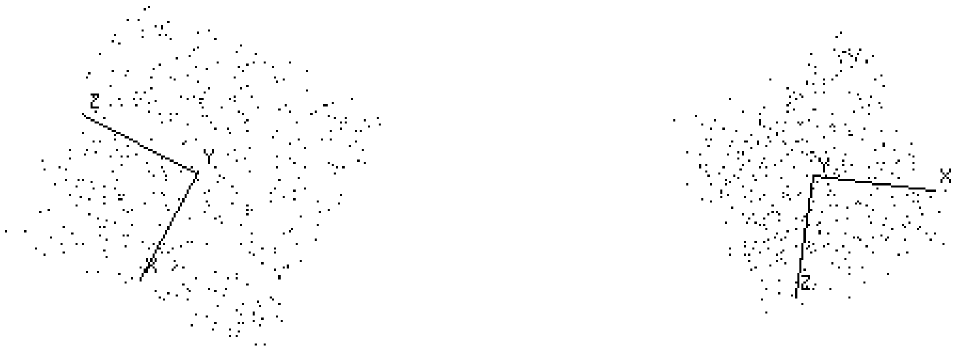


Figure 7.5: Distribution of x_1 and x_2(left), compared with distribution of first two p.h.d.'s

the model

$$y = \frac{1}{3}(\beta_1'\mathbf{x})^3 - (\beta_1'\mathbf{x})(\beta_2'\mathbf{x})^2$$

for generating the data. The surface of this function is known as the monkey saddle. We take $\beta_1 = (1, 0, \cdots)'$, $\beta_2 = (0, 1, 0, \cdots)'$, and generate $n = 300$ data points. First, a histogram of y suggests a long tail distribution. To avoid the dominance of large y in the analysis, we cut out those cases with the absolute value of y greater than 2. This leaves 261 points in the sample. We find the y-based and the residual-based methods unsuccessful, as indicated by the P-values. Then we take the absolute value transformation on the residuals, treat them as y, and proceed with the p.H.d. method. Two directions are found significant. The best views for y and the views based on the estimated directions are given in Figures 7.6, 7.7. Three branches going upward and downward in the monkey saddle can be identified well by spinning these plots on the computer. Other transformations and other methods of handling large y values are worth trying.

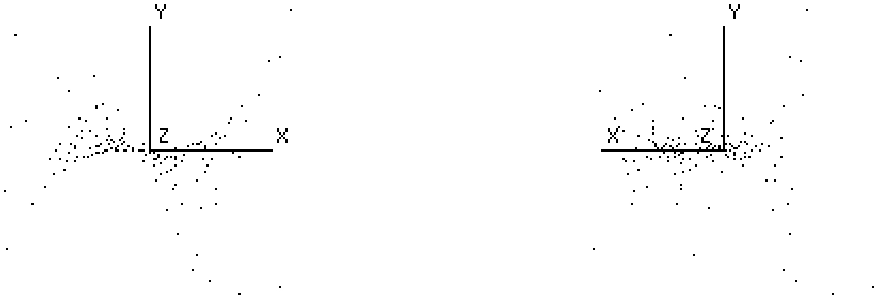


Figure 7.6: Best views of the monkey saddle.

Example 4.4.1 (continued). We continue the analysis of Ozone Data of Example 4.4.1 from chpater 4. Instead of using SIR to study the residuals, we apply by the p.H.d. method,

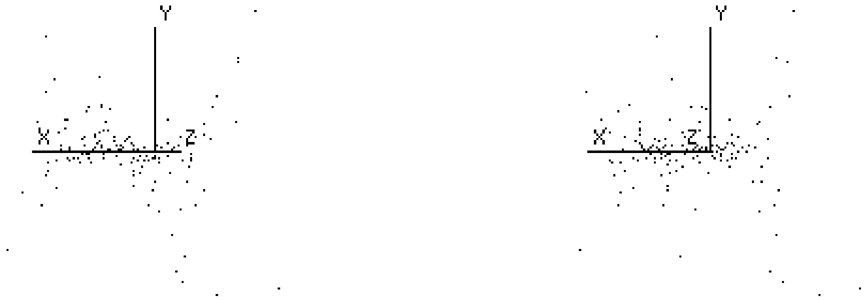


Figure 7.7: Views by the p.h.d. method for monkey saddle.

treating the residual as y. One component is found to be significant. We use a forward selection procedure to find that this component can be explained by x_3, x_5, x_6 with about 90% R-squared (if including x_8 then R-squared can be about 96%). We then run p.H.d. again, using only x_3, x_5, x_6 as the regressors. Again one component is found, denoted as $\hat{b}_{phd}$. Figure 7.8 gives the plot of the residual against this component. A quadratic pattern in this figure is detected by eyes and is confirmed by fitting a quadratic polynomial. The finding here is quite different from SIR plot. This indicates that it may be necessary to find a model that would use the directions from both SIR and PHD directions.

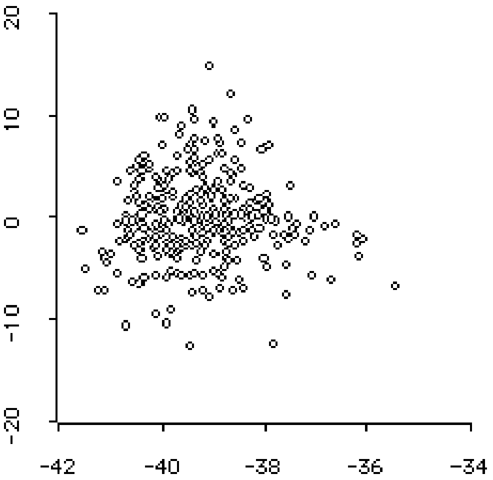


Figure 7.8: Ozone data(continued from Example 4.4.1), Residuals against the direction found by p.h.d.

Chapter 8

Linear Design Condition

As indicated earlier in Chapter 2, the linear design condition **(L.D.C.)** is crucial in the theoretic development of the SIR methodology. It is needed in PHD. Moreover, as to be discussed in later chapters, **(L.D.C.)** is also the key to achieve robustness for essentially all regression methods including the least squares linear regression or GLM estimates.

An important case for **(L.D.C.)** to hold is that the distribution of regressor $\mathbf{x}$ is elliptically symmetric. If data are not yet being collected and the levels of regressors can be set by us, then **(L.D.C.)** may be satisfied through a well planned design of experiment. Even in many observational studies involving survey sampling, if from other sources we have already had in mind some ideas about the distributions of several key regressors in the population, then instead of using representative sampling methods like simple random sampling, we may apply some purposive sampling techniques to induce elliptic symmetry in the selected sample.

Unfortunately, most data sets do not follow elliptically contoured distribution. One way to overcome this difficulty is to force them to behave so. Such ideas will be discussed in Section 8.1.

As mentioned before, elliptic symmetry is not a necessary condition for **(L.D.C.)** to hold. In section 8.2, we shall take a close look at **(L.D.C.)** when the regressor dimension p is large. As it turns out, we can prove a theorem which can be used to argue that for many high dimensional data, the violation of **(L.D.C.)** is negligible. Some practical guidance in dealing with **(L.D..C)** is given in Section 8.3.

8.1 Forcing elliptic symmetry.

8.1.1 A simple case: a 2-D square.

Consider that the regressor $\mathbf{x} = (x_1, x_2)$ is generated uniformly from the unit square $[o, 1]^2$; see Figure 8.1. Then the simplest way to achieve elliptic symmetry is to draw the largest circle inside the square and simply discard the points outside.

This procedure looks simple and effective. But suppose points are in a cube with ,say, ten-dimensional space. Then we can no longer do so effectively because most points will be

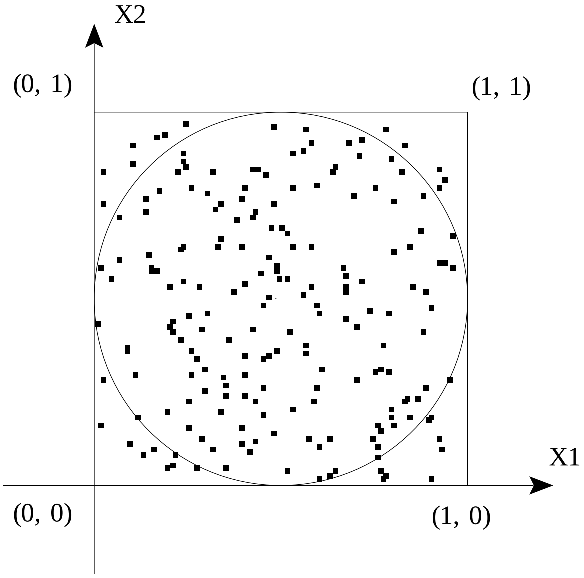


Figure 8.1: Achieving Elliptical Symmetry for Uniform Distribution on a 2D square.

cut out.

8.1.2 Brillinger's normal resampling.

Brillinger(1991) suggests the idea of forcing normality by resampling. To do this, we generate an i.i.d. normal sample of size *n* first. Then we use each simulated point to recruit the closest point in the data as one case in the artificial sample. Thus the constructed sample also has *n* observations, each one coming from the original **x** data. Some points in the original data will be used more than once while others won't be used at all. The resampled **x** is expected to follow closer to a normal distribution than the original one.

Which normal distribution to use ? One suggestion is to match the mean and covariance with the original sample values. But there is a more sophisticate choice given by Cook and Nachtsheim (1992) which will be discussed next.

8.1.3 Minimum volume ellipsoid and Voronoi tesselation.

Let Q_n denote a discrete probability measure that places mass only on the given $\mathbf{x}_i, i = 1, \cdots, n$. Instead of applying equal mass on each point, the idea is to construct an appropriate Q_n so that it is as close to a target elliptically contoured probability distribution, denoted by Q_t, as possible. This involves two problems that can be considered separately. The first one is to choose the target distribution Q_t. This will be discussed later.

The second one, how to construct a discrete Q_n for approximating an specified continuous distribution Q_t, has a natural solution using Dirichlet cells (i.e., a Voronoi tessellation). Let $d(\mathbf{x}_i)$ be the cell consisting of any x in R^p whose Euclidean distance to the point $\mathbf{x}_i$ is shorter than to any other point $\mathbf{x}_j$, $j \neq i$:

$$d(\mathbf{x}_i) = \{x \in R^p | \; ||x - \mathbf{x}_i|| \leq ||x - \mathbf{x}_j||, \, j \neq i\}$$

Then the probability mass on the point $\mathbf{x}_i$ will be set to $\int_{d(\mathbf{x}_i)} dQ_t(x)$. This integral can be obtained by the Monte Carlo method; namely generating a large number of points from the distribution Q_t and then enumerating the relative frequency of falling into each cell.

Choosing an appropriate target distribution Q_t is a harder issue. A simple suggestion is to use a normal distribution with mean and covariance matrix matching $\bar{\mathbf{x}}$, and $\hat{\Sigma}_{\mathbf{x}}$. Yet one limitation is that it won't work well if $\mathbf{x}$ is generated from a distribution with more than one modes. Furthermore, in order for Q_n to converge to Q_t as n tends to the infinity, a minimum requirement is that the support of Q_t be contained in the support of the true distribution that generates $\mathbf{x}$. Thus there are both practical and theoretical needs for using other elliptically symmetric distributions.

The target Q_t suggested by Cook and Nachtsheim is based on the minimum volume ellipsoid (MVE) that trims out a user-specified proportion α of $\mathbf{x}$' points. A MVE is the ellipsoid with the smallest volume that contains $(1 - \alpha) \cdot 100\%$ of the sample points. The target distribution Q_t has the mean and covariance equal to the sample mean and sample covariance for the trimmed data. One needs also to specifies the distribution of the radius, which could be the truncated chi-squares (to reflect the radius of MVE) or simply be the empirical distribution for the radii of the trimmed sample points. The use of MVE has the additional advantage of decreasing the influence of outliers. But the optimal choice of percentage for trimming is an unsettled issue. Cook and Nachtsheim suggest to try 3 or 4 trims ranging from 10 to 50 percent of the data.

Once a discrete Q_n is obtained, we can apply SIR for this Q_n by using the weight assigned by Q_n to each $\mathbf{x}_i$ in carrying out each step of SIR.

In principle, the amount of bias and variance of the SIR estimate depends on the choice of Q_n which in turn is controlled by the trimming percentage α. For a larger value of α, we may expect to have a smaller bias but a greater variance, and vise versa. How to achieve an optimal balance between bias and variance is open for further investigation.

8.2 Higher dimension.

8.2.1 Effectiveness of MVE.

It becomes increasingly hard to force elliptical symmetry as the dimension gets higher.

Example 8.1. Consider a ten-dimensional cube $[-1/2.1/2]^{10}$. The MVE's are balls centered at the origin. If we like to have all 2 D-projection to look good, then we can only retain a very small percentage of points. To get an idea about how bad things can be, we generate 1000 data points. Two 2-D scatter-plots are given in Figures 8.2 (a),(b). It shows square patterns as expected.

Now let's select 500 points that are closest to the origin. These are (approximately) the points inside the MVE with 50 percent trimming. The corresponding 2-D-scatterplots of

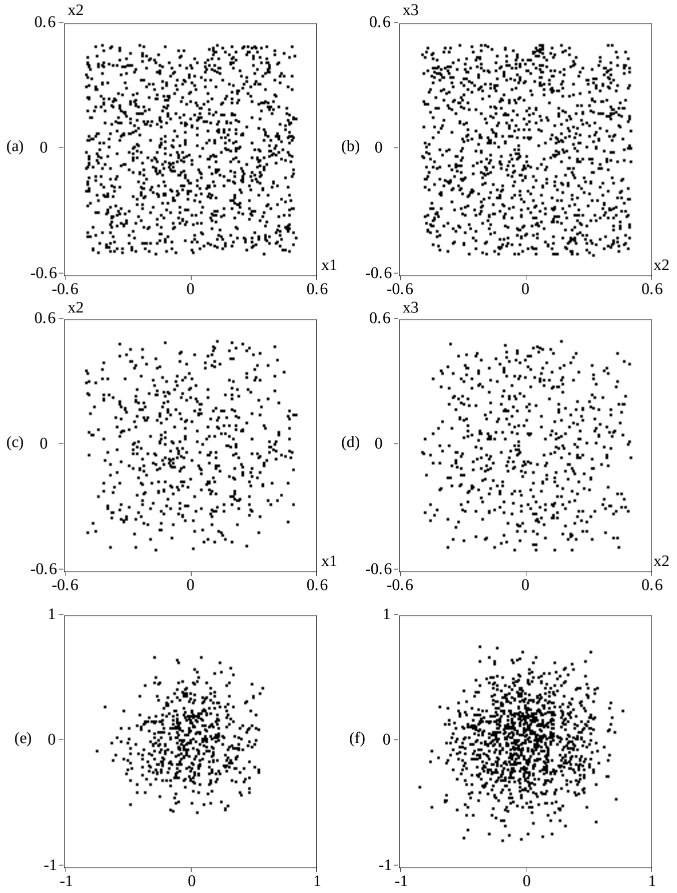


Figure 8.2: How well MVE works in 10 dimensional cube.

these 500 points are shown in Figure 8.2 (c) ,(d). We see that 50% MVE does very little in improving the elliptic symmetry along these low dimensional projections. How about other projections ? Since there are so many projections to look at, let us make a random choice. Two perpendicular random directions are generated:

$$(-0.2260.033 - 0.0950.175 - 0.1540.0370.746 - 0.033 - 0.4930.281)$$

$$(0.0920.278 - 0.267 - 0.0890.2240.225 - 0.269 - 0.654 - 0.2570.286)$$

Figure 8.2(e) gives the projection of the selected 500 data points. It does look like normal now. Does the normality come from the use of the 50% MVE? Not really so! Figure 8.2 (f) shows the 2-D scatterplot of all 1000 data points on the same projection. It is already quite normal.

Example 8.2. Suppose $p = 2, \mathbf{x} = (x_1, x_2)'$. Generate x_1 from a standard normal distribution and generate x_2, given x_1, by $x_2 = x_1^2 + 0.5\epsilon$, where ϵ is another standard normal random variable that is independent of x_1. 100 pairs of the covariate values is shown in Figure 8.3. A useful MVE should trim out the points on the upper portion of either wings.

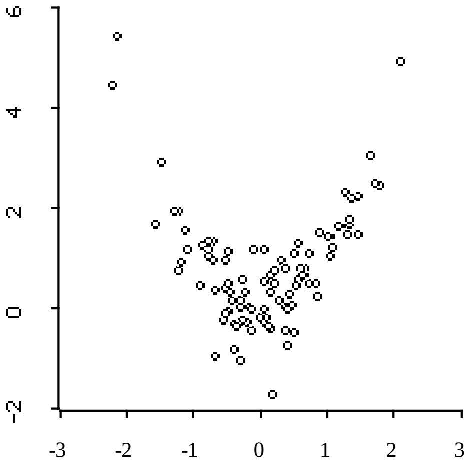


Figure 8.3: Will MVE work for this data?

Now consider the situation that p is much larger, say $p = 10$. Suppose that all other variables $x_3, \cdots, x_{10}$, are all from independent standard normal distributions. The distribution of $\mathbf{x}$ is not elliptically contoured and the projection along the directions , $(1, 0, \cdots, 0)'$ and $(0, 1, 0, \cdots, 0)'$, reveals the most serious departure from elliptical symmetry. This is the only projection we need to improve for elliptical symmetry. In particular, those points $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})'$ with x_{i2} larger than 2.5, for example, had better be trimmed out and the effective way to do so is to apply MVE only after projecting $\mathbf{x}_i$ on the first two coordinates, $(x_{i1}, x_{i2})'$. Without projection, the resulting ten-dimensional MVE is likely to contain many points outside the two-dimensional MVE. This problem becomes more serious as the dimension gets larger, because the Euclidean distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$ will be dominated eventually by the projection on the orthogonal complement of the $x_1 - x_2$ plane:

$$||\mathbf{x}_i - \mathbf{x}_j||^2 = (x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2 \approx (x_{i3} - x_{j3})^2 + \cdots + (x_{ip} - x_{jp})^2$$

Thus it becomes harder to trim out the outliers in the directions where the improvement is essential.

Fortunately, as we will see, there are only very few directions that may violate the (L.D.C). Thus there is still hope that these directions can be detected, and monitored in forcing elliptic symmetry.

8.2.2 Difference between conditional linearity and the elliptic contoured distribution.

The design condition needed in the SIR theorem are satisfied by the elliptic contoured distributions. But it is weaker than elliptic symmetry. We can find many such distributions. For instance, suppose $\mathbf{x}$ is uniform from a cube as in Example 1. If y is related to $\mathbf{x}$ only through

x_1, then the dimension reduction model hold with $k = 1$, $\beta_1 = (1, 0, \cdots, 0)'$. In this case, the design condition holds. Of course, it does not hold if we take $\beta = (1, 2, 0, \cdots, 0)'$ for example. But this is not the direction that generates the data so it does not affect the result of our theorem.

There is a catch here though. We do not know which one is the true direction! A gloomy point of view on this is that even if SIR may be doing the job well, we won't be able appreciate it because we don't know whether **(L.D.C)** does hold or not. This weakness appears rather damaging. But fortunately, as to be discussed in Section 8.2.3. and 8.2.4, there is another point of view which is much brighter.

There is a basic attitude worthwhile to keep before we move on. Like elsewhere, the dimension reduction theory presented in this book is at best an approximation of reality. We cannot expect the conditions assumed in the theory to be met exactly in the real data. A more sensible approach is to pretend they are O.K. in the first place and then conduct a diagnostic checking immediately after SIR/PHD is applied and primary analysis is performed. The effectiveness of any diagnostic checking procedure itself will require certain assumptions as well. This appears to be an endless loop, so we have to draw a cut at some point.

The above argument should not be used to award licenses to poor methods which rely on unrealistic assumptions. ”Realistic or not” is subjective, debatable, and often has to be dealt with on an individual basis. Perhaps it is better to follow the lead of Cook(1998), and replace this phrase by ”scope-limiting or not”. We shall examine how scope-limiting **(L.D.C.)** is by quantifying how often this condition can hold. A somewhat shocking result is that for very large dimension of **x**, **(L.D.C.)** holds with nearly 100 percent chance.

8.2.3 A simulation study.

It would be an ideal situation if **(L.D.c)** can be verified a priori. But this is not easy because it involves the unknown d.r. space which is yet to be estimated. Alternatively, Hall and Li(1993) shows that when the dimension p is large, for most low dimensional $\mathcal{B}$, the **(L.D.C)** can hold approximately. This strongly suggests that the hope is great that SIR may still do well even though we cannot verify **(L.D.C.)** directly. It is already appealing enough to justify the routine use of SIR at the exploratory stage of data analysis. Of course, precautionary measures and retrospective diagnostic checks on **(L.D.C)** should be encouraged as well.

The following simulation study from Li(1991) is helpful in illustrating the theoretical result of Hall and Li(1993).

Example 3.1. Take a rational function as the regression function:

$$y = \frac{\beta_1'\mathbf{x}}{.5 + (\beta_2'\mathbf{x} + 1.5)^2} + .5 \cdot \epsilon, \ \epsilon \sim N(0, 1) \tag{8.1}$$

A typical data pattern when viewed from the rotation plot for y against projections of **x** along the true d.r. directions looks like Figure 2 of Li(1991). A saddle-like surface is visible during the rotation. Suppose that **x** is generated from the uniform distribution on the $p = 10$ dimensional cube $[-\sqrt{3}, \sqrt{3}]^p$. This is not an elliptically contoured distribution. **(L.D.C)** holds only for some directions of β_1, β_2; for example, $\beta_1 = (1, 0, \cdots, 0)'$, $\beta_2 = (0, 1, \cdots, 0)'$.

But for most directions, the violation is not serious enough to induce a large bias. To see this, we first generate $n = 400$ independent $\mathbf{x}_i$'s. Then we repeat the following steps 100 times with these fixed x_i's:

(1). Generate an orthogonal pair of β_1, β_2, uniformly from the unit sphere of R^p.

(2). Generate independent y_i's from model (2.1) with the fixed $\mathbf{x}_i$'s and the β parameters provided by (1).

(3). Apply SIR to find $\hat{b}_1, \hat{b}_2$.

(4). For $i = 1, 2$, compute the closeness measure $R^2(\hat{b}_i)$ between $\hat{b}_i$ and the d.r. space $\mathcal{B}$, spanned by β_1, β_2:

$$\begin{aligned} R^2(\hat{b}_i) &= cos^2(\text{angle between } \hat{b}_i \text{ and } \mathcal{B}) \\ &= \text{R-squared value for regressing } \hat{b}_i'\mathbf{x} \text{ on } \beta_1'\mathbf{x} \text{ and } \beta_2'\mathbf{x} \end{aligned}$$

The histogram of the 100 values of the closeness measure for each direction is given in Figures R.1 and R.2 of Li (1991). From there, we see that most of the time, the R-squares is over .8. The angle of the estimated d.r. direction to the true d.r. space is most often less than $cos^{-1}\sqrt{.8}$, or 26.5 degrees. For comparison, the expected value of the closeness measure between a direction selected at random and the d.r. direction is equal to $2/10 = .2$; $cos^{-1}(\sqrt{.2} = 63.4$ degrees. Since the difficulty in sharpening the angle increases exponentially as the dimension increases, the reduction from 63.4 to 26.5 degrees means that SIR is $(63.4/26.5)^{10} = 245$ times more efficient than a random projection.

8.2.4 Most low dimension projections are nearly linear.

The theoretical foundation behind the above simulation can be found in Hall and Li(1993). Suppose that after normalization, $E\mathbf{x} = 0$, $cov\mathbf{x} = I$, and $B = (\beta_1, \cdots, \beta_k)$ is formed by k orthonormal random vectors from the uniform sphere. Then as the dimension p tends to the infinity, it can be shown that

$$||E(\mathbf{x}|B, B'\mathbf{x} = \mathbf{t})||^2 - ||\mathbf{t}||^2 \rightarrow 0 \tag{8.2}$$

in probability, as p tends to the infinity. The needed regularity conditions are mild. Note that the design condition implies that the left side of the expression is identical to zero. The proof of this result is quite lengthy and will not be presented here.

This result shows that the set of the directions for which the design condition is seriously violated should be small in the high dimension cases.

Earlier related results can be found in Diaconis and Freedman(1984). But our result has a much stronger implication.

8.3 Implication and some guidance.

8.3.1 Blind application.

We may liberally apply SIR/PHD without worrying about **(L.D.C)**. The results from the previous two sections promise that such blind application is not as dangerous as it may sound. But a follow-up diagnostic checking is strongly recommended.

8.3.2 Diagnostic checking.

To do so more effectively, we need to worry about the direction where the violations might be most severe. This is to be discussed next. Once the most dangerous directions are detected, then apply MVE or similar methods to force elliptic symmetry on these directions. Then apply SIR again to see if any change might occur.

8.3.3 The most dangerous directions.

Example 2 shows the need of identifying a small number of projection directions that may depart from elliptic symmetry most seriously first. There are already several projection pursuit density estimation methods for finding these special low dimension projections; see Huber (1985), Friedman(1987), Hall(1989).

But what seems to be more important are those directions that are related to the e.d.r. space; for example, the first few leading directions of SIR. We shall continue this discussion in Chapter 10.

Example 8.2 (continued). We take $p = 5$ and generate 100 cases of $\mathbf{x}_i$. Then generate the response $y = exp(x_1)$ as in Cook and Nachtsheim. Now apply SIR directly. The rotation plot of y against $\hat{b}_1'\mathbf{x}$, $\hat{b}_2'\mathbf{x}$, the first two projections found by SIR with $H = 10$ slices, is given in Figure 8.4(a)(c). Rotating the plot fast enough about the $y - axis$, we find that the data points form a shape that spins like a helix. The scatterplot of $\hat{b}_1'\mathbf{x}$ and $\hat{b}_2'\mathbf{x}$ is given in Figure 8.4(d). It looks similar to Figure 8.3. In fact, both $\hat{b}_1$, $\hat{b}_2$ are quite close to the x_1, x_2 plane. This can be explained by the contour argument.

Insert Figure 8.3(a)-(d) (see page 92 of original notes Figure 7.2(a)-(d) only) *

8.3.4 Bias bound.

Duan and Li(1991, Ann. Stat) gives a bound on the bias that might incur for SIR when the underlying component $k = 1$. Bounds for linear regression can also be obtained. We shall discuss these results later.

Chapter 9

Incorporating Discrete Input Variables

Discrete and continuous variables often call for different treatments in probability and statistics. In our context, this difference is less serious for the output variable. By slicing, we have essentially discretized the output variables anyway (see Chapter 13). But discreteness in the input variables can cause a lot of trouble for us. For continuous variables, linear combinations can be interpreted as projections. But for discrete variables, such kind of geometric interpretation is lost. In earlier chapters, we are mainly concerned with continuous input variables. Now it is time to discuss some ways of handling discrete input variables. The methods discussed here are designed to deal with the case where most of the input variables are still continuous. A different treatment would be called for if all (or most of) the input variables are discrete. For simplicity, we shall often assume that there is only one discrete input variable in the data; all other input variables are continuous. The discrete variable will be denoted by **S** and all other input variables will be put together as a p-dimension vector **x**. It is also useful to differentiate between ordinal (such as the number of cars in a family) and categorical (such as sex, occupation, or treatment) variables. Sometimes an ordinal variable can be treated as a categorical variable. For more detailed discussion, see Carroll and Li(1995)

9.1 Stratification.

Stratification is natural to pursue when **S** is categorical. We simply divide the whole data into several strata according the values of **S**. After that, we simply perform our analysis within each stratum as before. For example, we can run SIR (I or II) to find the best view of data for each stratum and compare how they differ between strata.

The model for dimension reduction (1.1) of chapter 1 can be used for each stratum separately :

$$y = f_s(B_s'\mathbf{x}, \epsilon_s), \ \ for\ \mathbf{S} = s \tag{9.1}$$

where B_s is a p by K_s matrix. Thus each stratum is allowed to have its own e.d.r. space, the space spanned by the columns of B_s with a possibly different dimension K_s. The function f_s and the distribution of ϵ_s may also vary for different strata. The **(L.D.C.)** needed for the consistency of SIR, has to hold for each stratum :

$$E(b'\mathbf{x}|B_s'\mathbf{x}, \mathbf{S} = s) is\ linear\ in\ \ B_s'\mathbf{x}\ \ for\ each\ b \tag{9.2}$$

In the following discussion, we shall concentrate on SIR-I although the general principle applies to the second moment based SIR as well.

9.2 Pooling estimates from different strata.

Important factors for determining the output Y may or may not be the same for different strata. This is why the e.d.r. spaces might be different for different strata. But in this subsection, we shall consider the case where it is reasonable to assume that the e.d.r. directions are the same for all strata; namely each B_s has the same dimension and spans the same column space. However, we still allow different f_s and different distribution for ϵ_s. We would like to pool the results from different strata to yield a better estimate of the e.d.r. space.

There are many ways of pooling the estimates from different strata. The situation is in a sense comparable to the inference for comparing two population means.

In most general situation, the means and the covariance matrices of **x** can vary for different strata. Let the sample mean and the (unbiased) sample covavriance matrix of **x** for stratum s be denoted by $\bar{\mathbf{x}}_s$ and $\hat{\Sigma}_{\mathbf{x},s}$ respectively. Follow steps **1-5** of the SIR algorithm in Chapter 2. Let $\hat{\Sigma}_{\eta,s}$ be the covariance matrix of slice means for stratum s.

If we continue to carry out step **6** separately for each stratum, we would end up with different sets of eigenvectors and we have to glue them together to find the most important K directions.

Our job is similar to that of common eigenvalue decomposition, but not entirely the same. It is related with Homals in Gifi's nonlinear multivariate Analysis (Gifi1989, Wiley). To give some flavor of what is involved, consider the case where $K = 1$. We aim at finding a common direction b such that

$$\hat{\Sigma}_{\eta,s} b = \lambda_s \Sigma_{\mathbf{x},s} b \tag{9.3}$$

for all s, and λ_s is the largest eigenvalue for each eigenvalue decomposition. Of course, this is impossible to achieve exactly. Therefore an approximate solution will be desirable. The real question now is about the sense of approximation. We have to define a reasonable measure to evaluate how much (9.3) is violated. For a general K, the problem gets even more complicated, and more variations are possible.

Another way to look at this problem is to consider the variance of the pooled estimate. The goal here is to make sure that the pooled estimate should have as small variance as possible. This can eventually lead to a sensible solution, although details remain to be carried out. For $K = 1$, the following procedure is not hard to implement : find the vector b that maximizes

$$n^{-1} \sum_s n_s \frac{b' \hat{\Sigma}_{\mathbf{x},s}^{-1/2} \hat{\Sigma}_{\eta,s} \hat{\Sigma}_{\mathbf{x},s}^{-1/2} b}{b' \hat{\Sigma}_{\mathbf{x},s} b}. \tag{9.4}$$

Remark. 9.1. When the number of strata becomes large, we might not find enough cases per stratum to estimate each covariance matrix of $\mathbf{x}$. But if we are willing to assume that this covariance matrix is the same for any stratum, then we can estimate this common covariance matrix by

$$\hat{\Sigma}_{\mathbf{x},c} = (n - S)' \sum_s \sum_i (\mathbf{x}_{is} - \bar{\mathbf{x}}_s)(\mathbf{x}_{is} - \bar{\mathbf{x}}_s)'.$$

Substitute this common covariance matix for each $\Sigma_{\mathbf{x},s}$ in (9.4). To find b, we only need to conduct the eigenvalue decomposition of the matrix $n^{-1} \sum_s n_s \hat{\Sigma}_{\eta,s}$ with respect to $\hat{\Sigma}_{\mathbf{x},c}$. This procedure works for any K.

Remark 9.2. The merit of a complete randomization. When a randomized experiment is conducted, $\mathbf{x}$ is independent of $\mathbf{S}$. Then $cov(\mathbf{x}|S = s)$ does not depend on s. We can combine estimate according to Remark 9.1. Moreover, since $E(\mathbf{x}|S = s)$ is also independent of s, we can even simply run SIR for Y against $(\mathbf{x}, S)$ to get the right directions.

9.3 Estimation of treatment effects.

Our focus in 9.2. is mainly on the variable $\mathbf{x}$. But there are situations where we need to assess the effect of $\mathbf{S}$.

For simplicity, in this section, we shall concentrate on the special case where $\mathbf{S}$ has only two values. We change the notation and use Z to replace $\mathbf{S}$. Despite of the simplicity, this setting arises quite commonly in treatment comparison; $Z = 0$ stands for the control group and $Z = 1$ stands for the treatment group.

Our dimension reduction model can be rewritten as

$$y = g(B'X + \theta Z, \epsilon), \tag{1}$$

where B is a $p - 1$ by k matrix and θ is a k-vector.

This model is quite different from the existing models for treatment comparison in the literature. Let us begin with the case that the dimension of X is one.

1. Models for treatment comparisons: univariate X.

Since we have in total of $p = 2$ regressors, the e.d.r. space can have $k = 0, 1,$ or 2 dimensions. The case $k = 0$ is trial. The case $k = 2$ does not offer any reduction. It does not bring any connection between the control and the treatment groups regarding the relationship between X and Y. We will consider $k = 1$. There are again two cases:

$$Y = g(Z, \epsilon) \tag{1.0}$$

and

$$Y = g(X + \theta Z, \epsilon) \tag{1.1}$$

where ϵ is independent of $(X, Z)'$. The first case implies that in both groups, the X variable is independent of Y. We shall concentrate on (1.1).

The parameter θ in (1.1) can be thought of as the treatment effect. For the control group, Y is related to X through

$$Y = g(X, \epsilon).$$

Without any assumption on g, ϵ. This allows Y and X to have any joint distribution. But for the treatment group, (1.1) gives

$$Y = g(X + \theta, \epsilon)$$

Thus being in the treatment group is equivalent to adding an amount θ in X.

(1.1) covers several interesting special cases of treatment comparison.

(1). Linear model:

$$Y = a + bX + cZ + \epsilon \tag{1.2}$$

For this case, parallel lines are anticipated in the plot of Y on X. The parameter c can be interpreted as the amount of vertical shift needed in order that the line for $z = 0$ to match that for $z = 1$. The matching can also be done by shifting the first line horizontally by the amount $\theta = c/b$.

(2). Two generalizations of the linear models:

$$Y = g(X) + cZ + \epsilon \tag{1.3}$$

$$Y = g(X + \theta Z) + \epsilon \tag{1.4}$$

These represents two different ways of shifting one curve to another. (1.3) is usually studied in the literature concerning partly linear models or partial splines; see Rice, Speckman, Chen, shiao, etc.. The model (1.4) are studied in Haerdle and Marron, Carroll and Hall, Gasser, Kneip, etc.. Note that (1.3) is not covered under (1.1), however.

(3). Multiplicative error and Taguchi's method for industrial Statistical problems:

$$Y = \mu + g(X + \theta Z)\epsilon$$

The mean represents a fixed target, and the uncertainty depends on the regressors. The goal usually is then to reduce the variance. This is a case where all methods mentioned in (2) do not work because the treatment does not effect the mean function at all.

(4.) Generalized linear.

$$Y \sim P_\tau,\ \tau = a + bx + tcz$$

where P_τ from an exponential family indexed by τ. This allows Y to be discrete, such as binomial or Poisson.

2. Identifiability.

Before we present the method for estimating θ, there is an identifiability problem to resolve first. Consider the case that X has a bounded support.

We can always choose a θ large enough in absolute value so that there will be no overlapping bewteen the support of $X + \theta$ with $Z = 1$ and that for X with $Z = 0$. We can construct a g function for each part of the support without restriction. Thus in order to have a meaningful comparison, we should restrict θ to those values for which the support of $X+\theta$ does overlap at least partly with that of X. From now on we shall assume that the value of θ is unambiguously defined to be the smallest (in absolute value) number needed to shift X in the control group so that the conditional distribution of Y given X will be the same as that for the treatment group.

From now on, we shall denote $T_\theta = X + \theta Z$. Note that

$$\begin{aligned} E(Z|T_\theta = t) &= P\{Z = 1|X + \theta Z = t\} \\ &= \frac{f_{X|z=1}(T - \theta)P\{Z = 1\}}{f_{X|Z=0}(t)P\{Z = 0\} + f_{X|z=1}(t-\theta)P\{Z = 1\}} \end{aligned}$$

where all f with subscripts denote conditional densities. It is clear that the design condition cannot hold unless $\theta = 0$ and Z and X are independent.

Note in general the definition of e.d.r. space could be ambiguous sometimes if the design space is not convex. As an example, one may imagine two separate balls on the x_1, x_2 plane, with one y value associated with each ball. Any projection direction on the x_1, x_2 plane that separates the two balls can serve as a one-dimensional e.d.r. space.

3. The method.

One way to estimate θ is to treat it as curve comparison model (1.3) and methods used by others. But this works only when $E(g(X,\epsilon)|X = x)$ is not a constant. Thus it cannot handle the multiplicative error model. A different method will be presented here.

Our method is based on the observation that

$$E(Z|T_\theta = t, Y = y) = E(Z|T_\theta = t) \tag{3.1}$$

(3.1) is equivalent to (1.1). (c.f. Remark 4.1 below)

Lemma 3.1. *Under model (1.1), θ is the unique solution of the minimization problem:*

$$\min_c E(E(Z|T_c, Y) - E(Z|T_c))^2.$$

Remark 3.1. Another equivalence is

$$\text{conditional density of } Y|T_\theta = t, Z, \text{ is equal to conditional density of } Y|T_\theta = t$$

Both are stating the same : Y and Z are independent conditioned on T_θ.

The procedure for implementing Lemma 3.1 can be done simply by kernel estimate of each expectation for example, followed by the search for the minimum of θ. Let

$$\hat{f}_{T,Y}(t, y, \theta) = \frac{(nh_1h_2)^{-1}\sum_{i=1}^n z_i K_1(\frac{y_i - y}{h_1})K_2(\frac{x_i + \theta z_i - t}{h_2})}{\hat{\psi}(y, t)} \tag{3.1}$$

where

$$\hat{\psi}(y,t) = (nh_1h_2)^{-1}\sum_{i=1}^{n} K_1(\frac{y_i - y}{h_1})K_2(\frac{x_i + \theta z_i - t}{h_2}) \tag{3.2}$$

Let

$$\hat{f}_T(t,\theta) = \frac{(nh_1h_2)^{-1}\sum_{i=1}^{n} z_i K_2(\frac{x_i+\theta z_i - t}{h_2})}{(nh_1h_2)^{-1}\sum_{i=1}^{n} K_2(\frac{x_i+\theta z_i - t}{h_2})}$$

which is equal to

$$\int \hat{\psi}(y,t)\,\hat{f}_{T,Y}(y,t,\theta)dy$$

We then estimate θ by minizing

$$\sum_{j=1}^{n}(\hat{f}_{T,Y}(t_j,y_j,\theta) - \hat{f}_T(t_j,\theta))^2$$

If Y is discrete, then the task is simpler. Only univariate kernel is required in (3.1) and (3.2).

4. Multivariate X.

Treatment comparison for high dimensional X rarely goes beyond linear models in the literature. The dimension recuction model we consider here provides new thoughts on how treatment and control groups should be compared.

A quick glance at (1.1) shows that the space spanned by columns of B is the e.d.r. space for both the treatment group and the control group. After we reduce the $p-1$ dimensional X to the k dimensional $W = B'X$, then the treatment effect amounts to shifting W by θ. Unlike the univariate case, this shift involves not only the size but also the direction. Thus for example, the treatment may have an effect equivalent to certain increment in one covariate but not in others.

To study (1.1) more closely, let's rewrite it as

$$Y = g(v'X + Z\theta, B_c'X, \epsilon) \tag{4.1}$$

where B_c is a p by $k-1$ matrix, v is a p-vector with $B_c'v = 0$, and θ is now a real number. The e.d.r. space is spanned by v and columns of B_o for both the treatment and the control groups. The treatment effect amounts to adding θ to the variable $v'x$ while setting $B_c'X$ fixed; unless $v = 0$.

We may look at the problem from another direction. First the e.d.r. space (with smallest dimension) for the control group may be different from that of the treatment group. Then we can write it as

$$Y = g_j(B_j'x, \epsilon_j),\ \ for\ Z = j(= 0,\ or\ 1)$$

In general, B_0 and B_1 can hav different column spaces and even with different dimensions. Now let B_I be a joint and B_u be a union of B_1 and B_2. The case that v is not equal to zero is the special case that $B_1 = B_2$.

We see that the case that v is zero or not has quite different interpretation.

In general, if X satisfies the same condition as $\mathbf{x}$ given before, then we can estimate the e.d.r. space for each group. If the reduced variable has one dimension, then we can apply the method outlined before to W. If W has more than one dimension, but small, then we can still extend the method to estimating the shift parameter θ.

Chapter 10

Quasi-Helices in High Dimensional Regression

It is not uncommon to find nonlinear patterns in the scatterplots of regressors variables. But how the detected nonlinearity affects standard regression analysis remains largely unexplored. The diversity of nonlinear patterns makes it difficult to find a comprehensive approach. This chapter brings out a special nonlinear structure, a helix/spiral/curved slide-like pattern, for study. The importance of quasi-helical structures is addressed by taking a fresh new look on central issues such as the validity of prediction and inference, diagnostics, functional approximation, model uncertainty, Fisher information, robustness, and adaptiveness. A method of finding such quasi-helical objects is proposed. This chapter extends the discussion of **(L.D.C.)**. The material is largely taken from Li(1997).

10.1 Quasi-helical confounding.

Throughout most part of this chapter, we shall take $k = 1$ in the model (1.1) of Chapter 1. We can rewrite it in the form :

$$y = f(\alpha + \beta' \mathbf{x}, \epsilon) \tag{1.1}$$

where y is the output variable and $\mathbf{x}$ is a $p-$dimensional regressor. The ”intercept” term α may seem a bit redundant at this moment because it can be absorbed into f. But we shall see soon why it is needed. As we have mentioned before, various specifications on the structure of f and the distribution of ϵ lead to linear regression, the Box-Cox power transformation model, the generalized linear model with a canonical link, and many others. Once the model assumptions are specified, maximum likelihood estimates of the unknown vector β can be obtained and their sampling properties are well studied.

How does the distribution of $\mathbf{x}$ affects the standard analysis ? In multiple linear regression, this issue has received a good deal of attention. There are largely two areas of study. The first one focuses on the detection of unusual observations (outliers) and assessing the importance of individual cases. The notion of leverage and other regression diagnostics have become popular. They are implemented in almost every regression package. (In nonlinear regression, the study seems just to begin; see Laurent and Cook,1992). The second area is on the linear relationship between input variables. It is well-known that near collinearity inflates the variances of regression coefficients and induces inaccuracy in computing due to rounding errors. Diagnostic methods are available and a rich literature can be found, for example, from the recent article of Stewart(1987) with discussions and the references given there.

We are interested in studying the impact of various nonlinear patterns in **x** on regression analysis. This seems to be a new area. If the regression model holds exactly, then the impact of regressor nonlinearity may amount to only a second order effect. But it could become a dominant one if the model is only approximately correct or even completely unknown.

A comprehensive approach for addressing this issue is hard to pursue. In part, this is due to the diversity of nonlinear structures to be encountered once we depart from linearity . We shall concentrate on one prototype of nonlinearity : a helix/spiral/curved slide- like structure.

Example 1.1 Consider two regressors, $\mathbf{x} = (x_1, x_2)'$, with a strong nonlinear relationship :

$$|x_2 - \log x_1| \leq \delta \tag{1.2}$$

for a small number δ, data generated from the simple linear model

$$y = x_2 + \epsilon \tag{1.3}$$

can also be expressed as

$$y \approx \log x_1 + \epsilon$$

Conversely, data generated from

$$y = \log x_1 + \epsilon \tag{1.4}$$

can be approximated by

$$y \approx x_2 + \epsilon$$

In either case, the 3-D scatterplot of y against x_1 (the x-axis) and x_2 (the z-axis) has almost the same pattern. Data points cling around a 3 dimensional curve which draws a descending arc, bent and twisted about the y axis. When the plot is spin about the y axis in fast motion, it appears like a portion of a helical/spiral/curved-slide object screwing around the computer screen. Figures 10.1(a)-(d) exhibit a couple of static angles of a 3-D rotation for 100 observations generated from model (1.4) with $x_1 \sim uniform[0, 1]$ and $\epsilon \sim Normal(0, 1)$. Here x_2 is generated by adding a uniform random variable between $[-.5, .5]$ to $\log x_2$ so that (1.2) holds with $\delta = .5$. Notice the drastic geometric change from different projections. This never occurs in the collinearity study. Suppose instead of (1.2), x_1 and x_2 satisfies the linear constraint $|x_2 - a - bx_1| \leq \delta$. Then the rotation plot for (1.3) always appears linear, while for (1.4) it always shows a logarithmic curve from any projection. Scale is the only change that takes place in the collinearity study.

We shall use the descriptive word "*quasi-helix*" rather liberally. It refers to *any three-dimensional curved object which appears like a portion of helix/spiral/curved-slide.* The object needs not necessarily cling closely to a cone or cylinder.

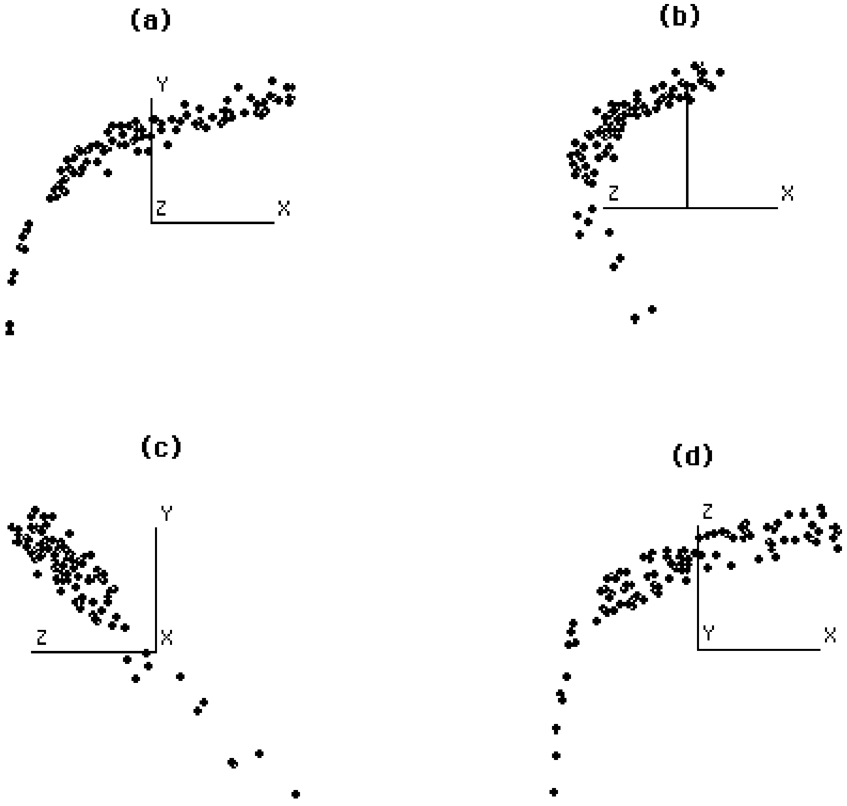


Figure 10.1: How well MVE works in 10 dimensional cube.

The presence of a quasi-helix poses a major difficulty in modeling. In Example 1.1, it is hard to distinguish between the data generated from (1.3) and those from (1.4). A straightforward application of linear models may end up with selecting the wrong variable if (1.4) is indeed the true model. Yet this danger is hard to detect with any available regression diagnostic methods.

A generic terminology "quasi-helical confounding" will be used for referring to the ambiguity in modeling due to the shape change when rotating the plot of y against two curvilinearly associated factors, $\beta'\mathbf{x}$ and $\mathbf{b}'\mathbf{x}$.

In the extreme case that $\beta'\mathbf{x} = h(\mathbf{b}'\mathbf{x})$ for some nonlinear function $h(\cdot)$, model (1.1) can also be written as $y = f(\alpha + h(\mathbf{b}'\mathbf{x}), \epsilon)$ and we cannot tell which direction is correct without knowing the functional form of f. Factor $\beta'\mathbf{x}$ and factor $\mathbf{b}'\mathbf{x}$ are thus totally confounded, a situation similar to (but not the same as) the confounding between interactions (and/or main effects) in the fractional factorial design literature.

Quasi-helical confounding is geometrically different from the confounding due to collinearity. If $\beta'\mathbf{x} = c_1 + c_2\mathbf{b}'\mathbf{x}$, then plotting y against $\mathbf{b}'\mathbf{x}$ is the same as plotting against $\beta'\mathbf{x}$ except for a location shift and a scale change in the horizontal axis. A straight line remains linear and a parabolic pattern remains quadratic, etc.. Because collinearity does not incur the shape change, it does not make the task of specifying the form of f in (1.1) more difficult.

A companion problem associated with model uncertainty is in predicting the future outcome at a given $\mathbf{x} = \mathbf{x}_o$ level. In general, as $\mathbf{x}_o$ gets closer to the existing data points, better accuracy in prediction is anticipated. In multiple linear regression, this is well represented by the Mahalanobis distance which uses the inverse of the sample covariance matrix to define a metric. The center $\mathbf{x}_o = \bar{\mathbf{x}}$, the sample mean of $\mathbf{x}$, is the most accurate cite for prediction. But when we have a curvilinear relationship in $\mathbf{x}$ like (1.2), $\bar{\mathbf{x}}$ may not be even close to the data points. Thus unless the linear model is extremely accurate, it would be very misleading to rely on the Mahalanobis distance for assessing the accuracy of prediction.

Example 1.1 has only two regressors. We can visualize the entire distribution by a single plot. This helps us to detect any pathology, to speculate possible causes and to find ways of remedy. However, it becomes harder to visualize the entire distribution as the regressor dimension gets larger. Scrutiny on plots of y against each of $\binom{p}{2}$ pairs of coordinates is already laborious. Yet there is even more room for quasi-helices to hide in other projections. We need a systematic approach in order to reveal them effectively.

10.2 The κ measure of nonlinearity.

We begin with a bivariate regessor $\mathbf{x} = (x_1, x_2)'$. Consider the following decomposition :

$$\begin{aligned} x_2 &= E(x_2|x_1) + \\ &= l(x_1) + \mathbf{k}(x_1) + e \end{aligned} \tag{2.1}$$

The random part, $e = x_2 - E(x_2|x_1)$, is uncorrelated with the deterministic part $E(x_2|x_1)$. The latter is further decomposed into a linear function $l(x_1)$ and a nonlinear function $\mathbf{k}(x_1)$ in the least squares sense :

$$\begin{aligned} E(E(x_2|x_1) - l(x_1))^2 &= \min_{a,b} E(E(x_2|x_1) - a - bx_1)^2 \\ \mathbf{k}(x_1) &= E(x_2|x_1) - l(x_1) \end{aligned}$$

Definition 2.1. For any two random variables with the decomposition (2.1), define the κ measure of nonlinearity for x_2 on x_1 as

$$\kappa_{x_2|x_1} = \frac{var\ \mathbf{k}(x_1)}{var(x_2 - l(x_1))} = \frac{var\ \mathbf{k}(x_1)}{var\ \mathbf{k}(x_1) + var\ e}$$

The κ measure is undefined for the case that x_2 is a linear function of x_1 because of the zero in the denominator. In linear regression, the terminology “R-square” refers to the proportion of variation in the dependent variable explained by the regression. The R-square of x_2 on x_1 is $var\ l(x_1)/var\ x_2$. Similarly, $\kappa_{x_2|x_1}$ is just the proportion of the variation in the residual $x_2 - l(x_1)$ explained by nonlinear regression on x_1. Thus the κ measure may be considered as “nonlinear R-square”. It reflects only the strict curvillinearity and is free from the influence of collinearity. In particular, we see that for any constants a_0, a_1, a_2,

$$\kappa_{a_2x_2+a_1x_1+a_0|x_1} = \kappa_{x_2|x_1}, \tag{2.2}$$

Unlike R-squares, the κ measure is asymmetric; $\kappa_{x_1|x_2}$ is not the same as $\kappa_{x_2|x_1}$. But if x_1, x_2 are highly correlated, then we can expect $\kappa_{x_1|x_2} \approx \kappa_{x_1|x_2}$. More precisely, given a fixed bivariate random variable $(z_1, z_2)'$, suppose $x_1 = a_1 z_1 + a_2 z_2, x_2 = a_1' z_1 + a_2' z_2$. Then as $\rho(x_2, x_1)$ tends to 1, we have

$$\kappa_{x_1|x_2} = \kappa_{x_2|x_1} + o(1) \tag{2.2'}$$

To see this, let $\tilde{\mathbf{x}}$ be a combination of z_1, z_2, which is uncorrelated with x_1. Then from from (2.2), we can derive $\kappa_{x_2|x_1} = \kappa_{\tilde{\mathbf{x}}|x_1} \approx \kappa_{\tilde{\mathbf{x}}|x_2} = \kappa_{x_1|x_2}$.

We can generalize this measure of nonlinearity to multivariate regressors by considering one-dimensional projections.

Definition 2.2. For any p dimensional random variable $\mathbf{x}$, the κ measure of $\mathbf{x}$ in the direction $\mathbf{b}$ is defined as

$$\kappa_{\mathbf{b}} = \max_{v \in R^p} \kappa_{v'\mathbf{x}|\mathbf{b}'\mathbf{x}}$$

Any direction v that achieves $\kappa_{\mathbf{b}}$ is considered as a most nonlinear direction against $\mathbf{b}$. Because of (2.2), any linear combination of a most nonlinear direction v and $\mathbf{b}$ itself is also a most nonlinear direction for projection. But the degree of linear association varies. For the clearest exhibition of nonlinearity, we may choose a most nonlinear direction from those that are uncorrelated with the direction $\mathbf{b}$:

$$\kappa_{\mathbf{b}} = \max_{\rho(v'\mathbf{x}, \mathbf{b}'\mathbf{x})=0} \kappa_{v'\mathbf{x}|\mathbf{b}'\mathbf{x}} \tag{2.3}$$

There is an easy way to find $\kappa_{\mathbf{b}}$. Let $\mathbf{r}$ be the residual for the linear regression of $\mathbf{x}$ on $\mathbf{b}'\mathbf{x}$:

$$\begin{aligned} \mathbf{r} &= \mathbf{x} - L_{\mathbf{b}}(\mathbf{b}'\mathbf{x}) & (2.4) \\ L_{\mathbf{b}}(\mathbf{b}'\mathbf{x}) &= E\mathbf{x} + (var(\mathbf{b}'\mathbf{x}))^{-1} \cdot \Sigma_{\mathbf{x}}\mathbf{b} \cdot \mathbf{b}'(\mathbf{x} - E\mathbf{x}) & (2.5) \end{aligned}$$

where $\Sigma_{\mathbf{x}} = cov\ \mathbf{x}$. Take the eigenvalue decomposition of the matrix

$$\Sigma_{\mathbf{b}} = cov(E(\mathbf{r}|\mathbf{b}'\mathbf{x})) \tag{2.6}$$

with respect to $\Sigma_{\mathbf{x}}$:

$$\begin{aligned} \Sigma_{\mathbf{b}}\gamma_i = \lambda_i \Sigma_{\mathbf{x}}\gamma_i, \qquad & i = 1, \cdots, p & (2.7) \\ & \lambda_1 \geq \cdots \geq \lambda_p & \end{aligned}$$

Lemma 2.1. *The nonlinear R-square $\kappa_{\mathbf{b}}$ is equal to the largest eigenvalue λ_1 given in (2.7) and the first eigenvector γ_1 is a most nonlinear direction.*

This lemma can be shown from a decomposition analogous to (2.1) :

$$\begin{aligned} \mathbf{x} &= E(\mathbf{x}|\mathbf{b}'\mathbf{x}) + \mathbf{e} & \\ &= L_{\mathbf{b}}(\mathbf{b}'\mathbf{x}) + \mathcal{K}_{\mathbf{b}}(\mathbf{b}'\mathbf{x}) + \mathbf{e} & (2.8) \end{aligned}$$

where the linear part $L_{\mathbf{b}}(\mathbf{b}'\mathbf{x})$ is from linear least squares (2.5) and the nonlinear part $\mathcal{K}_{\mathbf{b}}(\mathbf{b}'\mathbf{x})$ is obtained by substracting the linear part from $E(\mathbf{x}|\mathbf{b}'\mathbf{x})$. It is easy to see that

$$\Sigma_{\mathbf{b}} = cov(\mathcal{K}_{\mathbf{b}}(\mathbf{b}'\mathbf{x})) \tag{2.9}$$

Left-multiplying both sides of (2.8) by v', we get

$$v'\mathbf{x} = v'L_{\mathbf{b}}(\mathbf{b}'\mathbf{x}) + v'\mathcal{K}_{\mathbf{b}}(\mathbf{b}'\mathbf{x}) + v'\mathbf{e} \tag{2.10}$$

This is exactly the same decomposition as (2.1) if we take x_1, x_2 to be $\mathbf{b}'\mathbf{x}$ and $v'\mathbf{x}$ respectively. For any direction v uncorrelated with $\mathbf{b}$, the linear part $v'L_{\mathbf{b}}(\mathbf{b}'\mathbf{x})$ vanishes, implying that

$$\kappa_{v'\mathbf{x}|\mathbf{b}'\mathbf{x}} = \frac{var(v'\mathcal{K}_{\mathbf{b}}(\mathbf{b}'x))}{var\ v'\mathbf{x}} = \frac{v'\Sigma_{\mathbf{b}}v}{v'\Sigma_{\mathbf{x}}v}$$

The maximization problem of (2.3) is thus reduced to the eigenvalue decomposition (2.7), completing the proof of Lemma 2.1. Note that the first eigenvector γ_1 is automatically uncorrelated to $\mathbf{b}$ because $\mathbf{b}$ is in the null space of $\Sigma_{\mathbf{b}}$:

$$\mathbf{b}'\Sigma_{\mathbf{b}}\mathbf{b} = var(E(\mathbf{b}'\mathbf{r}|\mathbf{b}'\mathbf{x})) = var\ 0 = 0$$

10.3 Searching for quasi-helices.

We can easily implement the procedure suggested by Lemma 2.1 for finding the most nonlinear direction against a given direction $\mathbf{b}$. The residual $\mathbf{r}$ and the covariance matrix $\Sigma_{\mathbf{x}}$ can be replaced by their natural sample estimates, $\hat{\mathbf{r}}$, $\hat{\Sigma}_{\mathbf{x}}$. For $\Sigma_{\mathbf{b}}$, a simple estimate can be formed by slicing :

(1) Divide the range of $\mathbf{b}'\mathbf{x}$ into H slices.

(2) For each slice find the sample mean of $\mathbf{r}$.

(3) Find the covariance of the slice means, weighted by the proportion of cases in each slice.

The eigenvalue decomposition (2.1) can be carried out to obtain estimates $\hat{\gamma}_i$, $\hat{\lambda}_i$. If the number of slices is large so that the number of cases in each slice is small, then the variance in estimating the sliced mean becomes sizable and $\hat{\lambda}_i$ tends to over-estimate λ_i. We propose the following rule for modification :

$$\tilde{\lambda}_i = \max\ ((n\hat{\lambda}_i - H)/(n-H), 0)$$

This is motivated from the ANOVA identity used to justify the slice-two method of SIR (Li1991).

Example 3.1 Consider Example 1.1 again and set $u_1 = x_1, u_2 = x_2$. Expand the set of input variables by generating u_3, u_4, u_5, from the standard normal distribution. To make the problem look more difficult, instead of $u_1, \cdots, u_5$, data-analysts are given the following transformed variables as the input variables :

$$x_1 = u_1 + u_3, x_2 = u_2 + u_4 + u_5, x_3 = u_3 - u_4, x_4 = u_4, x_5 = u_5$$

The y variable is the same as in Example 1.1. Thus the quasi-helix is hidden in these two projections :

$$u_1 = (1, 0, -1, 1, 0)\mathbf{x};\ u_2 = (0, 1, 0, -1, -1)\mathbf{x}$$

It would be hard to find it from any scatterplot of coordinate variables. The linear regression gives

$$\hat{\beta}_{ls} = (.667, .756, -.729, -1.49, -.699)'$$

The projection $\hat{\beta}_{ls}'\mathbf{x}$ is shown in the x-axis of Figure 10.2(a). The quasi-helix-searching procedure is carried out with $H = 15$ slices. All but the first eigenvalues are zero and $\tilde{\lambda}_1 = .578$. The most nonlinear direction is $\hat{\gamma}_1 = (-1.36, -.553, 1.41, 1.96, .509)'$. The projection $\hat{\gamma}_1'\mathbf{x}$ is given by the $z-$axis. A quasi-helix is now revealed; see Figure 10.2(a)-(e). Both least squares direction and the most nonlinear direction are highly correlated with u_1 and u_2 :

$$\begin{aligned}\hat{\beta}'\mathbf{x} &\approx\ -.64 + .70u_1 + .75u_2\\ \hat{\gamma}_1'\mathbf{x} &\approx\ 5.59u_1 - 1.30u_2\end{aligned}$$

with a correlation coefficient of more than .99 in each case. The scatterplot of the two approximations is given in Figure 10.2(f), which looks almost identical to Figure 10.2(e).

Example 3.2. We apply our quasi-helix-searching method to the Boston Housing Data(Harrison, D. and Rubinfeld, D.L. 1978; Breiman and Friedman, 1985) for a low-crime rate group which consists of 374 cases. Take y to be the housing price and the remaining 13 variables as the regressor. A quasi-helix is exhibited; Figure 10.3 (a)-(e). We use $H = 15$ slices here but other values also give similar results. The least squares direction $\hat{\beta}_{ls}'\mathbf{x}$ is highly correlated with x_6, a correlation coefficent of about .95; the most nonlinear direction $\hat{\gamma}_1'\mathbf{x}$ against the least squares direction is highly correlated with

$$0.32x_1 + 1.43x_6 + 19.6x_{13}$$

with a correlation coefficent of about 0.94. Figure 10.3(f) shows the scatterplot of these two variables, which looks very similar to Figure 10.3(e). The first eigenvalue is much larger than others :

$$(.44, .08, .05, .03, .00, \cdots)$$

It is interesting to notice that x_6, the number of rooms, is a physical measurement variable, while x_1, x_{13} are the socio-economic variables. Users of multiple regression may unduely down-weight the importance of the socio-economic variable without the aid of the quasi-helix found here.

In the above two examples, the direction $\mathbf{b}$ is taken to be the slope vector $\hat{\beta}_{ls}$ from the linear least squares regression. Our procedure finds the most nonlinear direction $\hat{\gamma}_1$ for supplementing $\hat{\beta}_{ls}$. This yields a three dimensional plot, y against projections along these two directions, for scrutiny. This plot is likely to exhibit a quasi-helix-like pattern if $\tilde{\lambda}_1$ is large.

An immediate use of this plot is in prediction. A well-accepted perception is that interpolation is generally more reliable than extrapolation. But for high dimension regression, it

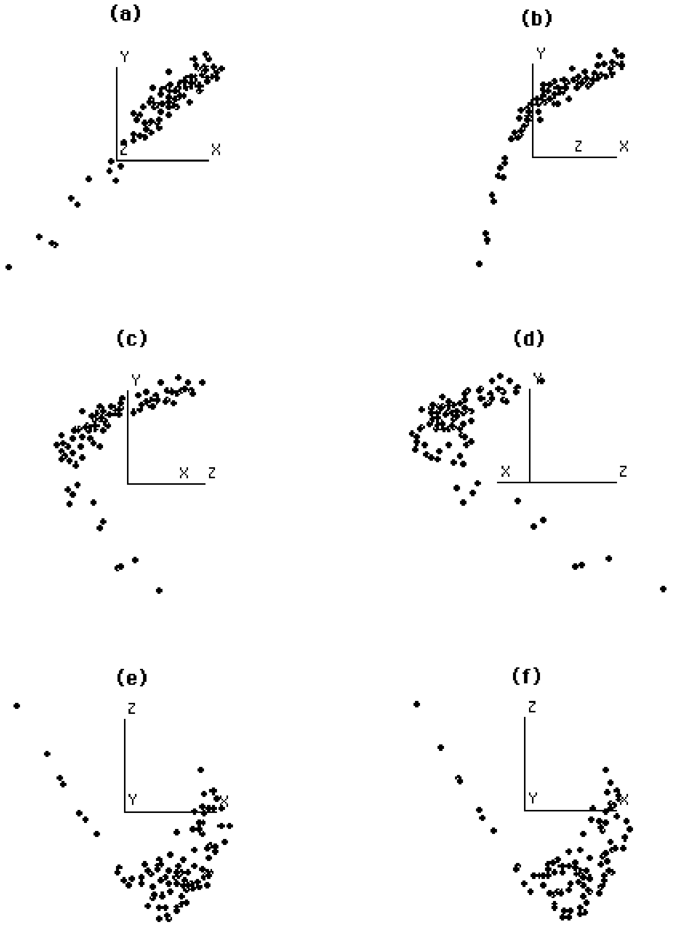


Figure 10.2: Rotation plot for Example 3.1

is not easy to tell if a new case $\mathbf{x}_o$ is inside or outside the data cloud. Standard regression practice is to rely on the Mahalanobis distance. As we have argued before, this practice can be misleading if quasi-helical structures exist. Our three-dimensional plot provides a simple remedy. Just compute $\hat{\beta}_{ls}'\mathbf{x}_o$, $\hat{\gamma}_1'\mathbf{x}_o$ and mark its position in the $x-z$ plane of our plot. We can now visually assess how close it is to the existing data points. If desired, we could even apply simple two-dimensional nonparametric regression methods such as the nearest neighbor or kernel estimates to obtain alternative predictions and compare with the result from the linear model.

Our procedure can also be applied to find a complementary direction for any regression estimate $\hat{\beta}$ for β in (1.1). We need only to take $\mathbf{b} = \hat{\beta}$ and carry out the same procedure as described in the beginning of this section for finding the most nonlinear direction $\hat{\gamma}_1$ for projection. We may also consider the weighted version of Lemma 2.1 because $\hat{\beta}$ is often obtained by iterative weighted least squares.

In the subsequent sections, the issue of quasi-helical confounding will be examined more

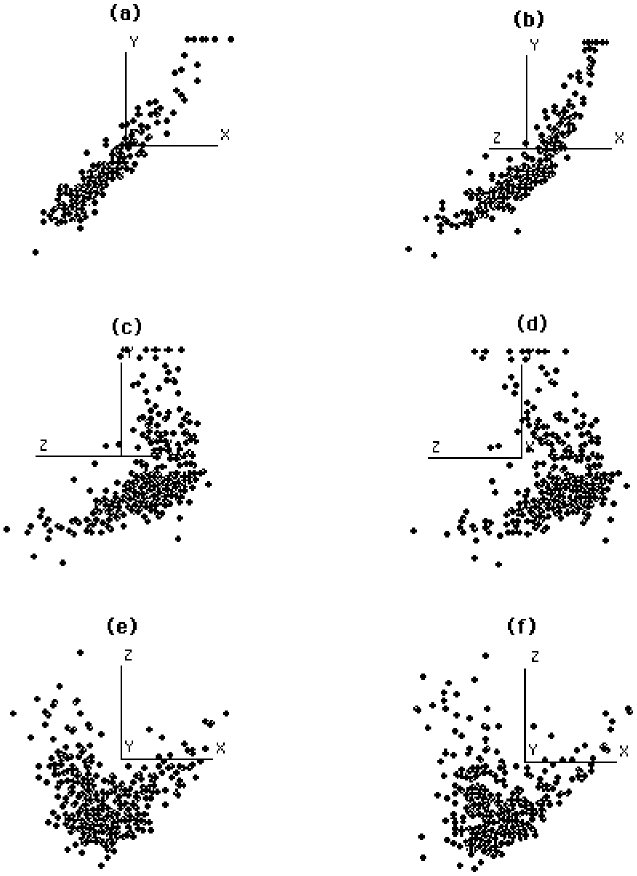


Figure 10.3: Rotation plots for Boston Housing data

closely from several perspectives. The κ measure is found to play a key role in each case.

10.4 Sensitivity of geometric shape change.

As claimed earlier, geometric shape change in the regression function poses a major difference between collinearity and quasi-helical confounding. A closer look at this issue will be taken here.

We begin with a linear regression function $E(y|\mathbf{x}) = a + v'\mathbf{x}$. If $E(y|\mathbf{x})$ is plotted against a different projection $\mathbf{b}'\mathbf{x}$, then from (2.10) we see that in addition to the linear pattern from $v'L_{\mathbf{b}}(\mathbf{b}'\mathbf{x})$, the plot contains a nonlinear trend $v'\mathcal{K}_{\mathbf{b}}(\mathbf{b}'\mathbf{x})$, blurred by the noise term $v'\mathbf{e}$. When the linear trend is removed, the residual becomes

$$v'\mathcal{K}(\mathbf{b}'\mathbf{x}) + v'e$$

We may measure the clarity of the nonlinear trend in the residual plot by the signal to noise ratio

$$\frac{E[\ v'\mathcal{K}(\mathbf{b}'\mathbf{x})]^2}{E\ (v'\mathbf{e})^2}$$

which equals to

$$\frac{\kappa_{v'\mathbf{x}|\mathbf{b}'\mathbf{x}}}{1-\kappa_{v'\mathbf{x}|\mathbf{b}'\mathbf{x}}}$$

Thus the larger the κ measure is, the sharper the change in the shape of the regression function may be seen.

10.5 Over-linearization in linear approximation.

Simplicity is a good reason for the popularity of linear least squares regression. When the true model is linear with $i.i.d.$ normal errors, it offers the best unbiased estimation of the regression coefficients. Yet in practice, linear models are rarely exact. They are often interpreted as a simple approximation to the true regression function. It is commonly believed that if the L_2 norm of the approximation error is small, then the analysis is still valid. Furthermore, we can use residual plots to detect if serious departure from linearity is present or not.

However, very little attention has been paid to the fact that the distribution of **x** affects the sense of linear approximation. As we shall argue in this section, large κ measure can incur a substantial degree of overlinearization, thus weakening the power of residual plots in detecting any lack of fit.

Linear least squares regression approximates a general regression surface $G(\mathbf{x}) = E(y|\mathbf{x})$ in the L_2 sense with respect to the distribution of **x** :

$$\min_{a,\mathbf{b}} E(G(\mathbf{x}) - a - \mathbf{b}'\mathbf{x})^2 \tag{5.1}$$

Denote the least squares solution by a_{ls}, β_{ls} and the approximation error as

$$H_{ls}(\mathbf{x}) = G(\mathbf{x}) - a_{ls} - \beta_{ls}'\mathbf{x}. \tag{5.2}$$

The approximation error in linearizing the nonlinear surface $G(\mathbf{x})$ can be quantified by $EH_{ls}(\mathbf{x})^2$.

The adequacy of L_2 norm in approximation has met some challenges and other norms, e.g., the L_1 norm, have been considered. Our interest is different. We adhere to the L_2 norm but will examine closely the role of the distribution of **x** in the approximation.

Suppose that the regression surface $G(\mathbf{x})$ is indeed only one dimensional:

$$G(\mathbf{x}) = g(\alpha + \beta'\mathbf{x}) \tag{5.3}$$

for some function $g(\cdot)$. Consider the linear approximation for this function :

$$\min_{a,b} E(g(\alpha+u) - a - bu)^2,\ u = \beta'\mathbf{x} \tag{5.4}$$

where the expectation is taken with respect to the distribution of $u = \beta'\mathbf{x}$. Denote the approximation error by

$$h(u) = g(\alpha + u) - a_o - b_o u, \tag{5.5}$$

where (a_o, b_o) is the least square solution of (5.4). In general, β_{ls} is not in the same direction as β. The lack of fit from the least square direction, $EH_{ls}(\mathbf{x})^2$, is smaller than the true approximation $Eh(u)^2$. The plot of y against $\beta_{ls}'\mathbf{x}$ typically appears more linear than the plot of y against $u = \beta'\mathbf{x}$.

Definition 5.1. For a nonlinear function of the form (5.3), the over-linearization of multiple linear regression is defined to be the ratio :

$$OL = \frac{var(h(u)) - var(H_{ls}(\mathbf{x}))}{var(h(u))}$$

The OL measure is strongly affected by the nonlinearity in the regressor. If κ_β is zero, then Brillinger's surprising result, which shows that the least square β_{ls} is proportional to β, implies that $OL = 0$. The relationship between over-linearization and the κ measure in general is explored in the following.

First, fitting (5.4) can be viewed as fitting a submodel of (5.1). Let $\mathbf{b}_r$ be the slope vector for regressing $h(u)$ on $\mathbf{x}$:

$$\min_{a,v} E(h(u) - a - v'\mathbf{x}))^2 \tag{5.6}$$

We can obtain the least squares direction β_{ls} from

$$\beta_{ls} = \mathbf{b}_r + b_o\beta \tag{5.7}$$

The overlinearization measure OL is just the R-square in the regression (5.6) :

$$OL = \frac{var(\mathbf{b}_r'\mathbf{x})}{var(h(u))} = \rho(h(u), \mathbf{b}_r'\mathbf{x})^2 = \max_v \rho(h(u), v'\mathbf{x})^2 \tag{5.8}$$

For any direction v, consider the decomposition (2.10) with $\mathbf{b} = \beta$. Since $h(u)$ is uncorrelated with $u = \beta'\mathbf{x}$, the maximization in (5.8) can be restricted to those directions v that are uncorrelated with β. The first term in (2.10) thus vanishes. Consequently we have

$$\begin{aligned}\rho(h(u), v'\mathbf{x})^2 &= \frac{cov(h(u), v'\mathbf{x})^2}{var(h(u)) \cdot var(v'\mathbf{x})} \\ &= \frac{cov(h(u), v'\mathcal{K}_\beta(u))^2}{var(h(u)) \cdot var(v'\mathbf{x})} \\ &= \rho(h(u), v'\mathcal{K}_\beta(u))^2 \cdot \kappa_{v'\mathbf{x}|\beta'\mathbf{x}} \end{aligned} \tag{5.9}$$

where the second identity is due to the decomposition (2.10).

Now from (5.7), (5.8), (5.9), and (2.2), we obtain the key identity

$$\begin{aligned} OL &= \max_v \ \rho(h(u), v'\mathcal{K}_\beta(u))^2 \cdot \kappa_{v'\mathbf{x}|\beta'\mathbf{x}} \\ &= \rho(h(u), \beta_{ls}'\mathcal{K}_\beta(u))^2 \cdot \kappa_{\beta_{ls}'\mathbf{x}|\beta'\mathbf{x}} \end{aligned} \tag{5.10}$$

Dropping the correlation coefficient term, we obtain an upper bound for OL. This is given in the following theorem.

Theorem 5.1. *For a nonlinear regression function of the form (5.3), the over-linearization of multiple linear regression is bounded by the κ measure of nonlinearity in regressor:*

$$OL \le \kappa_{\beta_{ls}'\mathbf{x}|\beta'\mathbf{x}} \le \kappa_\beta$$

This theorem shows how the curvillinearity between regressors can affect the amount of over-linearization. Over-linearization may not be serious unless there is a strong curvillinearity among regressors. The result of Hall and Li(1993) assures that in many cases the κ measure is small. This helps in explaining why residual plots in multiple regression are often useful. But the other side of the coin is also important to bear in mind. If OL is large, then not much signal will be left in the residuals for detecting the lack of fit. The performance of residual plots can be seriously impaired by quasi-helical confounding. We illustrate this with the setting of Example 3.1.

Example 5.1. We generate the data again from the model considered in Example 3.1, but with a larger noise level, $\sigma = .5$. Figure 10.4(a) shows the standard residual plot for linear regression. No obvious nonlinear pattern can be found from this plot. By comparison, the residual plot from regressing y against the correct projection u_1, Figure 10.4(e), shows a clear nonlinear trend. Multiple linear regression has over-linearized the true regression function to an extent that the residual plot has lost power of detecting the nonlinearity. We now carry out our quasi-helix-hunting procedure as done in Example 3.1. The first eigenvalue is large, about .45, indicating a serious quasi-helical confounding. Figures 10.4(b)-(d) show some angles from the rotation plot of y against $\hat{\beta}_{ls}'\mathbf{x}$ and $\hat{\gamma}_1'\mathbf{x}$. Figure 10.4(f) is the residual plot of regressing y linearly against a special angle, $.27\hat{\beta}_{ls}'\mathbf{x} + .15\hat{\gamma}_1'\mathbf{x}$. This is the angle that has the highest correlation, about .99, with u_1. This plot is almost identical to Figure 10.4 (e).

We turn to a related question: how close is the least squares direction β_{ls} to the true direction β ? We measure the closeness by the square of the correlation coefficient between $\beta'\mathbf{x}$ and $\beta_{ls}'\mathbf{x}$. The quantity $1-\rho(\beta'\mathbf{x}, \beta_{ls}'\mathbf{x})^2$ shows the size of bias of the regression estimate β_{ls}.

Theorem 5.2 *(maximum bias for least squares). For any $\delta > o$, consider the class $\mathcal{F}_\delta$ of regression functions with the form (5.3), such that the proportion of the variance of the approximation error, $var(h(u))/var(g(\alpha+u))$ does not exceed δ. Then the maximum of the bias in the linear least square estimate is related to the nonlinear R-square of the regressors :*

$$\max_{G(\cdot)\in\mathcal{F}_\delta} 1 - \rho(\beta_{ls}'\mathbf{x}, \beta'\mathbf{x})^2 = \frac{\delta\kappa_\beta}{1-\delta(1-\kappa_\beta)}$$

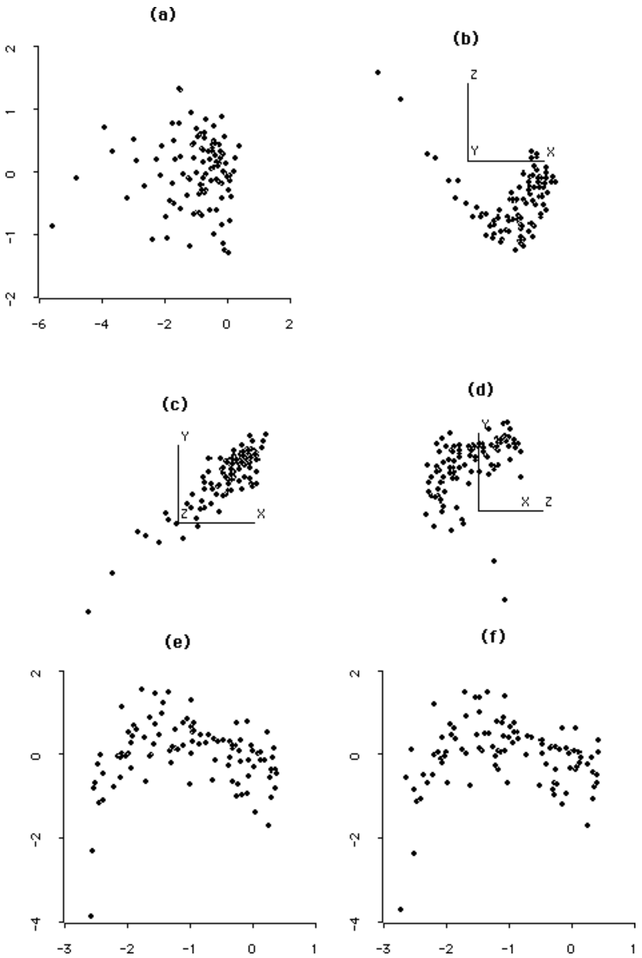


Figure 10.4: Residual plots and helix.

Denote $\rho = \rho(\beta'\mathbf{x})\beta_{ls}'\mathbf{x}$. We can also relate OL, ρ to the R-square of the multiple linear regression, $R^2 = var(\beta_{ls}'\mathbf{x})/var\ y$:

$$OL = \frac{(1-\rho^2)R^2}{(1-\rho^2R^2-\frac{var(\epsilon)}{var(y)})} \geq \frac{(1-\rho^2)R^2}{1-\rho^2R^2}$$

$$\rho^2 = \frac{R^2 - OL(1-\frac{var(\epsilon)}{var(y)})}{R^2(1-OL)} \geq \frac{R^2-OL}{R^2(1-OL)} \geq \frac{R^2-\kappa_\beta}{R^2(1-\kappa_\beta)}$$

The same lower bound for ρ^2 is also derived in Duan and Li(1991) which uses a different argument without the overlinearization measure.

10.6 Over-fit in nonlinear approximation.

In many applications, nonlinear functions are also used for approximation. Sigmoidal functions such as the logistic $G_o(t) = (1 + e^{-t})^{-1}$, for example, are popular in neural network modeling (e.g., White 1989). Lik in linear regression, overfitting can be severe when there is κ mesure is large unless the nonlinear model is exact. For better understanding, we need a weighted version of the κ measure. This will not be discussed here.

10.7 Model uncertainty and information loss

. Linear regression estimate is efficient when the true regression function is linear and the errors are i.i.d. normal. The Fisher information matrix for the slope vector β per observation is equal to $\mathcal{I} = \sigma^{-2}\Sigma_{\mathbf{x}}$ and the covariance matrix of $\hat{\beta}_{ls}$ is $n^{-1}\mathcal{I}^{-1}$. If the true function is only approximately linear, how much information will be lost in estimating β? To answer this question, we may parametrize any reasonable patterns of departure from linearity, treat the additional parameters as the nuisance parameters, and then eliminate their influence when finding the information matrices for β. A least favorable submodel is the one which yields the smallest information $\mathcal{I}_{min}$. The information loss, $\mathcal{I} - \mathcal{I}_{min}$, due to the model uncertainty will be examined in this section. As it turns out, a major source of the loss comes from nonlinearity in the regressor. If the distribution of **x** is elliptically symmetric, there will be no loss of information in estimating β up to a proportionality constant. On the other hand, the loss may be grave if a serious quasi-helical confounding condition is present. The consideration of least favorable submodels is essential in the semiparametrics and adaptiveness studies; see Bickel et al (1993).

10.7.1 Least favorable submodel.

Model uncertainty can be addressed in a more general context to cover nonlinear regression as well. Consider

$$y = g(\alpha + \beta'\mathbf{x}) + \epsilon \tag{7.1}$$

where ϵ is normal with mean 0 and variance σ^2. We will construct a least favorable submodel in the neighborhood of a given function $g_o(\cdot)$ for any direction β_o.

Define the function

$$\eta(u) = E(\mathbf{x}|\beta_o'\mathbf{x} = u) \tag{7.2}$$

Then our least favorable submodel is :

$$y = g_o(\alpha + \beta'\mathbf{x} - \delta'\eta(\beta'\mathbf{x})) + \epsilon \tag{7.3}$$

with δ being a p-dimensional nuisance parameter. When $\delta = 0$, we are reduced to (7.1) with $g(\cdot) = g_o(\cdot)$.

Generically for any parametric family of distributions with density functions $\{f(y; \beta, \delta)\}$, the Fisher information for the parameter of interest β, in presence of the nuisance parameter

δ, can be founded by first computing the Fisher scores,

$$S_\beta = \frac{\partial log f(y;\beta,\delta)}{\partial\beta};\ S_\delta = \frac{\partial log f(y;\beta,\delta)}{\partial\delta}$$

S_β and S_δ are correlated in general. Elimination of the influence of the nuisance score can be done by orthogonalization. A matrix M can be found so that $S = S_\beta - MS_\delta$ is uncorrelated with S_δ. In other words, S is the residual for regressing S_β on S_δ linearly. After eliminating the nuisance parameter, the Fisher's information for β is just the covariance matrix of S.

Following the above recipe and treating both α and δ as nuisance parameters in model (7.3), at $\alpha = \alpha_o, \beta = \beta_o, \delta = 0$, we find

$$\begin{aligned}
S_\alpha &= \sigma^{-2}\epsilon\dot{g}_o(\alpha_o + \beta_o'\mathbf{x}) \\
S_\beta &= \sigma^{-2}\epsilon\dot{g}_o(\alpha_o + \beta_o'\mathbf{x})\mathbf{x} \\
S_\delta &= \sigma^{-2}\epsilon\dot{g}_o(\alpha_o + \beta_o'\mathbf{x})\eta(\beta_o'\mathbf{x}) \\
S &= \sigma^{-2}\epsilon\dot{g}_o(\alpha_o + \beta_o'\mathbf{x})(\mathbf{x} - \eta(\beta_o'\mathbf{x}))
\end{aligned}$$

By conditioning, it is easy to see that S is uncorrelated with S_α, S_δ :

$$\begin{aligned}
cov(S, S_\alpha) &= \sigma^{-2}E[\dot{g}_o(\alpha_o + \beta_o'\mathbf{x})^2 E(\mathbf{x} - \eta(\beta_o'\mathbf{x})|\beta_o'\mathbf{x})] = 0 \\
cov(S, S_\delta) &= \sigma^{-2}E[\dot{g}_o(\alpha_o + \beta_o'\mathbf{x})^2 \eta(\beta_o'\mathbf{x}) E(\mathbf{x} - \eta(\beta_o'\mathbf{x})|\beta_o'\mathbf{x})] = 0
\end{aligned}$$

Theorem 7.1. *For the parametric model (7.2), the Fisher information of β per observation is equal to*

$$\mathcal{I}_{min} = \sigma^{-2}E\dot{g}_o(\alpha_o + \beta_o'\mathbf{x})^2(\mathbf{x} - \eta(\beta_o'\mathbf{x}))(\mathbf{x} - \eta(\beta_o'\mathbf{x}))' \tag{7.4}$$

This Fisher information is no greater than the information matrix for β derived from any parametric model

$$y = g(\alpha + \beta'\mathbf{x}; \tilde{\delta}) + \epsilon \tag{7.5}$$

at $\beta = \beta_o, \tilde{\delta} = 0, \alpha = \alpha_o$, where $\tilde{\delta}$ is a finite dimensional nuisance parameter with $g(\alpha_o + \beta_o'\mathbf{x}; 0) = g_o(\alpha_o + \beta_o'\mathbf{x})$

We briefly describe how our least favorable submodel is found. Consider a direction β in a small neighborhood of β_o. The principle of least favorable model is to find a g function so that $g(\alpha + \beta'\mathbf{x})$ can be as close to $g_o(\alpha + \beta_o'\mathbf{x})$ as possible because this would make the task of discriminating β from β_o the most difficult. For the extreme case that $\beta_o'\mathbf{x}$ is a function of $\beta'\mathbf{x}$, say $\beta_o'\mathbf{x} = h(\beta'\mathbf{x})$, the least favorable g is obvious. We need only to set $g(\alpha + u) = g_o(\alpha + h(u))$, in which case we cannot identify between β and β_o at all, a completely confounded case as mentioned in Section 1. For the general case, the idea is to replace $h(u)$ by the condition expectation $E(\beta_o'\mathbf{x}|\beta'\mathbf{x} = h)$. This leads to consider

$$g(\alpha + u) = g_o(\alpha + E(\beta_o'\mathbf{x}|\beta'\mathbf{x} = u))$$

Observe that

$$\begin{aligned}
E(\beta_o'\mathbf{x}|\beta'\mathbf{x} = u) &= E(\beta'\mathbf{x}|\beta'\mathbf{x} = u) - (\beta - \beta_o)'E(\mathbf{x}|\beta'\mathbf{x}) \\
&= u - (\beta - \beta_o)'E(\mathbf{x}|\beta_o'\mathbf{x} = u) + o(\beta - \beta_o)
\end{aligned} \tag{7.6}$$

Dropping the little o term and taking $\beta - \beta_o$ as δ, we have

$$g(\alpha + u) \approx g_o(\alpha + u - \delta'\eta(u))$$

This leads to the least favorable submodel (7.3).

10.7.2 Information loss for nearly linear regression.

The information loss when the regression function is known only to be approximately linear can be assessed by taking

$$g_o(u) = u$$

in Theorem 7.1. Thus the minimum Fisher information is

$$\mathcal{I}_{min} = \sigma^{-2} E(\mathbf{x} - \eta(\beta_o'\mathbf{x}))(\mathbf{x} - \eta(\beta_o'\mathbf{x}))' \tag{7.7}$$

To compare this with the Fisher information $\mathcal{I}$ from the linear model theory, we recall the notations in (2.4)- (2.9) with $\mathbf{b} = \beta_o$ and decompose $\sigma^2\mathcal{I}$ as

$$\begin{aligned} \sigma^2\mathcal{I} &= cov\ L_{\beta_o}(\beta_o'\mathbf{x}) + cov\ \mathcal{K}_{\beta_o}(\beta_o'\mathbf{x}) + cov\ \mathbf{e} \\ &= (var(\beta_o'\mathbf{x}))^{-1} \cdot \Sigma_{\mathbf{x}}\beta_o\beta_o'\Sigma_{\mathbf{x}} + \Sigma_{\beta_o} + \mathcal{I}_{min} \end{aligned} \tag{7.8}$$

The first term in (7.8) is a rank one matrix. The loss of this term is because we cannot distinguish β from its scalar multiple under (7.1); any scalar multiple of β can be absorbed into the g. The least favorable submodel also reflects this limitation because for δ in the direction of β_o, the function $\delta'\eta(\cdot)$ becomes a multiple of the identity function and we only identify β up to a proportion. The second and the third terms in (7.8) constitute the information for estimating β up to a proportionality in the linear model. To see this, consider any homogeneous functions of β :

$$\phi(c\beta) = \phi(\beta), \ \ for\ any\ c \neq 0$$

One example is $\phi(\beta) = \frac{\beta}{||\beta||} \cdot sign(\beta)$, where $sign$, taking values $+1$ or -1, can be set in any convenient way so that the function will be symmetric under the sign change of the argument. The asymptotic covariance matrix for the least squares estimate is $\phi(\hat{\beta}_{ls})$ equals

$$\nabla\phi(\beta_o)\mathcal{I}^{-1}\nabla\phi(\beta_o)' = \nabla\phi(\beta_o)(\Sigma_{\beta_o} + \mathcal{I}_{min})^{-}\nabla\phi(\beta_o)'$$

where the superscript $-$ denotes a generalized inverse matrix, and the gradient operator returns a row vector. The equality comes from the fact that $\nabla\phi(\beta_o)$ is orthogonal to β_o.

We are ready to claim the following observation.

Observation 7.1. When g is approximately linear, the information loss in estimating the regression slope vector up to a proportionality constant is equal to $\sigma^{-2}\Sigma_{\beta_o}$.

The most nonlinear direction γ_1 against β_o given by Lemma 2.1 (with $\mathbf{b} = \beta_o$) is the direction where the relative information loss is the greatest. This observation depicts well the relationship between the quasi-helical confounding and the information loss. If κ_{β_o} is zero then there is no information loss. But when the quasi-helical confounding is serious, κ_{β_o} is large and so is the information loss. The scatterplot of the projection $\gamma_1'\mathbf{x}$ against $\beta_o'\mathbf{x}$ reveals the nonlinear pattern that causes the most serious information loss. Suppose we are allowed to collect more data. Then the information on this direction can be fortified if the projections of the added data points on this plot are distributed in a way to flatten out the curvilinearity as much as possible. This opens up a new aspect in guiding the selection of a future sample from any given set of possible cites. The same consideration should be given simultaneously to the ith most nonlinear direction $\hat{\gamma}_i'\mathbf{x}$ as well if the information loss on that direction, reflected by the associated eigenvalue, is still large.

Remark 7.1. Let $\hat{\beta}_{min}$ denote any efficient estimate under the least favorable submodel. We can verify that the ratio of $var(u'\hat{\beta}_{min})$ to $var(u'\hat{\beta}_{ls})$ is the largest when $u = \Sigma_{\mathbf{x}}\gamma_1$, among all $u'\beta_o = 0$.

10.8 Hypothesis testing for nearly linear regression.

This section discusses the impact of quasi-helical confounding in testing the null hypothesis that β is in the direction of a given $\beta_o \neq 0$:

$$H_o : \beta = c\beta_o, \;\; for\ some\ c \in R$$

when the regression model is only approximately linear. We will examine the level and the power of the standard F-test in Sections 8.1-8.2. Then an alternative test will be given in section 8.3.

10.8.1 Significance levels for F-tests.

If the model (7.1) is known to be exactly linear, $g(t) = t$, an F-test can be conducted in the usual way. However, if the model is only approximately linear, by how much will the level of the test exceed the nominal value under H_o ? As it turns out, this problem becomes more serious as the nonlinearity measure κ_{β_o} gets larger. Thus quasi-helical confounding can substantially increase the chance of falsely rejecting the null hypothesis. Likewise, in the next subsection, we shall show that the power of the F-test is also seriously impaired.

For a given sample, $(y_i, \mathbf{x}_i)$, $i = 1, \cdots, n$, the usual F-test is based on the ratio :

$$\begin{aligned} F &= \frac{(SSE_o - SSE_1)/(p-1)}{SSE_1/(n-p-1)} \qquad (8.1) \\ SSE_o &= \text{residual sum of squares under } H_o \\ SSE_1 &= \text{residual sum of squares under } H_1 \end{aligned}$$

Under the exact linear model assumption, SSE_o/σ^2 and SSE_1/σ^2 are chi-square random variables with $(n-2)$ and $(n-p-1)$ degrees of freedom respectively when the null hypothesis is true. When the sample size is large, $SSE_1/(n-p-1)$ converges to σ^2, and we can approximate the F-test by a chi-square test with $p-1$ degrees of freedom. The null hypothesis is rejected if $(p-1)F > C_{p-1,\alpha}$ where $C_{p-1,\alpha}$ denotes the $(1-\alpha)$ quantile of a chi-square distribution with $p-1$ degrees of freedom.

When $g(\cdot)$ is not linear, both SSE_1/σ^2 and SSE_o/σ^2 follow noncentral chi-square distributions, conditional on $\mathbf{x}$. The overlinearization measure defined in section 5 is a major source for the noncentrality parameters. Set $\beta = \beta_o$ and recall the two lack of fit functions $H_{ls}(\cdot)$, and $h(\cdot)$, from (5.2), (5.5). Let

$$var\ h(\beta_o'\mathbf{x})/\sigma^2 = \delta^*$$

be the measure for the departure of $g(\cdot)$ from linearity. It is easy to see that, SSE_1/n converges to

$$var(\epsilon + H_{ls}(\beta_o'\mathbf{x})) = \sigma^2 + var\ H_{ls}(\beta_o'\mathbf{x}) = \sigma^2(1+(1-OL)\delta)$$

Similarly, SSE_o/n converge to $\sigma^2(1+\delta^*)$. Thus the noncentrality parameter for $(SSE_o - SSE_1)/\sigma^2$ is approximately equal to $n\delta^* \cdot OL$. If follows that approximately

$$(p-1)F \sim (1+(1-OL)\delta)^{-1}\chi^2_{p-1}(\phi), \tag{8.2}$$

where $\chi^2_{p-1}(\phi)$ denotes a noncentral chi-square distribution with the noncentrality parameter $\phi = n\delta^* \cdot OL$. Denote the level of the test at $g(\cdot)$ by

$$\alpha_g = P\{(p-1)F > C_{p-1;\alpha}\ |H_o\}$$

Using the approximation (8.2), we see that

$$\alpha_g \approx P\{\chi^2_{p-1}(n\delta^* \cdot OL) > (1+(1-OL)\delta^*)C_{p-1;\alpha}\}$$

which increases as OL gets larger. By Theorem 5.1, the maximum over all functions with a given δ^* is obtained when $OL = \kappa_{\beta_o}$. Therefore the worst level is approximately equal to

$$\max_g \alpha_g \approx P\{\chi^2_{p-1}(n\delta^*\kappa_{\beta_o}) > C_{p-1;\alpha}(1+(1-\kappa_{\beta_o})\delta^*)\} \tag{8.3}$$

The last expression shows that the chance of falsely rejecting H_o is an increasing function of the nonlinearity measure in $\mathbf{x}$ along the direction β_o. This is another ill-effect of quasi-helical confounding. From (8.3), we see that the worst level easily exceeds 50% if the noncentrality parameter, $\phi = n\delta^*\kappa_{\beta_o}$, is greater than $(1+(1-\kappa_{\beta_o})\delta^*)C_{p-1;\alpha}$. In other words, in order to keep the α level under 50%, we absolutely cannot accommodate the nonlinearity with

$$\delta^* > \frac{C_{p-1;\alpha}}{(n+C_{p-1;\alpha})\kappa_{\beta_o} - C_{p-1;\alpha}} \tag{8.4}$$

Thus we see that stricter linearity condition must be imposed when the quasi-helical confounding gets more serious. On the other hand, the most favorable situation is when $\kappa_{\beta_o} = 0$, which implies $OL = 0$. The significant levels can be guaranteed as long as $\delta = o(1)$ (as opposed to the rate $o(n^{-1})$ inferred from (8.4)). For more discussion, see Remark 8.1 at the end of this subsection.

Serious quasi-helical confounding can make the assumption (8.4) too stringent to be practical. Here is a simple way for tempering this problem. First follow Sections 2-3 and find the most nonlinear direction γ_1 against β_o. Then conduct an usual F-test for the null hypothesis

$$H_o^* : \beta \text{ is in the space spanned by } \gamma_1, \beta_o$$

We then reject our original hypothesis H_o only if the intermediate hypothesis H_o^* is rejected. This procedure definitely makes the probability of falsely rejecting H_o much smaller. In fact, the worst α-level can be approximated by (8.3) with κ_{β_o} replaced by the second eigenvalue $\lambda_2 = \kappa_{\gamma_2'\mathbf{x}|\beta_o'\mathbf{x}}$ from (2.7). If this is still too big, we can enlarge the space for H_o by including γ_i's successively.

The drawback of using the intermediate hypothesis H_o^* is the lack of power in differentiating the direction γ_1 from the direction β_o. But the standard F test for H_o does not have much power either if κ_{β_o} is large. This is discussed in the next subsection. Another testing procedure is considered in subsection 8.3.

Remark 8.1. The nominal level of the F-test is better maintained if $OL = 0$. This can be argued from a couple of viewpoints. First, if $\mathbf{x}$ is elliptically symmetric, the asymptotic distribution for $(p-1)F$ statistic (without conditioning on $\mathbf{x}$) is a chi-square distribution, with $p-1$ degrees of freedom. This follows easily from Li and Duan (1989). If $\mathbf{x}$ is not elliptically symmetric, then the result still holds under the condition that

$$Eh(\beta_o'\mathbf{x})^2(\mathbf{x} - E\mathbf{x})(\mathbf{x} - E\mathbf{x})' = o(1) \tag{8.5}$$

Condition (8.5) is the same as $\delta^* = o(1)$ if the support of $\mathbf{x}$ is bounded. If (8.5) does not hold, then the noncentrality parameter does not vanish. Unlike the case that $OL \neq 0$, the magnitude of the noncentrality parameter remains bounded as the sample size n increases.

10.8.2 Loss of power for F-tests.

The power of an F-test under exact linear model assumption can be computed using noncentral F-distributions or noncentral chi-squares for a large sample size. In our case, the power at β is approximately equal to

$$P_\beta \approx P\{\chi^2_{p-1}(\phi) > C_{p-1;\alpha}\}$$

where the noncentrality parameter ϕ is just the residual sum of squares for regressing $\beta'\mathbf{x}$ on $\beta_o'\mathbf{x}$. Thus

$$\phi \approx n(\beta'\Sigma_\mathbf{x}\beta - (\beta'\Sigma_\mathbf{x}\beta_o)^2/\beta_o'\Sigma_\mathbf{x}\beta_o)/\sigma^2$$

which increases at rate n. If the true regression function $g(\alpha+\beta'\mathbf{x})$ deviates from the nonlinearity, then the noncentrality parameter can get either smaller or larger. The worst situation is when the (population) least squares estimate β_{ls}, defined in (5.1) with $G(\mathbf{x}) = g(\alpha+\beta'\mathbf{x})$, is proportional to β_o :

$$\beta_{ls} = c\beta_o, \text{ for some } c \tag{8.2.1}$$

In this case, the noncentrality parameter will not tend to the infinity as n increases. The power of the test remains bounded away from one even for very large sample sizes; in other words, the F-test will not be consistent.

Consider $g(\cdot)$ with the decomposition

$$\begin{aligned} & g(\alpha + \beta'\mathbf{x}) = \alpha + \beta'\mathbf{x} + h(\beta'\mathbf{x}) \\ & var\ h(\beta'\mathbf{x}) \leq \delta^o \end{aligned} \tag{8.2.2}$$

where $h(\beta'\mathbf{x})$ is uncorrelated with $\beta'\mathbf{x}$. We want to find the smallest δ^o, denoted by δ_β, so that there exists such a g for which (8.2.1) holds. We may consider δ_β as the largest amount of nonlinearity that can be incorporated into the linear model to maintain the consistency of the F-test.

Recall the notation $\Sigma_\mathbf{b}$ again from section 2, with $\mathbf{b} = \beta$. It can be shown that

$$\delta_\beta = (c\beta_o - \beta)'\Sigma_\mathbf{x}\Sigma_\beta^-\Sigma_\mathbf{x}(c\beta_o - \beta) \tag{8.2.3}$$

$$c = \beta'\Sigma_\mathbf{x}\beta/\beta'\Sigma_\mathbf{x}\beta_o \tag{8.2.4}$$

Geometrically, if we take $\Sigma_\mathbf{x} = I$, then $c\beta_o$ is just the vector from the β_o line whose projection on the β line is equal to β. In general, (8.2.4) is satisfied if $(c\beta_o - \beta)'\mathbf{x}$ is uncorrelated with $\beta'\mathbf{x}$. We can derive a lower bound for (8.2.3) from the eigenvalue decomposition (2.7) with $\mathbf{b} = \beta$:

$$\delta_\beta \geq \kappa_\beta^{-1}(c\beta_o - \beta)'\Sigma_\mathbf{x}(c\beta_o - \beta) \tag{8.2.5}$$

We see that for maintaining consistency of the F-test, larger deviation from linearity can be allowed if the nonlinearity in $\mathbf{x}$ along the direction β is less serious.

When we are concerned with local powers of the $F-$test, β is assumed to be close to β_o. Put $\beta = \beta_o + v$, where v is a direction uncorrelated with β_o, $v'\Sigma_\mathbf{x}\beta_o = 0$. We can approximate Σ_β by Σ_{β_o}, and $c \approx 1$, as $v'\Sigma_\mathbf{x}v$ tends to zero. Thus we can approximate (8.2.3) by

$$\delta_\beta \approx v'\Sigma_\mathbf{x}\Sigma_{\beta_o}^-\Sigma_\mathbf{x}v \geq \kappa_{\beta_o}^{-1}v'\Sigma_\mathbf{x}v \tag{8.2.6}$$

This again shows how the nonlinearity in $\mathbf{x}$ along the direction β_o may impair the power of the test locally. The most nonlinear direction γ_1 is the direction where the power loss can be the most serious.

10.8.3 A quasi-helix-robust test based on the least favorable submodel.

In this subsection, we consider an alternative test based on the least favorable submodel (7.3) with $g_o(\cdot)$ being the identify function. By making the task of differentiating between β_o and β the hardest, the least favorable submodel gives the null hypothesis H_o the maximum benefit of doubt. This explains why the significance level of the derived test can be guaranteed even if the true $g(\cdot)$ does not belong to the model (7.3).

(7.3) involves nonlinear regression. It may be laborious to carry out a likelihood ratio test. To simplify the task, we propose the following model as an approximation :

$$y = \alpha + \beta'\mathbf{x} - \delta'\eta(\beta_o'\mathbf{x}) + \epsilon \tag{8.3.1}$$

Here we simply replace β in the argument of the η function by β_o in (7.3) and ignore the approximation error $\delta'(\eta(\beta'\mathbf{x}) - \eta(\beta_o'\mathbf{x}))$.

(8.3.1) is a linear model with parameters α, β, δ. But there is an estimability problem due to the collinearity between $\mathbf{x}$ and $\eta(\beta_o'\mathbf{x})$, because of the identity $\beta_o'\mathbf{x} = \beta_o\eta(\beta_o)$. To avoid this problem, we may require the nuisance parameter δ to be uncorrelated $\beta_o : \delta'\Sigma_{\mathbf{x}}\beta_o = 0$. Rewrite (8.3.1) as

$$y = \alpha + \beta'\mathbf{x} - \delta'\mathcal{K}_{\beta_o}(\beta_o'\mathbf{x}) + \epsilon \tag{8.3.2}$$

Now we can conduct a standard F-test for the null hypothesis H_o :

$$\begin{aligned} F^* &= \frac{(SSE_o^* - SSE_1^*)/(p-1)}{SSE_1^*/(n-2p)} \\ SSE_1^* &= \text{residual sum of squares under (8.3.2)} \\ SSE_o^* &= \text{residual sum of squares under (8.3.2) and } H_o \end{aligned}$$

We then reject H_o if F exceeds the $(1-\alpha)$ quantile of an F-distribution with degrees of freedom $(p-1, n-2p)$.

When $g(\cdot)$ is nonlinear, the significance level of this test is much better protected than the standard F-test outlined in section 8.1. Under H_o, we can show that even if $g(\cdot)$ is nonlinear, the β parameter for the best fit under the linear model (8.3.2) is always in the direction of β_o: the solution β^*, δ^* for the following minimization is achieved at $\beta^* = c^*\beta_o$:

$$\min_{a,\mathbf{b},\delta} E(g(\alpha + \beta_o'\mathbf{x}) - a - \mathbf{b}'\mathbf{x} - \delta'\mathcal{K}_{\beta_o}(\beta_o'\mathbf{x}))^2 \tag{8.3.3}$$

An easy way to see this is to consider the decomposition (2.8) with $\mathbf{b} = \beta_o$ again. We have

$$\begin{aligned} & E(g(\alpha + \beta_o'\mathbf{x}) - a - \mathbf{b}'(L_{\beta_o}(\beta_o'\mathbf{x})) - (\mathbf{b}+\delta)'\mathcal{K}_{\beta_o}(\beta_o'\mathbf{x}) - \mathbf{b}'\mathbf{e})^2 \\ \geq \quad & E(g(\alpha + \beta_o'\mathbf{x}) - a - \mathbf{b}'(L_{\beta_o}(\beta_o'\mathbf{x})) - (\mathbf{b}+\delta)'\mathcal{K}_{\beta_o}(\beta_o'\mathbf{x}))^2 + E(\mathbf{b}'\mathbf{e})^2 \end{aligned}$$

The last term in the preceding expression vanishes if $\mathbf{b}$ is in the direction of β_o. Therefore this shows that $\beta_o^* = c^*\beta_o$.

The nuisance parameter δ in the linear model (7.3) serves as an outlet for absorbing the bias of multiple linear regression due to the nonlinearity in $g(\cdot)$. The immediate consequence is that the expectation of $SSE_o^* - SSE_1^*$ remains bounded as n increases. The significance level is under much better protection.

To carry out this test, we need to estimate $\mathcal{K}_{\beta_o}(\cdot)$. This is not hard to implement because involves only p one-variable against one-variable nonparametric regressions. However, we prefer to carry out the quasi-helix searching procedure as suggested in Section 3 first. Then keep only a small number of non-zero eigenvectors, $\hat{\gamma}_1, \cdots, \hat{\gamma}_k$. We estimate only $\hat{\gamma}_i'\mathcal{K}_{\beta_o}(\cdot)$; this reduces the number of nonparametric regressions to just k. For each i, $1 \leq i \leq k$, let z_i denote any nonparametric regression fit of $\hat{\gamma}_i\mathbf{x}$ against $\hat{\beta}_{ls}\mathbf{x}$. Instead of (8.3.2), we can now implement the F test for the model

$$y = \alpha + \beta'\mathbf{x} - \sum_{i=1}^{k}\delta_i z_i + \epsilon \tag{8.3.4}$$

Our experience indicates that the result does not change much for a wide range of smoothing parameter. We shall refer to this testing procedure as a quasi-helix-robust test.

Example 8.3.1 For the data generated in Example 5.1., consider the hypothesis

$$H_o : \beta = c\beta_o = c(1, 0, -1, -1, 0)', \text{ for some } c$$

In this case, β_o is the true direction that generates the data. The standard F-test (8.1) gives a large value of $F = 12.6$; with R-squares of .587 and .730 respectively under the null and the alternative hypotheses. Thus H_o is falsely rejected by the standard F-test. Recall also that the standard residual plot fails to recognize any nonlinear pattern, which only reinforces the misleading conclusion. We now carry out the quasi-helix-hunting procedure for the direction $\mathbf{b} = \beta_o$. Only one nonzero eigenvalue is found, $\tilde{\lambda}_1 = 0.628$, $\hat{\gamma}_1 = (4.60, -1.47, -4.66, -3.02, 1.46)'$. This suggests the possibility of quasi-helical confounding. We use only the first nonlinear direction $\hat{\gamma}_1$ to carry out our quasi-helix-robust test. The LOWESS function from Xlips.stat (Tierney 1991) is used for smoothing. Figure 8.1 shows the result for $z_1 = \hat{\gamma}_1'\mathbf{x}$ against $\hat{\beta}_{ls}'\mathbf{x}$. The F value is reduced to 1.57, with the R-squares of .729 and .746. The p-value of the test is .19. Hence the correct null hypothesis won't be rejected by our test. Note that Figure 10.5 is produced using the default values. Improvement on smoothing is certainly quite feasible. We report this suboptimal situation to indicate that our testing procedure is not sensitive to the choice of smoothing parameters. Other values have been tried and yield similar results.

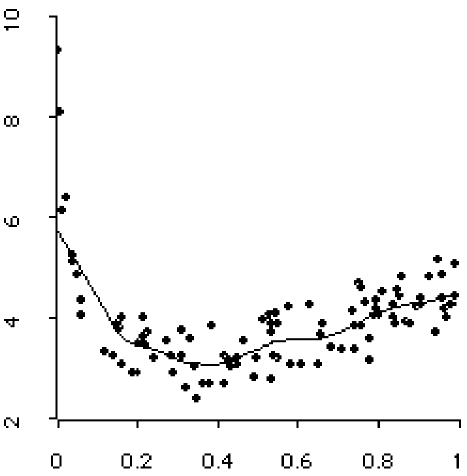


Figure 10.5: Smoothing by LOWESS for the most nonlinear direction against the null direction.

Chapter 11

Errors in Regressors

This chapter is taken from Carroll and Li(1992).

Errors in the regressors can cause severe bias in estimation. For linear models, there are many references and techniques available, see Fuller (1987). The standard least squares estimate tends to flatten the slope parameters. Thus adjustment is needed to correct the bias.

For nonlinear models, bias of course persists for the standard maximum likelihood estimation. Yet it gets even harder to make correction and relatively less attention has been given until recently, see Carroll (1989), Carroll and Stefanski (1990) and Carroll and Wand (1990) for recent reviews.Applications of binary measurement error models are discussed by Carroll, et al. (1984), Rosner et al. (1989) and Tosteson, et al. (1989), among others.

We will address the nonlinear case from the viewpoint of dimension reduction and data visualization as in the regressor-error-free situation.

11.1 The setting.

With errors in the regressor variables, nonlinear regression becomes much more complicated. For example, the likelihood function generally involves multiple integration and issues of model sensitivity and robustness are not well understood. Let Y be the response, $\mathbf{x}$ the true predictor(which is typically unobervable) and $\mathbf{w}$ the observed surrogate. Likelihood analysis requires specifying functional forms for the distributions of Y given $\mathbf{x}$ and $\mathbf{x}$ given $\mathbf{w}$. Part of our goal is to reduce the necessity for fully specifying these functional forms.

We adopt the dimension reduction model for Y and $\mathbf{x}$ from Chapter 1. Like Chapter 10, we consider only the case that the e.d.r. space has only one dimension:

$$Y = g(\alpha + \beta'\mathbf{x}, \epsilon), \tag{1.1}$$

The surrogate $\mathbf{w}$ is related to $\mathbf{x}$ via the linear model :

$$\mathbf{w} = \gamma + \Gamma\mathbf{x} + \delta \tag{1.2}$$

where Γ is a q by p matrix, which can be known , unknown, or partly known. We assume that δ is independent of $\mathbf{x}$ and ϵ, although this can be relaxed to δ and Y independent given $\beta'\mathbf{x}$. An important case is when $p = q$ and Γ equals the identity.

We shall present some methods of estimating β, for $p \geq 2$, up to a constant of proportionality **without knowledge of the functional form of g.**

11.1.1 The basic idea.

Similar to SIR, the key is to consider the inverse regression curve $\eta(y) = E(\mathbf{w}|Y = y)$. For the special case that

$$\mathbf{w} = \mathbf{x} + \delta$$

we see easily that

$$E(\mathbf{w}|Y = y) = E(\mathbf{x} + \delta|Y = y) = E(\mathbf{x}|Y = y)$$

Thus the inverse regression curve is exactly the same as when the regressor is without error. This immediate leads to the suggestion of applying SIR for Y on $\mathbf{w}$. But since the covariance matrix of $\mathbf{x}$ is not the same as the covariance matrix of $\mathbf{w}$ in general:

$$\Sigma_{\mathbf{w}} = \Sigma_x + \Sigma_\delta$$

we need correction when conducting eigenvalue decomposition. Suppose that knowledge of Σ_δ is available from other sources. Then an appropriate eigenvalue decomposition would be for $cov(E\mathbf{w}|Y)$ on $\Sigma_{\mathbf{w}} - \Sigma_\delta$. This is largely in line with (but not exactly the same as) what we shall pursue for the more general case of (1.2) in the next section.

11.2 Basic theory.

11.2.1 Preliminaries

If $\mathbf{x}$ were known, then we have shown that

$$E(\mathbf{x}|Y = y) - E\mathbf{x} \propto \Sigma_{\mathbf{x}}\beta \tag{2.1}$$

under the linear design condition: for any b,

$$E(b'\mathbf{x}|\beta'\mathbf{x}) = c_0 + c_1\beta'\mathbf{x} \tag{2.2}$$

(2.1) is the key to the theory of SIR. But since $\mathbf{x}$ is not available, this result cannot be applied immediately. Our strategy is to predict $\mathbf{x}$ from $\mathbf{w}$. Let

$$\begin{aligned} \mathbf{u} &= L\mathbf{w} \\ L &= cov(\mathbf{x}, \mathbf{w})\Sigma_{\mathbf{w}}^{-1} \end{aligned}$$

Clearly $\mathbf{u}$ is the best linear prediction of $\mathbf{x}$ from $\mathbf{w}$.

Theorem 2.1 below shows that $\mathbf{u}$ behaves just like $\mathbf{x}$. This allows us to apply methods available for $\mathbf{x}$ to $\mathbf{u}$.

Theorem 2.1: *Under (1.1), (1.2), (2.2), we have*

$$E(\mathbf{u}|Y=y) - E\mathbf{u} \propto \Sigma_{\mathbf{u}}\beta \qquad (2.3)$$

Proof: We can assume that $E\mathbf{w} = E\mathbf{x} = 0$. Now,

$$\begin{aligned} E(\mathbf{u}|Y) &= LE(\mathbf{w}|Y) = L\Gamma E(\mathbf{x}|Y) \\ &\propto L\Gamma\Sigma_{\mathbf{x}}\beta \\ &= Lcov(\mathbf{w},\mathbf{x})\beta \\ &= cov(\mathbf{x},\mathbf{w})\Sigma_{\mathbf{w}}^{-1}cov(\mathbf{x},\mathbf{w})'\beta \\ &= \Sigma_{\mathbf{u}}\beta \end{aligned}$$

This proves the theorem. □

There are two ways to apply this theorem, to be described below.

11.2.2 Linear regression type method

.

We can apply a standard linear least squares regression of Y against $\mathbf{u}$.

$$\min_{b_1\in R^q, a\in R} E(Y - a - b_1'\mathbf{u})^2.$$

Then the regression slope b_{ls} will be proportional to β:

$$b_{ls} \propto \beta \qquad (2.4)$$

Upon observing that

$$cov(Y,\mathbf{u}) = E(Y(\mathbf{u} - E\mathbf{u})) = E\{Y[E(\mathbf{u}|Y) - E\mathbf{u}]\} \propto \Sigma_{\mathbf{u}}\beta$$

the proof of (2.4) follows easily from the least squares formula, $b_{ls} = cov(y,\mathbf{u})\Sigma_{\mathbf{u}}$.

This result can be viewed as an extension of Brillinger (1977, 1983), where the usual linear least squares estimate was shown to be consistent for estimating the direction of the slope vector when the regressor variable is free of error. Li and Duan (1989) extended Brillinger's result to other regression estimates, including generalized linear models and M-estimates.

11.2.3 SIR type method

Denote $\zeta(y) = E(\mathbf{u}|Y = y)$ and define the covariance matrix $cov\{\zeta(Y)\}$ by Σ_ζ. We can find the nondegenerate direction by applying a suitable principal component analysis as the following Corollary suggests.

Corollary 2.1. *Under the same conditions as given in Theorem 2.1, the covariance matrix Σ_ζ has rank at most one. Assuming that Σ_ζ is of rank one, let v be the nonzero eigenvector* *for the eigenvalue decomposition of Σ_ζ with respect to $\Sigma_{\mathbf{u}}$: $\Sigma_\zeta v = \lambda \Sigma_{\mathbf{u}} v$, where λ is the nonzero eigenvalue. Then v is proportional to β : for some scalar c, $v = c\beta$.*

Note that the eigenvalue $\lambda = E\left[E\left\{\beta'(\mathbf{x} - E\mathbf{x})|Y\right\}^2\right]/\beta'\Sigma_{\mathbf{u}}\beta$.

Remark 2.1: If the joint distribution of **x** and **w** is normal, we can extend the result of Li and Duan (1989) to the error-in-regressors problem by pretending that we have observed an error-free regressor **u**. Specifically, for any function $\rho(Y, \theta)$ which is convex in θ, under (1.1) the solution (a_ρ, b_ρ) of the minimization

$$\min_{a\in R, b\in R^p} E\rho(Y, a + b'\mathbf{u})$$

satisfies the condition that b_ρ is proportional to β. In fact , the normality assumption can be weakened by assuming that conditional on $\beta'\mathbf{u}$, $\beta'\mathbf{x}$ is independent of **u**.

11.3 Estimation

Given an $i.i.d.$ sample, $(Y_i, \mathbf{w}_i)$, $i = 1, \cdots, n$, we discuss how to implement the two methods in Section 2. We distinguish among three different situations:

- L known;
- L unknown and estimated by an independent validation sample of $(\mathbf{x}, \mathbf{w})$;
- L unknown, but $p = q$ and Γ is known, so that L can be estimated by an independent sample containing replicates of **w**.

11.3.1 Least squares with known L

If L is known, define $\mathbf{u}_i = L\mathbf{w}_i$. The least squares estimate is

$$\tilde{b}_{\mathrm{ls}} = \tilde{\Sigma}_{\mathbf{u}}^{-1}(n-1)^{-1}\sum_{i=1}^{n} Y_i(\mathbf{u}_i - \bar{\mathbf{u}}), \tag{3.1}$$

where $\tilde{\Sigma}_{\mathbf{u}} = L\hat{\Sigma}_{1\mathbf{w}}L'$ is the sample covariance for **u**, and $\hat{\Sigma}_{1\mathbf{w}}$ is the sample covariance of **w**. It can be shown that this estimate is root n consistent and is asymptotically normal with mean b_{ls} defined by (2.4), and covariance matrix

$$\Sigma(\tilde{b}_{\mathrm{ls}}) = n^{-1}\Sigma_{\mathbf{u}}^{-1}\mathrm{cov}\Big\{e_i(\mathbf{u}_i - E\mathbf{u})\Big\}\Sigma_{\mathbf{u}}^{-1}. \tag{3.3}$$

where $e_i = Y_i - EY - b'_{ls}(\mathbf{u}_i - E\mathbf{u})$

11.3.2 SIR with known L

We can apply the SIR algorithm to **u**:

(I). Divide the range of Y into H slices and let $\hat{p}_h$ be the proportion of Y_i's falling into the h^{th} slice I_h.

(II). Within each slice compute the sample mean of **u**, $\bar{\mathbf{u}}_h = (n\hat{p}_h)^{-1}\sum_{Y_i \in I_h} \mathbf{u}_i$, $h = 1, \cdots, H$.

(III). Form the covariance matrix $\tilde{\Sigma}_\zeta = \sum_{h=1}^{H} \hat{p}_h(\bar{\mathbf{u}}_h - \bar{\mathbf{u}})(\bar{\mathbf{u}}_h - \bar{\mathbf{u}})'$. Then conduct an eigenvalue decomposition of $\tilde{\Sigma}_\zeta$ with respect to the sample covariance of **u**, $\tilde{\Sigma}_{\mathbf{u}}$:

$$\tilde{\Sigma}_\zeta \tilde{b}_{\text{sir}} = \tilde{\lambda}\tilde{\Sigma}_{\mathbf{u}}\tilde{b}_{\text{sir}}$$

where $\tilde{\lambda}$ is the largest eigenvalue.

The asymptotic distribution of $\tilde{b}_{\text{sir}}$ can be derived.
Theorem 3.1: *Under the same conditions as given in Theorem 2.1, the estimate $\tilde{b}_{sir}$ is root n consistent in estimating the direction of β. Moreover, there exists a version of $\tilde{b}_{sir}$ which is asymptotically normal with covariance matrix*

$$\Sigma(\tilde{b}_{sir}) = n^{-1}\Sigma_{\mathbf{u}}^{-1} cov\Big\{\tilde{e}_i(\mathbf{u}_i - E\mathbf{u})\Big\}\Sigma_{\mathbf{u}}^{-1}.$$

where $\tilde{e}_i = \tilde{Y}_i - E\tilde{Y}_i - b'_{sir}(\mathbf{u}_i - E\mathbf{u}))$ and b_{sir} is the regerssion slope of $\tilde{Y}$

11.3.3 Least squares with unknown L: validation

Often L is unknown and has to be estimated from another independent sample of size m, called a validation sample, $(\mathbf{x}_i, \mathbf{w}_i)$, $i = n+1, \cdots, n+m$. Typically, m is smaller than n. We now study the effect on our estimates in (3.1), due to the uncertainty from estimating L. We consider the estimate of L defined by

$$\hat{L} = \widehat{cov}(\mathbf{x}, \mathbf{w})\hat{\Sigma}_{2\mathbf{w}}^{-1}, \tag{3.8}$$

where $\widehat{cov}(\mathbf{x}, \mathbf{w})$ is the sample covariance between **x** and **w** and $\hat{\Sigma}_{2\mathbf{w}}$ is the sample covariance of **w**, all based on the validation sample $(\mathbf{x}_i, \mathbf{w}_i)$, $i = n+1 \cdots, n+m$. Each row of $\hat{L}$ is the usual least squares regression slope of the corresponding coordinate of **x** against **w**.

Denote $\hat{\mathbf{u}}_i = \hat{L}\mathbf{w}_i$ for $i = 1, \cdots, n$ and define the associated sample covariance

$$\hat{\Sigma}_{\mathbf{u}} = \hat{L}\hat{\Sigma}_{1\mathbf{w}}\hat{L}'. \tag{3.9}$$

We shall replace $\mathbf{u}_i$'s by $\hat{\mathbf{u}}_i$'s in constructing our estimates. The resulting estimates of β will be denoted by $\hat{b}_{\text{ls}}$ and $\hat{b}_{\text{sir}}$ respectively. Now let

$$\mathbf{r}_i = \mathbf{x}_i - \mathbf{u}_i = \mathbf{x}_i - L\mathbf{w}_i; \qquad \Lambda_{\mathbf{rw}i} = (\mathbf{r}_i - E\mathbf{r}_i)(\mathbf{w}_i - E\mathbf{w})'.$$

Theorem 3.2: *The asymptotic covariance matrix of $\hat{b}$ is given by*

$$\Sigma(\hat{b}_{ls}) = cov\big\{I(b_{ls})\big\} + \Sigma(\tilde{b}_{ls}),$$

where $\Sigma(\tilde{b}_{ls})$ is given by (3.3) and

$$cov\big\{I(b_{ls})\big\} = m^{-1} cov\Big\{\Sigma_{\mathbf{u}}^{-1}(\mathbf{u}_i - E\mathbf{u})(\mathbf{r}_i - E\mathbf{r}_i)b'_{ls}\Big\}.$$

Compared to the result in section 3.1, the first term reflects the cost of estimating L.

11.3.4 SIR with unknown L: validation

After estimating L by $\hat{L}$, the SIR type estimate can be carried out in the same way as described in Section 11.3.2. The matrix $\tilde{\Sigma}_\zeta$ will be replaced by

$$\hat{\Sigma}_\zeta = \hat{L}\hat{\Sigma}_\eta \hat{L}', \quad \text{where} \tag{3.13}$$

$$\hat{\Sigma}_\eta = \sum_{h=1}^{H} \hat{p}_h(\bar{\mathbf{w}}_h - \bar{\mathbf{w}})(\bar{\mathbf{w}}_h - \bar{\mathbf{w}})', \tag{3.14}$$

and where $\bar{\mathbf{w}}_h$ is the slice mean for $\mathbf{w}$, $(n\hat{p}_h)^{-1}\sum_{Y_i \in I_h} \mathbf{w}_i$. The maximum eigenvector of the eigenvalue decomposition $\hat{\Sigma}_\zeta$ with respect to $\hat{\Sigma}_{\mathbf{u}}$, see (3.9), is our estimate $\hat{b}_{\text{sir}}$. We shall find the asymptotic distribution for $\hat{b}_{\text{sir}}$ below.

Theorem 3.3: *There exists a version of $\hat{b}_{sir}$ with the asymptotic mean b_{sir} given by (3.7) and the covariance matrix*

$$\Sigma(\hat{b}_{sir}) = cov\{I(b_{sir})\} + \Sigma(\tilde{b}_{sir}),$$

where $\Sigma(\tilde{b}_{sir})$ is given in Theorem 3.1, and $cov\{I(b_{sir})\}$ is the same as the term $cov\{I(b_{ls})\}$ given in Theorem 3.2, with b_{ls} replaced by b_{sir}.

Remark 3.2: If Γ is known, then we can estimate L by $\hat{\Sigma}_{\mathbf{x}}\Gamma'\hat{\Sigma}_{2\mathbf{w}}^{-1}$. The derivation of the asymptotic distribution will be similar.

Remark 3.3: Note that the term cov$\{I(b_{\text{sir}})\}$ (respectively, cov$\{I(b_{\text{ls}})\}$) is the asymptotic covariance matrix for the estimated slope when we regress $b'_{\text{sir}}\mathbf{x}$ (respectively, $b'_{\text{ls}}\mathbf{x}$) against $\mathbf{u}$ based on the validation sample, if b_{sir} (respectively, b_{ls}) were known. Thus the additional uncertainty in our estimate due to estimating L is easy to assess. This information may be particularly useful in planning of the sample sizes m and n.

Remark 3.4: In Theorem 3.3, we have assumed that the slices are fixed. In practice, it is more convenient to choose approximately the same number of observations for each slice, unless Y is discrete. This makes our procedure invariant under monotone transformations of Y. The case that H increases as the sample size increases, in particular two observations per slice, can be treated using the methods of Hsing and Carroll (1990).

11.3.5 Least squares with unknown L: replication

An important special case occurs when Γ is known and $p = q$. Without loss we will take $\Gamma = I$, in which case $\mathbf{w}$ is an unbiased surrogate for $\mathbf{x}$. In many experiments, instead of validation we will have a replicated data set, i.e.,

$$\mathbf{w}_{ij} = \alpha + \mathbf{x}_i + \delta_{ij}, \quad j = 1, 2; \; i = n+1, ..., n+m. \tag{3.17}$$

If Σ_δ is the covariance of δ_{ij}, then we find that $L = I - \Sigma_\delta \Sigma_{\mathbf{w}}^{-1}$. Define $\widehat{\Sigma}_\delta$ and $\widehat{\Sigma}_{\mathbf{w}}$ to be (1/2) the sample covariance matrices of $w_{i1} - w_{i2}$ and $w_{i1} + w_{i2}$ respectively, and define

$$\widehat{L} = I - \widehat{\Sigma}_\delta \widehat{\Sigma}_{\mathbf{w}}^{-1}. \tag{3.18}$$

With this choice of L, the results similar to those in Section 11.3.3 can be derived.

11.3.6 SIR with unknown L: replication

In the replication model (3.17), with the estimate (3.18), the results similar to those in Section 11.3.4 go through analogously.

11.4 Statistical inference

We shall show how to apply the results of Section 3 to hypothesis testing and confidence interval problems.

In the nonlinear measurement error model literature, hypothesis testing has not been much discussed. An exception is the case of testing for a simultaneous null effect in all components of $\mathbf{x}$ measured with error, where score test ideas can be used; see Tosteson and Tsiatis (1988) and Stefanski and Carroll (1990). These methods do not apply for testing components of $\mathbf{x}$ which are measured precisely, see Carroll (1989) for examples. We can handle such problems using the techniques in Section 3.

Now consider the hypothesis testing problem of the form

$$H_0 : M\beta = 0 \; v.s. \; H_1 : M\beta \neq 0$$

where M is a given r by p matrix of rank $r \leq p$. For instance, if we take $M = (1, 0, \cdots, 0)$, then $r = 1$ and we are testing if the first variable in $\mathbf{x}$ affects the response Y or not. Let $\hat{\beta}$ denote any estimator constructed in Section 3. In order to construct a Wald test for the hypothesis, we need a consistent estimate $\widehat{\Sigma}(\widehat{\beta})$ of $\Sigma(\widehat{\beta})$, the asymptotic covariance of $\widehat{\beta}$. For the least squares estimates of Sections 3.1, 3.3 and 3.5, consistent estimation of $\Sigma(\widehat{\beta})$ is easy: just substitute population quantities by their estimates. The only point that needs a little case is that in the population version of (3.19), the term $\mathbf{w}_{i1} - \mathbf{w}_{i2}$ should be replaced by $\mathbf{w}_{i1} - \mathbf{w}_{i2} - m^{-1} \sum_{k=1}^{m} (\mathbf{w}_{k1} - \mathbf{w}_{k2})$.

For the SIR estimates, we can use Theorems 3.1 and 3.3 to construct estimate of the needed covariance.

The Wald test at level α rejects the hypothesis if

$$\widehat{\beta}' M' \left\{ M \widehat{\Sigma}(\widehat{\beta}) M' \right\}^{-1} M \widehat{\beta} > \chi_r^2(1-\alpha),$$

where $\chi_r^2(1-\alpha)$ is the appropriate percentage point of the chi–squared random variable with r degrees of freedom.

11.5 Generalizations

The method based on SIR is easy to generalize to other settings. First instead of (1.1) we may consider a model

$$Y = g(\alpha + \beta'\mathbf{x}, T, \epsilon),$$

where T is a stratification variable such as sex, $age\ group$, etc, so that g can be an arbitrary function with three arguments. We assume the design distribution to satisfy the linear conditional expectation condition (2.2) after conditioning on T; namely, for any direction b in R^p, there are functions of T, $c_0(T)$, $c_1(T)$, such that

$$E(b'\mathbf{x}|\beta'\mathbf{x}, T) = c_0(T) + c_1(T)\beta'\mathbf{x}.$$

It is clear that the inverse regression curves, $\eta(y|T) = E(\mathbf{w}|Y = y, T)$ and $\zeta(y|T) = E(\mathbf{u}|Y = y, T)$, still have the same property as given in Theorem 2.1, and (2.7):

$$\eta(y|T) = E(\mathbf{u}|T) + c(y|T)\Sigma_{\mathbf{u}|T}\beta, \qquad (2.7')$$

where $c(y|T) = (\beta'\Sigma_{\mathbf{x}|T}\beta)^{-1}E(\beta'(\mathbf{x} - E\mathbf{x})|Y = y, T)$ and $\Sigma_{\mathbf{x}|T}$ is the conditional covariance of $\mathbf{x}$ given T. Thus when creating slices we need to use both Y and T. We can estimate β from each stratum of T and then combine the estimates.

Under the additional assumption that $\Sigma_{\mathbf{x}|T}$ does not depend on T, we can combine the estimate of the covariance Σ_ζ from each stratum:

$$\sum_t \sum_h \hat{p}_{h,t}(\bar{\mathbf{u}}_{h,t} - \bar{\mathbf{u}}_t)(\bar{\mathbf{u}}_{h,t} - \bar{\mathbf{u}}_t)'),$$

where $\hat{p}_{h,t}$ is the proportion of the cases falling into slice h at stratum $T = t$, $\bar{\mathbf{u}}_{h,t}$ is the sample average of $\mathbf{u}$ for $T = t$ and slice h, and $\bar{\mathbf{u}}_t$ is the sample average of $\mathbf{u}$ for $T = t$. Then we can estimate β by the largest eigenvector of this matrix with respect to $\Sigma_\mathbf{x}$ as before. The asymptotic property will be similar to what is discussed before. The result of Hsing and Carroll (1990) can be used to justify the consistency property of the resulting estimate even if the number of cases per slice is small.

Another generalization is to consider the K component model (1.2). We can use the largest K eigenvectors of SIR (Step III of Section 3.2) to estimate the space spanned by the β's. We can justify our method based on the generalization of (2.7):

$$\eta(y) - E(\mathbf{u}) \text{ falls into the space spanned by } \Sigma_\mathbf{u}\beta_1, \cdots, \Sigma_\mathbf{u}\beta_K.$$

The proof of this result follows along the same line as that in the proof of Theorem 3.1 of Li (1990a), incorporating the crucial property that $E(\delta|Y = y, \mathbf{x}) = E(\delta) = 0$ as employed in the proof of our Theorem 2.1.

11.6 Data visualization

When regressors are subject to error, even when $\mathbf{x}$ and $\mathbf{w}$ are scalar, plots of Y against $\mathbf{w}$ might give misleading information about the regression slope (or functions) of Y against $\mathbf{x}$ (Spiegelman, 1986; Fuller, 1987). This can be the case even if the measurement error is small (Carroll and Stefanski, 1990). Despite these possibilities, in many instances curvature in $E(Y|\mathbf{w})$ does reflect curvature in $E(Y|\mathbf{x})$, and plotting Y against $\mathbf{w}$ may still be valuable. This section presents a theory for informatively visualizing the data when regressors are subject to errors.

When $\mathbf{x}$ is observable, the best viewing angle is $\beta'\mathbf{x}$ under (1.1). But since $\mathbf{x}$ is not available, we can at best find a projection on $\mathbf{w}$ so that the projected variable has the highest correlation with $\beta'\mathbf{x}$ to ensure the best viewing angle obtainable from $\mathbf{w}$, i.e., the one that is closest to the best view from $\mathbf{x}$. From Remark 2.2 at the end of Section 2, we see that $\beta'\mathbf{u}$, or its scalar multiple, is the desired variable. Since we can estimate the direction of β by $\hat{\beta}$, which denotes any estimate obtained earlier, we suggest plotting Y against $\hat{\beta}'\mathbf{u}$. If the correlation between $\beta'\mathbf{x}$ and $\beta'\mathbf{u}$ is high, then our plot may be useful , for instance, in suggesting the appropriate functional form in (1.1). When a validation sample of $(\mathbf{x}, \mathbf{w})$ is available, we can use it to estimate corr$(\beta'\mathbf{x}, \beta'\mathbf{u})$ by considering the sample correlation between $\hat{\beta}'\mathbf{x}$ and $\hat{\beta}'\mathbf{u}$.

The following is a simulation example to see how effective our method is. We consider two models for generating the data:

$$Y = (\alpha + \beta'\mathbf{x} + \epsilon)^2; \tag{6.1}$$

$$Y = (\alpha + \beta'\mathbf{x})^2 + \epsilon. \tag{6.2}$$

The first falls into the Box-Cox transformation family while for the second, the transformation is taken only for the mean response function , a special case of a “Transform–Both–Sides”model considered by Carroll and Ruppert 1988 (c.f. Remark 6.1 below). Each coordinate of $\mathbf{x}$ and ϵ are independent standard normal random variables. We set the dimension parameters as $p = 6 = q$, the primary sample size $n = 200$, the validation sample size $m = 100$, and $\beta = (1, 1, 1, 0, 0, 0)', \alpha = 4$. The relationship between $\mathbf{w}$ and $\mathbf{x}$ will be governed by the linear model (2.1) with $p = q = 6$, $\Gamma = I$. The distribution of δ is normal with mean 0 and covariance being a diagonal matrix with diagonal element $(0, 1/3, 1/3, 1/3, 1/3, 1/3)$. Thus except for the first one, each coordinate of $\mathbf{x}$ is contaminated by an error of a size equal to .577 of its standard deviation.

Since our procedure does not require the knowledge of the functional form that generates the data, we can apply it to both (6.1) and (6.2). After a hundred simulation runs, we summarize the results for $\hat{b}_{\text{sir}}$ in Tables 6.1 and 6.2. Since we are only interested in estimating the direction of β, we have standardized our estimate to have unit length. For each model the mean of $\hat{b}_{\text{sir}}$ is very close to the theoretical value $(.577, .577, .577, 0, 0, 0)' = \beta/||\beta||$. This demonstrates that our procedure can avoid the bias that one normally would have anticipated to occur without proper model specification. For instance, the naive estimate of regressing the square root of Y against $\mathbf{w}$ for data generated by (6.1) gives the slope $(1, .75, .75, 0, 0, 0)'$ on the average, which is not proportional to β. The standard deviation of our estimate is reasonably small, compared, for instance, to the ideal value of $.07 \approx \frac{1}{\sqrt{n}}$, the standard deviation of the least squares estimate for β under (6.1) if the square transformation were known and if $\mathbf{x}$ were observed without errors. The cosine of the angle between $\hat{b}_{\text{sir}}$ and β, which is the same as the correlation coefficient between $\hat{b}'_{\text{sir}}\mathbf{x}$ and $\beta'\mathbf{x}$, is very close to one, with the lowest value in the neighborhood of .95, see the next to the last column in each table. We considered three choices of the number of slices $H = 5, 10, 20$, with an equal number of observation per slice. This illustrates the stability of our procedure in regard to the change in H at least for this example.

Table 6.1: Summary of $\hat{b}_{\text{sir}} = (\hat{b}_1, \cdots, \hat{b}_6)'$ for model (6.1). Standard deviation and minimum are in parentheses and brackets respectively.

# of slices	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	$\hat{b}_5$	$\hat{b}_6$	$cos_{\mathbf{x}}$	$corr_{\mathbf{w}}$
5	.571 (.069)	.567 (.059)	.569 (.059)	.001 (.075)	-.002 (.071)	.007 (.079)	.986 [.949]	.986 [.950]
10	.571 (.067)	.568 (.052)	.570 (.058)	-.002 (.069)	-.005 (.070)	.004 (.077)	.987 [.959]	.988 [.959]
20	.571 (.066)	.569 (.054)	.568 (.062)	.005 (.073)	-.005 (.067)	.006 (.078)	.986 [.950]	.987 [.949]

Table 6.2: Summary of $\hat{b}_{\text{sir}} = (\hat{b}_1, \cdots, \hat{b}_6)'$ for model (6.2). Standard deviation and minimum are in parentheses and brackets respectively.

# of slices	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	$\hat{b}_5$	$\hat{b}_6$	$cos_{\mathbf{x}}$	$corr_{\mathbf{w}}$
5	.569 (0.059)	.573 (.047)	.571 (.053)	.007 (.060)	.000 (.072)	.002 (.067)	.989 [.968]	.990 [.971]
10	.573 (.060)	.571 (.045)	.572 (.051)	.005 (.057)	-.001 (.065)	.004 (.062)	.990 [.974]	.991 [.972]
20	.570 (.059)	.573 (.045)	.573 (.053)	.006 (.057)	-.004 (.064)	.003 (.061)	.990 [.970]	.991 [.968]

Now to demonstrate the application of our method for data visualization, a single additional run is taken. For this realization, we found that

$$\begin{aligned} \hat{b}_{\text{sir}} = (-0.54, -0.64, -0.53, -0.03, -0.09, -0.01)' \qquad & \text{for model (6.1);} \\ (-0.57, -0.58, -0.56, 0.04, -0.03, -0.00)' \qquad & \text{for model (6.2).} \end{aligned}$$

The estimate of the transformation L is given in Table 6.3. Then we compute $\hat{b}'_{\text{sir}}\hat{\mathbf{u}}_i = (\hat{L}\hat{b}_{\text{sir}})'\mathbf{w}_i$ for $i = 1, \cdots, n$. The scatterplots of Y_i against $\hat{b}'_{\text{sir}}\hat{\mathbf{u}}_i$ are given in Figures 11.1.

Table 6.3: Estimated L.

$$\begin{pmatrix}
1.00 & .000 & .000 & .000 & .000 & .000 \\
.077 & .765 & .030 & .104 & .012 & .023 \\
.038 & -.006 & .785 & .013 & .013 & -.020 \\
.012 & .015 & .048 & .760 & -.048 & .107 \\
.059 & -.0619 & -.044 & -.063 & .704 & -.047 \\
.037 & -.010 & -.059 & .065 & .046 & .747
\end{pmatrix}$$

and 11.4, respectively for models (6.1) and (6.2). These plots suggest that analysis based on transformation is reasonable to pursue further.

In Figures 11.2 and 11.5, we plot Y_i against $\beta'\mathbf{u}_i$, the best view on Y from the surrogate variable if L and β were known. Compared to what we had seen in Figures 6.1 and 6.4, we see that very little has been lost due to the estimation. The correlation between the variable $\beta'\mathbf{u}$ and our estimated variable $\hat{\beta}'\mathbf{u}$ is as high as .99. The last column in each of Tables 6.1 and 6.4 gives the mean and the lowest possible value of this correlation over the same 100 simulation runs done earlier. The lowest number is still as high as .95. This suggests that approximately the same view would be obtained from other simulation runs.

We also provide plots of Y_i against $\beta'\mathbf{x}_i$ in Figures 11.3 and 11.6, the best view of the models from the uncontaminated regressor. We find that the views obtained by our estimate, Figures 11.1 and 11.4, are also very close to these best views. This can be attributed to the fact that for our parameter setting, in spite of the apparent heavy contamination rate, the correlation between $\beta'\mathbf{x}$ and $\beta'\mathbf{u}$ is high, equaling to about .91. Figures 11.3, and 11.6 are supposed to be close to Figures 11.2 and 11.5 respectively, which in turn have been shown to be similar to Figures 11.1 and 11.4.

If we have a small number of additional data points available on $(Y, \mathbf{x})$, we can plot Y against $\hat{\beta}'\mathbf{x}$ as well. To illustrate this, we generate ten new data points for $(Y, \mathbf{x})$ from (6.2). The plot is given in Figure 11.7, which shows a quadratic trend well.

Remark 6.1: Carroll and Ruppert (1988) considered the Transform–Both–Sides model, which allows another transformation on Y before getting a model like (6.2). Since our procedure is invariant under the monotone transformation, we would have obtained the same estimate if our data were given after applying any unknown monotone transformation to Y in the model (6.2).

11.7 A case study.

We have access to a restricted data set containing breast cancer incidence Y and $\mathbf{x}$ = (age, body mass, nutient intake), the latter in this case being the logarithm of satuated fat. The primary data set was of size $n = 2800$ while the validation data were of size $m = 650$. The fallible version of nutrient intake was assessed in the study by an interview detailing the previous day's diet, while the version of truth used here was the average of 4 such interviews.

The measurement error is quite large: fully more than 50% of the observed variability in fat is error. All measurement error analyses of these data performed previously have shown a large age effect and a negligible effect due to body mass. Some analyses show a significant effect due to the nutrient intake, while others do not: in all cases the coefficient has been negative.

Programming the methods discussed in this paper is very easy. This is particularly the case for binary regression, because the least squares and SIR methods discussed in Section 3 yield identical estimates of β, in terms of unit length. In this example, the ordinary logistic regression estimate of unit length obtained from regressing Y on $\mathbf{w}$ was $(0.97, -0.12, -0.20)$ with two–sided significance levels $(.00, .68, .06)$. Our methods yielded estimates $(0.87, -0.14, -0.47)$ with two–sided significance levels $(.00, .64, .02)$. Note that this measurement error analysis yields a larger estimated relative effect due to the nutrient than did the ordinary logistic regression. The difference in statistical significance for the saturated fat coefficient may be due to the use of information standard errors for the ordinary logistic analysis: an analysis using M–estimator techniques to construct standard errors yielded lover significance levels. For example, if we assume that given observed fat, the true fat is independent of age and body mass (checked by a linear regression analysis), then the method of Rosner, et al. (1989) applied to these data yielded estimates and significance levels $(0.91, -.11, -.40)$ and $(.00, .64, .01)$ respectively.

Appendix

A.1. Proof of Fisher consistency in convex regression, Remark 2.1

We use the conditional argument similar to that for the proof of Theorem 2.1 in Li and Duan (1989). For any $b \in R^p$, by Jensen's inequality, we shall have

$$\begin{aligned} E\rho(Y, a + b'\mathbf{u}) &= E\Big[E\Big\{\rho(Y, a + b'\mathbf{u})|\beta'\mathbf{x}, \epsilon, \beta'\mathbf{u}\Big\}\Big] \\ &\geq E\Big[\rho\Big\{Y, a + E(b'\mathbf{u}|\beta'\mathbf{x}, \beta'\mathbf{u})\Big\}\Big] \\ &= E\Big[\rho\Big\{Y, a + E(b'\mathbf{u}|\beta'\mathbf{u})\Big\}\Big] \\ &= E\Big\{\rho(Y, a + c_0 + c_1\beta'\mathbf{u})\Big\}, \quad \text{for some real numbers} \quad c_0, c_1. \end{aligned}$$

Here the second to the last equality is due to the fact that given $\beta'\mathbf{u}$, $\mathbf{u}$ is independent of $\beta'\mathbf{x}$, a consequence of the joint normality of $\mathbf{x}$ and $\mathbf{w}$. It is now clear that a minimizer can be found along the direction of β, proving our claim.

Chapter 12

Censored Regression via the SIR Approach

This chapter is taken from Li, Wang, and Chen(1999).

Censoring makes high-dimensional regression analysis even more complex. One may apply SIR directly to the regressor and see what happens. If the censoring time is independent of the lifetime, this is fine. Otherwise, some modification is needed to adjust for the censoring bias. A key identity leading to the bias correction is derived and the root-n consistency of the modified estimate is established. Patterns of censoring can also be studied under a similar dimension reduction framework.

12.1 Introduction.

Survival data are often subject to censoring. When this occurs, the incompleteness of the observed data may induce a substantial bias in the sample. Several approaches have been suggested to overcome the associated difficulties in regression, including the accelerated failure time model, censored linear regression, Cox proportional hazard model, and many others. Survival analysis becomes even more intricate when the dimension of the regressor increases. To apply any of the aforementioned methods, users are required to specify a functional form which relates the outcome variable to the input ones. However in reality, knowledge needed for an appropriate model specification are often inadequate. As a matter of fact, the acquisition of such information may well turn out to be just one of the primary goals of the study itself. Under such circumstances, it seems preferable to have exploratory tools that rely less on such model specification. We shall see how our dimension reduction approach can be extended to settings which allow for censoring in the data. We shall offer methods of finding low-dimensional projections of the data for visually examining the censoring pattern. We shall show how censored regression data can still be analyzed without assuming the functional form a priori.

Recall the dimension reduction model from Chapter 1 :

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

To incorporate censoring into the dimension reduction framework, let

$$Y^o \quad = \quad \text{the true (unobservable) lifetime,}$$

$$\begin{aligned} C &= \text{ the censoring time,} \\ \delta &= \text{ the censoring indicator; } \delta = 1, \text{ if } Y^o \le C, \text{ and } \delta = 0, \text{ otherwise.} \\ Y &= min\{Y^o, C\}, \text{ the observed time.} \end{aligned}$$

We assume that

$$Y^o \text{ follows model (1.1),} \tag{1.2}$$

$$\text{Conditional on } \mathbf{x},\ C \text{ is independent of } Y^o. \tag{1.3}$$

The observed sample consists of n i.i.d. observations, $(Y_i, \mathbf{x}_i, \delta_i), i = 1, ..., n$ from the distribution of $(Y, \mathbf{x}, \delta)$. The continuous random variables, Y^o, C, are not observed. Condition (1.3) is the usual independence assumption to ensure identifiability under the random censoring scheme. If (1.3) is violated, then one needs more information on the censoring mechanism to build an appropriate model. This is not considered in this paper.

For $k = 1$, our formulation may include the generalized linear model (McCullagh and Nelder 1989) and the linear transformation model (Doksum 1987) as special cases. The latter also includes several survival analysis models such as the accelerated failure time model, the proportional hazard model, the proportional odds model, and the logit and probit models (Doksum and Gasko 1990).

For exploratory purpose, we may want to temporarily ignore the presence of censoring and apply SIR on $(Y, \mathbf{x})$. But the question is what can be inferred from the output. How does censoring affect SIR estimates? The answer depends on the relationship between the censoring time C and the regressor $\mathbf{x}$. The independence case in which

$$C \text{ is independent of } \mathbf{x}, \text{ and } Y^o. \tag{1.4}$$

is easy and will be treated first. The dependence case is more complicated. Of special interest to us is the one that uses a dimension reduction assumption on the censored time C as a counterpart of (1.2):

$$C = h(\gamma_1'\mathbf{x}, \cdots, \gamma_c'\mathbf{x}, \epsilon'). \tag{1.5}$$

12.2 SIR under the independence assumption (1.4).

Denote the uncensored inverse regression curve by $\eta^o(y^o) = E(\mathbf{x}|Y^o = y^o)$. Without censoring (i.e., $Y = Y^o$), the population version of SIR is based on the following eigenvalue decomposition:

$$\begin{aligned} \Sigma_{\eta^o} b_i &= \lambda_i \Sigma_{\mathbf{x}} b_i, \\ \lambda_1 &\ge \cdots \ge \lambda_p, \end{aligned} \tag{2.1}$$

where

$$\Sigma_{\eta^o} = cov(E(\mathbf{x}|Y^o)), \tag{2.2}$$

and

$$\Sigma_{\mathbf{x}} = cov(\mathbf{x}).$$

The justification for using the first k eigenvectors b_i with nonzero eigenvalues to estimate the e.d.r. (lifetime) directions follows from Theorem 2.1 in Chapter 2, which can be stated as follows:

Lemma 2.1._Assume that the dimension reduction assumption (1.2) holds. Then for any y^o, $\Sigma_{\mathbf{x}}^{-1}(\eta^o(y) - E(\mathbf{x}))$ falls into the e.d.r. (lifetime) space under the condition that_

$$\textit{for any vector } b,\ E(b'\mathbf{x}|\beta_1'\mathbf{x}, \cdots, \beta_k'\mathbf{x}) \textit{ is linear.} \tag{2.3}$$

Design condition (2.3) is called the **(L.D.C.)** earlier and is extensively Discussed in Chapter 8.

Censoring alters the distribution of the observed time Y. Its effect on SIR can be studied by comparing the censored inverse regression curve $\eta(y) = E(\mathbf{x}|Y = y)$ with the uncensored one $\eta^o(y^o)$. By conditioning, we have

$$E(\mathbf{x}|Y = y) = E(E(\mathbf{x}|Y^o, C)|Y = y). \tag{2.4}$$

Under (1.4), $E(\mathbf{x}|Y^o, C)$ is equal to $E(\mathbf{x}|Y^o)$, implying that

$$E(\mathbf{x}|Y = y) = E(\eta^o(Y^o)|Y = y). \tag{2.5}$$

Since Lemma 2.1 applies to $\eta^o(y^o)$, the following result is obtained.

Lemma 2.2. _Assume that (1.2) and (1.4) hold. Then $\Sigma_{\mathbf{x}}^{-1}(\eta(y) - E(\mathbf{x}))$ falls into the e.d.r. lifetime space under (2.3)._

The sample version of SIR can be implemented as described before without change. With Lemma 2.2, we can obtain the root-n consistency of SIR estimates $\hat{b}_i$ for finding e.d.r. lifetime directions. Thus censoring does not introduce bias to SIR. But this is true only when the censoring time is independent of the regressors and the true lifetime. Without (1.4), this appealing result vanishes and substantial bias may be induced by censoring under the more general condition (1.3).

12.3 A strategy for modifying SIR under (1.3).

An ideal way of bypassing the difficulties caused by general censoring (1.3) is to slice the true survival time Y^o. At first sight, this does not appear feasible because under censoring, Y^o is unobservable. But the promise comes from an identity derived in Section 3.1, which relates the conditional expectation of $\mathbf{x}$ in each slice to the observed time Y and the censored indicator. This leads to a modified slicing step by a suitable weighting scheme for offsetting the censoring bias in estimating the slice means. The consistency of this new procedure is discussed in Section 3.2.

12.3.1 An identity.

Let $0 = t_1 < t_2 < \cdots < t_H < \infty = t_{H+1}$ be a partition on the survival time. The expected value of $\mathbf{x}$ in a slice, $\mathbf{m}_j = E\{\mathbf{x}|Y^o \in [t_j, t_{j+1})\}$, can be written as

$$\mathbf{m}_j = \frac{E\{\mathbf{x}1(Y^o \in [t_j, t_{j+1}))\}}{P\{Y^o \in [t_j, t_{j+1})\}} = \frac{E\{\mathbf{x}1(Y^o \geq t_j)\} - E\{\mathbf{x}1(Y^o \geq t_{j+1})\}}{E\{1(Y^o \geq t_j)\} - E\{1(Y^o \geq t_{j+1})\}}, \tag{3.1}$$

where $1(\cdot)$ is the indicator function. The two numerator terms take the same form which involves the unobservable indicator $1(Y^o \geq t)$. They can be converted into terms with Y and δ via the identity:

$$E\{\mathbf{x}1(Y^o \geq t)\} = E\{\mathbf{x}1(Y \geq t)\} + E\{\mathbf{x}1(Y < t, \delta = 0)w(Y, t, \mathbf{x})\}, \tag{3.2}$$

where for $t' < t$,

$$\begin{aligned} w(t', t, \mathbf{x}) &= \frac{S^o(t|\mathbf{x})}{S^o(t'|\mathbf{x})}, & (3.3)\\ S^o(t|\mathbf{x}) &= P\{Y^o \geq t|\mathbf{x}\} = \text{ conditional survival function for } Y^o, \text{ given } \mathbf{x}. & (3.4) \end{aligned}$$

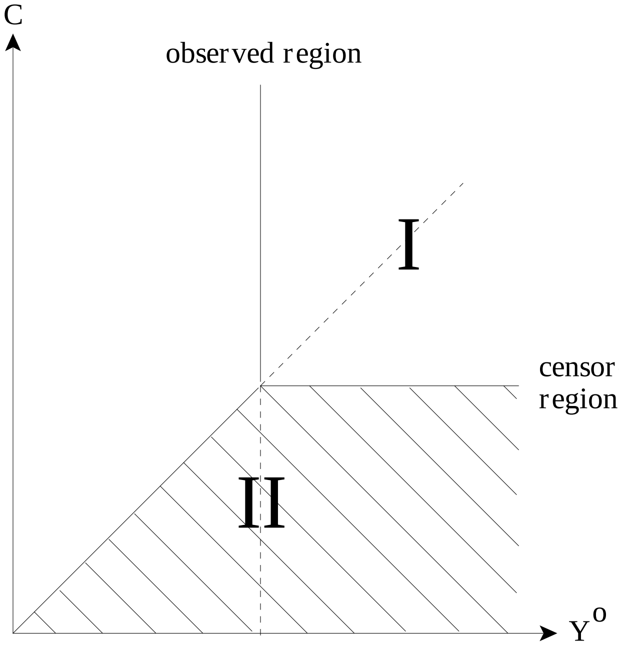


Figure 12.1: Integration regions.

Consider the plane of variables Y^o and C in Figure 12.1. The integration region $Y^o \geq t$ is decomposed into two parts. The first region (area **I** in Figure 12.1) with $Y^o \geq t, C \geq t$, or equivalently, $Y \geq t$, contributes to first term on the right side of (3.2). The second region with $Y^o \geq t, C < t$, falls into the censored area, $\delta = 0$. It is contained in the larger region (dashed area **II** in Figure 12.1) with $Y < t$ and $\delta = 0$. The second term on the right side of (3.2) comes from integration over this larger region with the weight adjustment $w(\cdot, \cdot, \cdot)$. Conditioning is the key to justify this term:

$$\begin{aligned} E\{\mathbf{x}1(Y^o \geq t, C < t)\} &= E\{\mathbf{x}1(Y < t, \delta = 0)1(Y^o \geq t)\} \\ &= E\{\mathbf{x}1(Y < t, \delta = 0)E[1(Y^o \geq t)|Y, \delta = 0, \mathbf{x}]\} \\ &= E\{\mathbf{x}1(Y < t, \delta = 0)E[1(Y^o \geq t)|C, Y^o > C, \mathbf{x}]\} \\ &= E\{\mathbf{x}1(Y < t, \delta = 0)w(C, t, \mathbf{x})\} \\ &= E\{\mathbf{x}1(Y < t, \delta = 0)w(Y, t, \mathbf{x})\}. \end{aligned}$$

Here the next to the last equality is due to the conditional independence assumption (1.3) which assures that conditional on **x**, the probability for the true survival time Y^o to exceed t given $C = t'$ and $Y^o \geq t'$ is equal to the conditional probability given by (3.3).

By a similar argument, the denominator terms can be converted via the identity:

$$E\{1(Y^o \geq t)\} = E\{1(Y \geq t)\} + E\{1(Y < t, \delta = 0)w(Y, t, \mathbf{x})\}. \tag{3.5}$$

The weight function (3.3) can be further expressed as

$$w(t', t, \mathbf{x}) = exp\{-\Lambda(t', t|\mathbf{x})\}, \tag{3.6}$$

where

$$\begin{aligned} \Lambda(t', t|\mathbf{x}) &= E\{\frac{1(t' < Y < t, \delta = 1)}{S_Y(Y|\mathbf{x})}|\mathbf{x}\}, \\ S_Y(\cdot|\mathbf{x}) &= \text{ the conditional survival function of } Y \text{ conditional on } \mathbf{x}. \end{aligned}$$

(3.6) follows from the well-known relationship between survival functions and cumulated hazards; for a proof, see Appendix A. The term $\Lambda(t', t|\mathbf{x})$ is simply the integrated conditional hazard (given **x**) function over the interval $[t', t]$.

12.3.2 Estimation.

To construct an estimate for $\mathbf{m}_j$, we replace each expectation term in (3.2) and (3.5) by the corresponding first sample moment:

$$\hat{\mathbf{m}}_j = \frac{\hat{E}\{\mathbf{x}1(Y^o \geq t_j)\} - \hat{E}\{\mathbf{x}1(Y^o \geq t_{j+1})\}}{\hat{P}\{Y^o \geq t_j\} - \hat{P}\{Y^o \geq t_{j+1}\}} \tag{3.7}$$

$$\hat{E}\{\mathbf{x}1(Y^o \geq t)\} = n^{-1} \sum_{i:\, Y_i \geq t}^{n} \mathbf{x}_i + n^{-1} \sum_{i:\, Y_i < t, \delta_i = 0}^{n} \mathbf{x}_i \cdot \hat{w}(Y_i, t, \mathbf{x}_i) \tag{3.8}$$

$$\hat{P}\{Y^o \geq t\} = \#\{i : Y_i \geq t\}/n + n^{-1} \sum_{i:\, Y_i < t, \delta_i = 0}^{n} \hat{w}(Y_i, t, \mathbf{x}_i) \tag{3.9}$$

where $\hat{w}(\cdot,\cdot,\cdot)$ denotes an estimate of the weight function (3.3) to be discussed later.

After estimating each slice mean by (3.7) , we can form the covariance matrix of the slice means in the usual way:

$$\begin{aligned}\hat{\Sigma}_{\eta_o} &= \sum_j (\hat{\mathbf{m}}_j - \bar{\mathbf{x}})(\hat{\mathbf{m}}_j - \bar{\mathbf{x}})'\hat{p}_j \\ \hat{p}_j &= \hat{P}\{Y^o \geq t_j\} - \hat{P}\{Y^o \geq t_{j+1}\}\end{aligned}$$

Finally, we may conduct the eigenvalue decomposition as before to find the SIR directions:

$$\begin{gathered}\hat{\Sigma}_{\eta^o}\hat{b}_i^o = \hat{\lambda}_i \hat{\Sigma}_{\mathbf{x}}\hat{b}_i^o, \\ \hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p.\end{gathered} \tag{3.10}$$

Smoothing is needed in estimating $w(t', t, \mathbf{x})$. There are several ways to proceed. For example, we can apply Beran's estimates for conditional survival functions and their variants. Under appropriate conditions, Beran(1981) and Dabrowska(1987, 1992) established the consistency of their estimates at convergence rates (slower than the root n rate) similar to those commonly found in nonparametric regression. These consistency results lead to the consistency of $\hat{\mathbf{m}}_h$ as an estimate of $\mathbf{m}_h$. It is easy to see that $\hat{\Sigma}_{\eta^o}$ is also consistent for the covariance matrix of the slice means $\mathbf{m}_h$'s. As in Li(1991), we can apply Lemma 2.1 to establish the consistency of $\hat{b}_i$ as estimates of e.d.r. lifetime directions.

Despite the slow rate of convergence in estimating conditional survival functions (hence the weight (3.3)), it is still possible to establish the root n convergence for $\hat{\mathbf{m}}_h$. We only consider the kernel smoothing method here for simplicity. Let $K_p(\cdot)$ be a kernel function on R^p and h_n be the bandwidth in each coordinate. We shall assume that $h_n = o(1)$ and nh_n^p tends to infinity. Further constraints will be imposed later. It is common for $K_p(\cdot)$ to take a product form : $K_p(x_1, \cdots, x_p) = K(x_1)\cdots K(x_p)$, for some one-dimension kernel function $K(\cdot)$. Our kernel estimate of (3.6) is defined by setting

$$\hat{\Lambda}(t', t|\mathbf{x}) = \frac{n^{-1}\sum_{i:\, t'<Y_i<t, \delta_i=1}^n (\hat{S}_Y(Y_i|\mathbf{x}_i))^{-1} h_n^{-p} K_p(h_n^{-1}(\mathbf{x}_i - \mathbf{x}))}{\hat{f}(\mathbf{x})}, \tag{3.11}$$

$$\hat{S}_Y(Y_i|\mathbf{x}_i) = \frac{n^{-1}\sum_{j:\, Y_j>Y_i}^n h_n^{-p} K_p(h_n^{-1}(\mathbf{x}_j - \mathbf{x}_i))}{\hat{f}(\mathbf{x}_i)}, \tag{3.12}$$

$$\hat{f}(\mathbf{x}) = n^{-1}\sum_i^n h_n^{-p} K_p(h_n^{-1}(\mathbf{x}_i - \mathbf{x})). \tag{3.13}$$

Under the regularity conditions , one can show that $\hat{\mathbf{m}}_h$ is a root-n consistent estimate for $\mathbf{m}_h, h = 1, \cdots, H$. This in turn implies that $\hat{b}_h$ is a root-n consistent estimate for an e.d.r. direction.

Assume that $f(\cdot)$ and $S(\cdot|\mathbf{x})$ are d-times continuously differentiable and that the kernel function satisfied the moment conditions : $\int x^i K_p(x)dx = 0$ for $i = 1, \cdots, d-1$, and $\int x^d K_p(x)dx$ is nonzero. Then the regularity assumptions needed can be satisfied with bandwidth $h \propto n^{-1/2d}$ provided $p \leq d$.

What we have presented so far in this section is a general strategy for offsetting the bias due to censoring. However, a direct application of this method may not help much in practice. The problem is that kernel smoothing only works well in the low dimensional case.

12.4 Dimension reduction model for censoring time.

Analyzing the censoring pattern is an important step in studying the censored data. It helps the recognition of the information-poor region in $\mathbf{x}$, the region where censoring is heavy and the regression structure is thus harder to explore. Sometimes, such analysis may even become a primary part of the study. In some industrial applications, Y^o may be the potential yield of a production process and censoring C may occur because of machine malfunctioning for example. In addition to learning how various input variables $\mathbf{x}$ may affect the potential yield, quality-control engineers may equally be interested in how they affect the censoring rate- they need such knowledge to prevent machine malfunctioning as much as possible.

Like its counterpart Y^o, we now assume that the censoring time C also has a dimension reduction structure given by (1.5). Again, the functional form of h and the distributional form of ϵ' are both unspecified. This model suggests only that the dimension of the regressor can be reduced from p to c. The relationship between the e.d.r. space for the censoring time and the e.d.r. space for the true lifetime is arbitrary. They can be either identical, partly overlapped, or disjoint. Linear combinations of their elements form a space which will be called the *joint* e.d.r. space. If Y^o and C were used for slicing, then by the same argument used in deriving Lemma 2.1, it is easy to see that

$$\Sigma_{\mathbf{x}}^{-1}(E(\mathbf{x}|Y^o, C) - E(\mathbf{x})) \text{ falls into the joint e.d.r. space.} \tag{4.1}$$

But instead of (Y^o, C), we can only observe Y and δ. This suggests that Y and δ can be used simultaneously for slicing. Let $\eta_d(Y, 0) = E(\mathbf{x}|Y = y, \delta = 0)$, and $\eta_d(Y, 1) = E(\mathbf{x}|Y = y, \delta = 1)$. We may replace (2.1) with

$$\begin{aligned} Cov(\eta_d(Y, \delta))b_i &= \lambda_i \Sigma_{\mathbf{x}} b_i, \\ \lambda_1 \leq \cdots &\leq \lambda_p. \end{aligned} \tag{4.2}$$

By conditioning, $E(\mathbf{x}|Y, \delta) = E(E(\mathbf{x}|Y^o, C)|Y, \delta))$. Thus from (4.1), we see that

$$\Sigma_{\mathbf{x}}^{-1}(\eta_d(y, \delta) - E(\mathbf{x})) \text{ falls into the joint e.d.r. space.}$$

This justifies the use of eigenvectors from (4.2) to estimate the joint e.d.r. space.

The sample version of (4.2) is easy to carry out. Denote the number of slices for the uncensored ($\delta = 1$) observations by H_1. Let I_{1j}, $j = 1, \cdots, H_1$ be a partition of the positive real line into non-overlapping intervals. Similarly, denote the number of slices for the censored ($\delta = 0$) observations by H_0, and let I'_{0j}, $j = 1, \cdots, H_0$ be another partition of the positive real line. We first form the individual slice means by taking

$$\bar{\mathbf{x}}_{lj} = (n\hat{p}_{lj})^{-1} \sum_{i=1}^{n} \mathbf{x}_i 1(\delta_i = l, Y_i \in I_{lj}),$$

where $\hat{p}_{lj}$ is the proportion of cases with $\delta_i = l$ falling into interval I_{lj}. Then we compute the covariance matrix for the slice means, $\hat{\Sigma}_d = \sum_l \sum_j \hat{p}_{lj}(\bar{\mathbf{x}}_{lj} - \mathbf{x})(\bar{\mathbf{x}}_{lj} - \mathbf{x})$. Finally we conduct the eigenvalue decomposition:

$$\begin{aligned} \hat{\Sigma}_d \hat{b}_{di} &= \hat{\lambda}_{di} \hat{\Sigma}_{\mathbf{x}} \hat{b}_{di}, \\ \hat{\lambda}_{d1} &\geq \cdots \geq \hat{\lambda}_{dp}. \end{aligned} \tag{4.3}$$

Note that the chi-squared test for determining the number of significant e.d.r. directions obtained by SIR can also be used for the double-slicing case.

Example 4.1. Take $p = 6$ and let $\mathbf{x} = (x_1, \cdots, x_6)'$ be generated from the standard normal distribution. Suppose

$$\begin{aligned} Y^o &= 4 - (|x_1 - 1|) + \sigma_1 \epsilon_1, \\ C &= 3 + \sigma_2 \epsilon_2, \text{ for } x_1 > 0, x_2 + x_3 > 0 \\ &= 10, \text{ otherwise} \end{aligned}$$

where $\sigma_1 = \sigma_2 = 0.1$, ϵ_1, ϵ_2 are normal random variables. Generate 300 cases. 66 observations in the data set are censored. Now apply double-slicing with the number of slices equal to 5 and 10, respectively for the censored and the uncensored groups. The eigenvalues of SIR are found to be .76, .35, .08, .06, $\cdots$, indicating that the first two eigenvectors, $\hat{b}_{d1} = (1.14, .05, -.03, -.00, -.04, .04)'$ and $\hat{b}_{d2} = (-.06, .69, .74, -.02, -.10, -.05)'$, are important. This is confirmed by the chi-squared test in Li(1991).

The joint e.d.r. directions $(1, 0, 0, 0, 0, 0)'$ and $\frac{1}{\sqrt{2}}(0, 1, 1, 0, 0, 0)'$ are captured successfully by $\hat{b}_{d1}$ and $\hat{b}_{d2}$. The censored cases are found to cluster in the first quadrant in the plot of the first two SIR variates; see Figure 12.2(c). The statistical information about the behavior of the true lifetime in that region is very little.

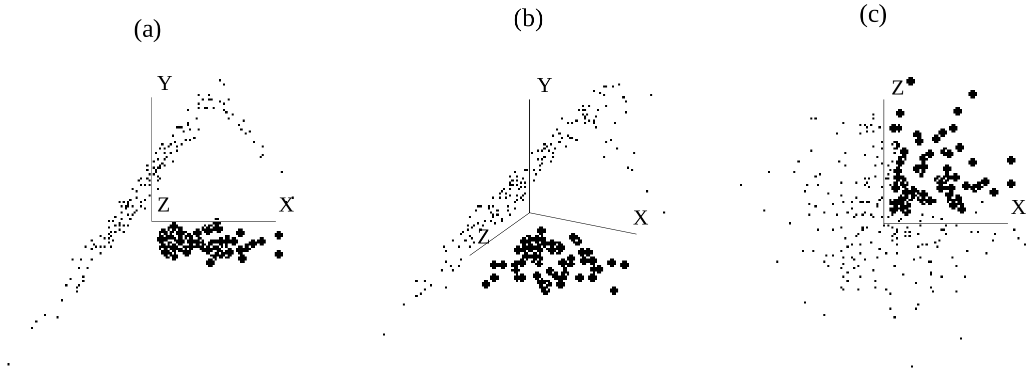


Figure 12.2: 3-D scatterplot of Y against the first two SIR variates found by double-slicing. The highlighted points are censored.

12.5 Implementation of modified SIR.

The directions found by double slicing can be used to relieve the difficulties encountered in Section 3 when kernel smoothing is to be applied for estimating the weight function (3.3). Under the dimension reduction assumptions for both the true lifetime and the censoring time, (1.2) and (1.5), it is easy to see that the dependence of the weight function (3.3) on $\mathbf{x}$ is only through joint e.d.r. variates. This suggests the following two-stage procedure:

(1) Apply double-slicing on (Y, δ) and find the joint e.d.r. directions, $\hat{b}_{di}$. Let $\hat{B}_r = (\hat{b}_{d1}, \cdots, \hat{b}_{dr})$ be the matrix formed by the first r significant directions.

(2) Apply r-dimensional kernel smoothing on $\hat{B}_r'\mathbf{x}$, to obtain the weight function $\hat{w}$:

$$\hat{w}(t', t, \mathbf{x}) = exp\{-\hat{\Lambda}(t', t|\mathbf{x})\} \tag{5.1}$$

where

$$\hat{\Lambda} * (t', t|\mathbf{x}) = \frac{n^{-1}\sum_{i:\,t'<Y_i<t,\delta_i=1}^{n}(\hat{S}_Y(Y_i|\mathbf{x}_i))^{-1}h_n^{-r}K_r(h_n^{-1}(\hat{B}_r(\mathbf{x}_i - \mathbf{x})))}{\hat{f}(\mathbf{x})} \tag{5.2}$$

$$\begin{aligned} \hat{S}_Y(Y_i|\mathbf{x}_i) &= \max\left\{\frac{n^{-1}\sum_{j:\,Y_j>Y_i}^{n} h_n^{-r}K_r(h_n^{-1}(\hat{B}_r(\mathbf{x}_j - \mathbf{x}_i)))}{\hat{f}(\mathbf{x}_i)}, c\right\} &(5.3)\\ \hat{f}(\mathbf{x}) &= n^{-1}\sum_{i}^{n} h_n^{-r}K_r(h_n^{-1}\hat{B}_r((\mathbf{x}_i - \mathbf{x}))). &(5.4)\end{aligned}$$

Note that a small positive number c (set to .05 in our examples) is used to bound $\hat{S}_Y(Y_i|\mathbf{x}_i)$ away from zero. This is needed in order to increase the stability of the factor $\hat{S}_Y(Y_i|\mathbf{x}_i)^{-1}$ in (5.2). After estimating the weight function, we can apply (3.7) $\sim$ (3.9) and then carry out the eigenvalue decomposition (3.10) to obtain estimates of e.d.r. lifetime directions.

We first conduct two simulation studies to illustrate how this strategy works. Then we apply our method to a data set concerning a study of the primary biliary cirrhosis in the liver (PBC).

Example 5.1. We take $p = 6$ and generate $\mathbf{x} = (x_1, \cdots, x_6)'$ from the standard normal distribution. The true survival time Y^o and the censoring time C are generated from

$$Y^o = -\frac{\log\epsilon_1}{e^{x_1}}; C = -\frac{\log\epsilon_2}{e^{x_2}}$$

where ϵ_1, ϵ_2 are independent uniform random variables from [0, 1]. Conditional on $\mathbf{x}$, Y^o and C are seen to follow the exponential distributions with the natural parameters λ_1, λ_2 linking to $\mathbf{x}$ via $\lambda_1 = e^{x_1}$; $\lambda_2 = e^{x_2}$, respectively.

We obtain 300 independent observations of (Y, δ), among them 138 cases are censored. We proceed with the SIR analysis. First, the method of double slicing on Y and δ as described by (4.3) gives eigenvalues $0.34, 0.27, 0.05, \cdots$. The first two eigenvectors, $\hat{b}_{d1} =$

$(-.67, -.70, -.08, .06, .11, 0.15)'$ and $\hat{b}_{d2} = (.69, -.73, .12, -.04, -.10, -0.12)'$, are close to the joint e.d.r. space for Y^o and C. We use these two directions to reduce the **x** dimension before estimating the weight function $w(\cdot, \cdot, \cdot)$. With the weight adjustment given by (5.1), we perform SIR as described by (3.7) $\sim$ (3.10) to find the e.d.r. lifetime directions. The eigenvalues are .40, .10, .03, $\cdots$, and the leading eigenvector is
$\hat{b}_1^o = (-.92, -0.12, -0.21, .08, 0.25, 0.11)'$. We see that $\hat{b}_1^o$ is quite close to the true e.d.r. lifetime direction $(1, 0, \cdots, 0)'$.

For comparison, we also carry out the SIR analysis on Y without weight adjustment as if the censoring were independent of **x**. The first direction,
$(-.68, -.69, -.058, .07, 0.13, 0.08)'$, does have a substantial bias. Therefore the weight adjustment is crucial in this example.

We used the bivariate normal kernel function here and the bandwidth is set at .18. The sensitivity to the bandwidth choice seems mild.

Example 5.2. Important prognostic variables affecting the hazard rate may be different at different survival stages. In this example, we assume that the true survival time Y^o follows an exponential distribution with the natural parameter equal to e^{2x_1} until time $\tau = log 2$. From time τ on, the additional survival time follows the exponential distribution with the natural parameter e^{3x_2}. More specifically, we assume

$$\begin{aligned} Y^* &\sim \text{exponential with parameter } \lambda = e^{2x_1}, \\ Y^{**} &\sim \text{exponential with parameter } \lambda = e^{3x_2}, \\ Y^o &= Y^* 1(Y^* < \tau) + (\tau + Y^{**}) 1(Y^* > \tau). \end{aligned}$$

The censoring time C follows an exponential distribution with parameter equal to $e^{x_3 - 1}$.

Again 300 independent observations of (Y, δ) are obtained. Among them, 98 cases are censored. The output of the double slicing procedure is given in Table 5.1. The first three eigenvectors, which have relatively larger eigenvalues compared to the rest, are then used in estimating the weight function for finding the true e.d.r. lifetime directions. After the weight adjustment, the final output of SIR is given in Table 5.2. Now we see that only the first two eigenvectors stand out and the important variables x_1 and x_2 can be identified.

Table 5.1: The first three eigenvectors and eigenvalues of SIR for Example 5.2 with the double-slicing procedure.

first vector	(-.93, -.11, .03, .03, .04, -.06)
second vector	(.09, -.76, -.60, -.01, .03, -.13)
third vector	(-.10, .55, -.73, -.03, -.02 .27)
eigenvalues	(.52, .21, .15, .03, .01, .01)

Example 5.3. The PBC data set collected at the Mayo Clinic between 1974 and 1986 has been analyzed in the literature. The data set and a detailed description can be found in Fleming and Harrington(1991). There are originally seventeen regressors. Fleming and

Table 5.2: The first two eigenvectors and eigenvalues of SIR for final result of Example 5.2 with weight adjustment.

first vector	(-.97, -.15, .10, -.04, .10, -.15)
second vector	(.16, -.95, -.18, -.02, -.06, -.20)
eigenvalues	(.66, .34, .05, .04, .02, .02)

Harrington selected five of them in their final equation for fitting a Cox proportional model. These five regressors plus another variable, the platelet count (x_5 below), will be used in this illustration:

Y = number of days between registration and the earlier of death or censoring

δ = 1 if Y is due to death; 0 otherwise

x_1 = Age in years

x_2 = presence of edema

x_3 = Serum bilirubin, in mg/dl

x_4 = Albumin, in gm/dl

x_5 = Platelet count

x_6 = prothrombin time

Cases with missing values are ignored and there are 308 cases remaining. We first apply double censoring with slice numbers $H_1 = H_0 = 10$. The first two directions are significant, as judged from the sequence of output eigenvalues $0.55, 0.15, 0.05, 0.0, 0.0, 0.0$. Figure 12.3 shows the scatterplot of the first two SIR variates. Two outliers labeled as 104 and 276 are found from the 3-D plot(not shown here) of Y against the first two SIR variates. They are removed. We apply double slicing again to the remaining 306 cases. The SIR output essentially remains the same. This suggests that the dimension of the joint e.d.r. space is two.

We proceed to find the true e.d.r. lifetime directions. We take $r = 2$ and use the two SIR directions reported in Table 5.3 to reduce the $\mathbf{x}$ dimension before estimating the weight function. The kernel function and the bandwidth are the same as in Example 5.1. The output of the weighted SIR is given in Table 5.4. Judging from the eigenvalue sequence, the first direction $\hat{b}_1^o$ is clearly important. The second direction is also worth further examination. Figure 12.4(a)(b) show the scatterplots of Y against $\hat{b}_1^{o\prime}\mathbf{x}$ and against $\hat{b}_2^{o\prime}\mathbf{x}$.

Earlier analysis in Fleming and Harrington(1991) yields that the true lifetime depends on $\mathbf{x}$ through the variate $Q = 0.0333x_1 + 0.7847x_2 + 0.8792 log x_3 - 3.0553 log x_4 + 3.0157 log x_6$. This variate turns out highly correlated with the first SIR variate $\hat{b}_1^{o\prime}\mathbf{x}$; the correlation coefficient is $\sqrt{0.858}$. The correlation between Q and $1.3\hat{b}_1^{o\prime}\mathbf{x} - 0.25\hat{b}_2^{o\prime}\mathbf{x}$ is equal to $\sqrt{0.89}$. Variable x_5 makes very little contribution to the first two SIR variates, with a squared multiple correlation of only 0.11. This is consistent with Fleming and Harrington's finding that platelet count is not important.

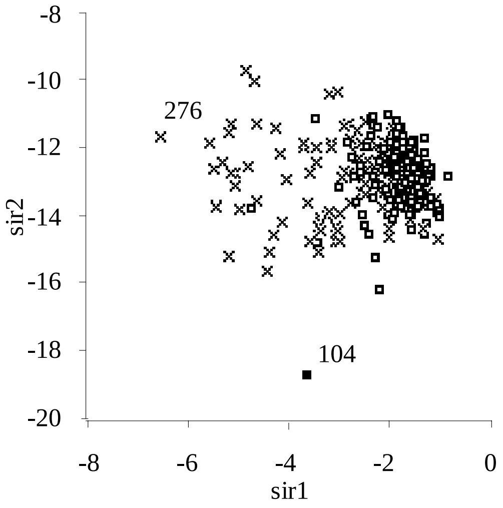


Figure 12.3: Scatterplot of the first two SIR variates found by double-slicing. x = observed, square = censored cases.

Table 5.3: The first two eigenvectors and eigenvalues of SIR for the PBC data in Example 5.3.

first vector	(.02, 1.04, .10, -.50, -.00, .39)
second vector	(.02, -1.62, .17, -.97, -.00, -.87)
eigenvalues	(.54, .16, .05, .01, .00, .00)

Finally, we estimate the censoring e.d.r. directions by reversing the roles of censoring time and the true lifetime. This amounts to replacing δ with $1 - \delta$ throughout our estimation procedure. The output is given by Table 5.5 and Figure 12.5. The assumption of independent censoring (1.4) is seen to be invalid for this data set. We further notice that the first censoring time direction is quite close to the first lifetime direction. The correlation coefficient between the first lifetime SIR variate and the first censoring SIR variate turns out to be $\sqrt{0.93}$.

Some caution needs to be taken regarding the design condition. Of the special concern is the second regressor (presence of edema) which is discrete and takes only three values (0, 0.5, 1). Nevertheless, the corresponding regression coefficient from Table 5.4 is 0.90, which is quite close to the coefficient 0.7847 based on Cox proportional hazard model. A further study would be to carry out another SIR analysis by focusing on the group with $x_2 = 0$. The other groups have only 29 and 19 cases and thus it is not feasible to carry out separate analyses for them.

Remark 5.1. In both of our simulation examples, we take $p = 6$. As the regressor dimension p gets larger, the problem certainly gets harder and one might expect the performance of our procedure to deteriorate as well. To study this effect, we vary p from 6 to 10, 15, and 20. The

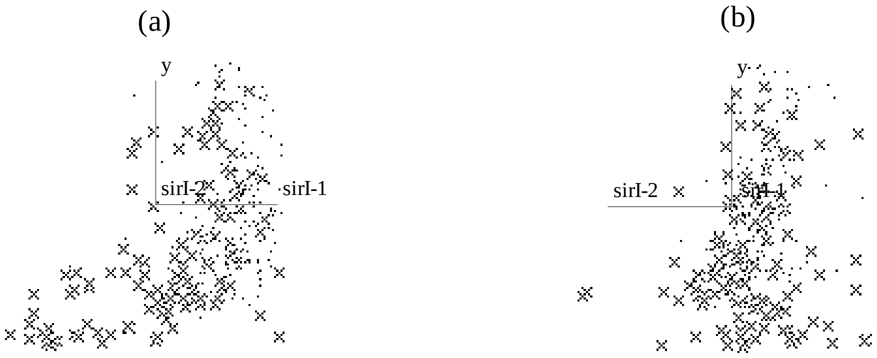


Figure 12.4: Scatterplot of Y against the first two lifetime SIR variates. x = observed, dot = censored cases.

Table 5.4: The first two eigenvectors and eigenvalues of the lifetime SIR directions for the PBC data in Example 5.3.

first vector	(.02, .90, .09, -.62, -.00, .38)
second vector	(.03, -2.3, .20, -.28, -.00, -.68)
eigenvalues	(.54, .16, .05, .02, .01, .00)

sample size is kept the same, $n = 300$. For each simulation run, we compute an R-squared term for evaluating how close to the true e.d.r. life time directions the estimated directions are. For the setup of Example 5.1 which has only one true e.d.r. life time direction, the R-squared term is simply the squared correlation coefficient between $\hat{b}_1^{o\prime}\mathbf{x}$ and $\beta_1'\mathbf{x}$. Since $\beta_1'\mathbf{x} = x_1$, the R-squared term is equal to the square of the first coordinate of $\hat{b}_1^o$. Table 5.6 (left-side panel) gives a summary of the R-squared values for 100 simulation runs in each case. For comparison, the R-squared values for the SIR estimate without the weight adjustment are given in the right-side panel. We can see that the improvement for the modified SIR procedure is still substantial for p as large as 20.

The setup of Example 5.2 has two true e.d.r. life time directions. For the first modified SIR direction, the R-squared term is just the R-squared value for regressing $\hat{b}_1^{o\prime}\mathbf{x}$ against $\beta_1'\mathbf{x}$

Table 5.5: The first two eigenvectors and eigenvalues, of the censoring time SIR variates for the PBC data in Example 5.3.

first vector	(.01, 1.43, .05, -.42, -.00, .55)
second vector	(.02, .38, -.15, 1.22, .00, .85)
eigenvalues	(.39, .22, .05, .03, .01, .00)

and $\beta_2'\mathbf{x}$ linearly. This is equal to the sum of the square of the first two coefficients in $\hat{b}_o$. The R-squared value for the second modified SIR direction is defined similarly. A summary for 100 simulation runs is given in Table 5.7.

Table 5.6: Performance of modified SIR as the number of regressors(p) increases under the setting of Example 5.1 with 100 runs.

mean(standard deviation) for R^2		
p	Modified SIR	Original SIR
6	.9172(0.0599)	.4751(0.1100)
10	.8630(0.0632)	.4736(0.0937)
15	.7963(0.0899)	.4322(0.0915)
20	.7576(0.0815)	.4152(0.0881)

Table 5.7: Performance of modified SIR as the number of regressors(p) increases under the setting of Example 5.2 with 100 runs.

mean(standard deviation) for R^2		
0	First modified SIR direction	Second modified SIR direction
6	.9730(0.0237)	.9132(0.0689)
10	.9434(0.0270)	.8455(0.0739)
15	.9239(0.0267)	.7911(0.0755)
20	.8933(0.0340)	.7149(0.1017)

12.6 Conclusion.

We have demonstrated how to extend the dimension reduction method of SIR to censored data. The extension is straightforward if censoring time is independent of the regressor. SIR can be applied to the observed data $(Y_i, \mathbf{x}_i)$ directly. But if censoring time depends on the regressor, then SIR needs to be modified. We introduce a weight function in estimating the slice means. The estimation of the weight function requires nonparametric smoothing. There are two options. The first one is to apply the kernel smoothing method of Section 3. This is feasible only if the number of regressors is small (e.g. $p \leq 3$) or if the sample size is substantially large. The other option, which seems more realistic, is the two-stage procedure of Section 5. We conduct a double-slicing SIR first to reduce the dimension of $\mathbf{x}$ before applying kernel smoothing. This two-stage procedure relies on condition (1.5) which assumes that the censoring variable also has a dimension reduction structure with respect to the regressor. This assumption appears reasonable and it offers the possibility to examine the censoring pattern visually.

The main feature that distinguishes our approach from most other methods in survival analysis is that it does not require the estimation of g at the dimension reduction stage of data analysis. Instead, after the dimension is reduced, the estimation of g can be pursued by applying any low dimensional smoothing methods. Furthermore, our approach can be used to check if a popular survival model is appropriate by examining the eigenvalues and the low dimensional plots generated by SIR. These plots provide valuable information about the general pattern of censoring, possible presence of outliers, and the shape of the regression surface.

Imputation is a powerful way of dealing with the incomplete censored observation. We can impute the censored Y observation first and then apply SIR directly to the imputed data. One possible imputation method is given in Fan and Gijbels(1995). While their method is effective for one or two regressors, it is not appropriate in the higher dimensional situation. A feasible alternative is to apply the dimension reduction method as outlined in this article first, and then apply imputation to the reduced variables. This prospect merits further study.

The proof of root n consistency as outlined in Appendix B can perhaps be improved with less strenuous assumptions. While this requires further theoretical investigation, it should not affect the applicability of the procedure proposed here.

Appendices

A. Derivation of (3.6).

It suffices to show that

$$S^o(t|\mathbf{x}) = exp\{E[\frac{1(t < Y, \delta = 1)}{S_Y(Y|\mathbf{x})}|\mathbf{x}]\}. \qquad (A.1)$$

First, the conditional independence assumption (1.3) implies that $S_Y(y|\mathbf{x}) = S^o(y|\mathbf{x})S_C(y|\mathbf{x})$, where $S_C(y|\mathbf{x}) = P\{C > y|\mathbf{x}\}$. Using this relationship, the expectation term in (A.1) can be written as

$$\begin{aligned} E\{\frac{1(t < Y, \delta = 1)}{S_Y(Y|\mathbf{x})}|\mathbf{x}\} &= E\{\frac{1(t < Y^o)1(Y^o < C)}{S^o(Y^o|\mathbf{x})S_C(Y^o|\mathbf{x})}|\mathbf{x}\} \\ &= E\{\frac{1(t < Y^o)}{S^o(Y^o|\mathbf{x})S_C(Y^o|\mathbf{x})}E(1(Y^o < C)|\mathbf{x}, Y^o)|\mathbf{x}\} \end{aligned}$$

By (1.3) again, we have $E(1(Y^o < C)|\mathbf{x}, Y^o) = S_C(Y^o|\mathbf{x})$. The last expression is seen to become

$$E\{\frac{1(t < Y^o)}{S^o(Y^o|\mathbf{x})}|\mathbf{x}\}$$

The rest of the derivation is straightforward from the relationship between the hazard and the survival functions. ◇

Chapter 13

PHD-aided Analysis of Variance for Two Level Factorial Designs

While the method of principal Hessian direction is primarily developed for dealing with continuous input variables, its application for data obtained by designed experiments is also noteworthy. We shall discuss the important case of two-level complete factorial designs. This chapter is taken largely from Filliben and Li(1997).

13.1 Traditional analysis for 2^p factorial designs.

The primary components of a traditional analysis consist of

(1). least squares estimated effects–ranked via magnitude;

(2). half-normal probability plot of estimated effects;

(3). residual standard deviation for various models;

(4). parsimonious fitted model.

The default general model for 2^p full factorial experiments is

$$Ey = \mu + \frac{1}{2}(\sum_{i}^{p} \alpha_i x_i + \sum_{i<j} \alpha_{ij} x_i x_j + \sum_{i<j<k} \alpha_{ijk} x_i x_j x_k + \cdots) \tag{13.1}$$

Here each factor x_i takes on the coded values +1 and -1, μ is the general additive constant, α_i are main effects, α_{ij} are two-term interactions, etc. The magnitude of α_i is the expected change in the response as we proceed from the -1 setting of factor i to the $+1$ setting of factor i. The advantage of this model is that (1). its coefficients have a natural interpretation (= expected change in the response); (2) the form for the maximum likelihood and least squares estimates of the coefficients is trivial (= difference of 2 averages); and (3) the prediction equation yields a perfect fit (at all 2^p sample points). In addition, all effects can be estimated by the Yates algorithm. The disadvantage of this model is that it has a large number (2^p) of terms, and so even for moderate p (such as $p = 10$), the model becomes unwieldy. Fortunately in practice, methods exist for pruning out most terms from the original default model so as to produce a more parsimonious and more manageable prediction equation.

13.1.1 Digital-to-analog converter error.

A p-bit DAC (digital to analog converter) (see, Soulders and Stenbakken 1985, Stenbakken, G. N. et. al. 1985, Stenbakken, G.N. and T. M. Souders 1987, Stenbakken, G.N., T. M. Souders, and G.W. Stewart 1989, Souders T.M., and G. N. Stenbakken 1990) is a device which has a total of 2^p code states, each code state being a combination of 0 (low level voltage) or 1 (high level voltage) for each of p bits. Such systems have a quite natural 2-level factorial design representation: the response variable Y being the converter's measured error (= the difference between the pre-conversion digital voltage and the post-conversion analog voltage at a given code state), the control factors being the individual p bits, and the factor levels being the individual bit settings (low and high).

The data set for DAC to be analyzed here came from a 2^{10} full factorial design. The sample size is of course $n = 2^{10} = 1024$. The response variable y is converter error–the difference between the pre-conversion digital voltage and the post-conversion analog voltage. The 10 factors are the 10 bits in the DAC.

Table 13.1 presents the ranked effects for the first 25 out of the 1023 estimated effects. Note that several of these large effects are 2-term and even 3-term interactions. The normal probability plot of all 1023 estimated effects is given in figure 13.1. It appears that all of these top 25 (and perhaps more) effects are significant. Any reasonably fitted models would thus include many terms. The presence of 2-way and 3-way interactions makes it hard to interpret the predicting equation.

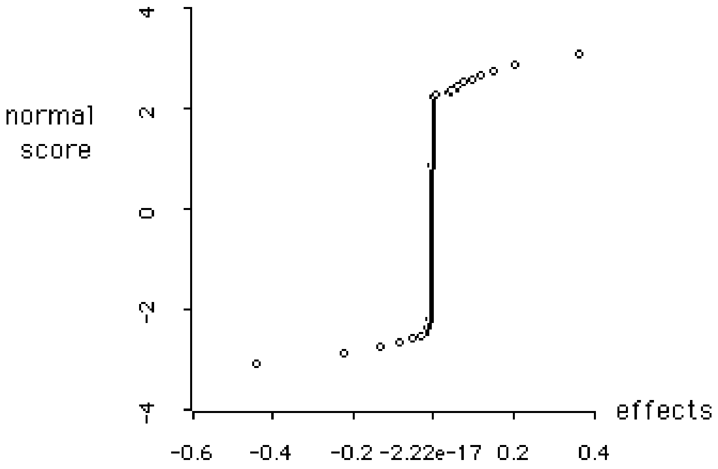


Figure 13.1: Normal plot of estimated effects for A-D cnverter

Table 1.1: Largest 25 effects from Yates analysis for the D/A converter.

Indentifier	Effect	T-value
Mean	- .24024	
9	- .87372	- 130.7
5	.72651	108.7
6	- .43429	-65.0
4	.41247	61.7
7	.30054	45.0
10	- .25472	-38.1
9,10	.23835	35.7
3	.20031	30.0
8,9,10	-.16659	-24.9
8,10	.16114	24.1
2	.12872	19.3
1	.12473	18.7
8,9	.09420	14.1
7,9,10	- .09387	-14.0
7,9	.09313	13.9
7,10	.08108	12.1
5,10	- .05616	-8.4
4,10	-.02798	-4.2
3,10	-.02303	-3.3
7,8,10	.02195	3.3
5,9	-.02097	-3.1

13.1.2 Effect abundance.

Research on methods of identifying significant effects in the analysis of factorial designs has been rekindled recently. Alternatives to the usual normal probability plot and half-normal probability plot have been proposed; e.g., Box and Meyer (1986), Lenth (1989), Dong (1992). When viewed as a regression model fitting/selection problem, then one could also include well-known criteria such as Mallows' C_p (Mallows 1973), cross-validation (Stone 1974, Geisser 1975), generalized cross-validation (Golub, Heath, and Wahba 1979), or the recent Bayesian method of Mitchell and Beauchamp (1988).

The aforementioned new procedures are based on the assumption of ”effect sparsity” (Box and Meyer 1986), namely that there are only a few non-zero main or interactions. Thus once these significant effects are found and the response function (= prediction equation) is constructed, we can easily provide answers to the typical subsequent questions: What is the predicted response at new level combinations? What are the optimal settings for achieving a specified goal? Where should we run our next experiment?

But the problem that we faced in the D-to-A converter example is quite the opposite of "effect sparsity". Since many interaction terms are non-zero, we have, by definition, "effect abundance". Effect abundance raises many important new questions.

The first group of questions concern the merit of the single prediction equation which presumably is valid for all level combinations. Can this global equation be improved upon? Are the predicted values equally good over the entire domain? How reliable are the results if in some applications, the equation is to be applied for interpolating and extrapolating to other regions of input variables? Are the coefficients in the prediction equation highly sensitive to which sub-domain of the data we are in? Are certain coefficients more stable, invariant, and robust over the entire data domain than others? Could improved (parsimonious and better-fitted) sub-models be derived for certain sub-domains? Would our insight be improved by replacing (or at least augmenting) the global prediction equation with a series of simpler local sub-domain prediction equations? If so, what is the least number of sub-domains that we could derive that would have linear (the ultimate in parsimony) sub-models.

The second group of questions concern the interpretability and readability of a final analysis from the scientist's point of view . The popularity of two level complete factorial designs is often attributed to its simplicity in interpreting the analysis results which are based on the notions of main effects and interactions of any order. Main effects require nothing but common sense to understand. The notions of interactions are still crystal clear in definition. One by one, they can be understood well by disciplined scientists so that at the end of analysis, a clear picture on how factors interact with each other in affecting the outcome variable can be formed in mind. But this is so, only when the number of interactions is small. Under effect abundance, will the picture still be as clear? If not, what additional efforts should be made? Are there any systematic ways in mapping out the relationship between the detected interactions? During such attempts, are some factors more important than others, at least for the purpose of reaching a parsimonious and crystal clear picture ?

Our methodology will answer these two groups of questions at the same time.

13.2 pHd for 2^p factorial designs

When applied to complete factorial designs with two levels, all three versions of pHd turns out to be identical. We need only to conduct the eigenvalue decomposition of $\hat{B}$, the p by p matrix of estimated two-term interactions :

$$\hat{B} = \frac{1}{2}\begin{bmatrix} 0 & \hat{\alpha}_{12} & \cdots & \cdots & \hat{\alpha}_{1p} \\ \hat{\alpha}_{21} & 0 & \hat{\alpha}_{23} & \cdots & \hat{\alpha}_{2p} \\ \cdots & \cdots & 0 & \cdots & \cdots \\ \cdots & \cdots & \cdots & 0 & \cdots \\ \hat{\alpha}_{p1} & \cdots & \cdots & \hat{\alpha}_{p\,(p-1)} & 0 \end{bmatrix} \tag{13.2}$$

In fact, it is easy to see that $\hat{B} = \Sigma_{yxx} = \Sigma_{rxx}$. For the $q-$based pHd, $\hat{B}$ is already equal to the matrix for the estimated quadratic terms. Note that for orthogonal designs (as is a 2^p full factorial), the covariance matrix of $\mathbf{x}$ is an identity matrix. Thus it can be ignored in implementing pHd.

Let $\hat{v}_1, \cdots, \hat{v}_p$ be the eigenvectors associated with the eigenvalues $\hat{\lambda}_1, \cdots, \hat{\lambda}_p$, in decreasing order of absolute value:

$$\hat{B}\hat{v}_i = \hat{\lambda}_i \hat{v}_i, \; i = 1, \cdots, p$$

The length of the eigenvectors are normalized to have unit length.

13.2.1 pHd plots.

The first few eigenvectors are important directions for studying the nonlinear structure in the data. In particular, we will plot the "main effects model" residual $\hat{r} = y - \hat{\mu} - \frac{1}{2}(\sum_{i=1}^{p} \hat{\alpha}_i x_i)$ against the projection along first two pHd directions, $\hat{v}_1'\mathbf{x}$, $\hat{v}_2'\mathbf{x}$. We will examine what PHD plots should look like for some specific cases.

Case 1: Underlying models with 2-term interactions.

First consider we may consider an intermediate model:

$$Ey = \mu + \frac{1}{2}(\sum_{i}^{p} \alpha_i x_i + \sum_{i<j} \alpha_{1j} x_1 x_j) \tag{13.3}$$

This model involves $2p$ coefficients: the additive constant + p main effects + the $(p-1)$ interactions. We shall refer to this model as the single source interaction model because all interactions are associated with factor 1, the common-source factor. In application, which factor plays the role of source factor has to be decided from the data; methods to be discussed later. The expectation of the matrix of two-factor interactions takes the form:

$$E\hat{B} = \frac{1}{2}\begin{bmatrix} 0 & \mathbf{a}_1{}' \\ \mathbf{a}_1 & \mathbf{O} \end{bmatrix} \tag{13.4}$$

where $\mathbf{a}_1$ denotes the column vector $(\alpha_{12}, \cdots, \alpha_{1p})'$, and $\mathbf{O}$ is a $p-1$ by $p-1$ matrix of zeros. In this case, there are only two nonzero eigenvalues, $\lambda_1 = \frac{1}{2}||\mathbf{a}_1||, \lambda_2 = -\frac{1}{2}||\mathbf{a}_1||$, with eigenvectors

$$v_1 = \frac{1}{\sqrt{2}}(1, \mathbf{a}_1'/||\mathbf{a}_1||)', \; v_2 = \frac{1}{\sqrt{2}}(1, -\mathbf{a}_1'/||\mathbf{a}_1||)'.$$

It follows that

$$E\hat{B} = \lambda_1 v_1 v_1' + \lambda_2 v_2 v_2'.$$

The key factor x_1 is in the projection of $\mathbf{x}$ along the direction of $v_1 + v_2 = (\sqrt{2}, 0, \cdots, 0)'$. This indicates that the pHd plot can be used to find the common-source-factor. For the ideal situation, the direction $v_1 - v_2$ reveals two perpendicular lines, Figure 13.2. This is an important characteristic for the single-source-interaction model.

Now consider the general model with all two-term interactions and main effects. The "main effects model" residual r has the expectation

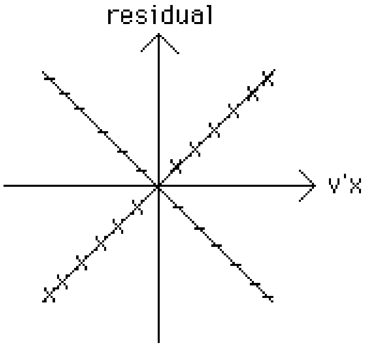


Figure 13.2: PHD plot for an ideal single interaction source model.

$$E\hat{r} = \sum_{i<j} \alpha_{ij} x_i x_j$$

The method of pHd explores the structure of this quadratic surface in the same way as in the canonical analysis of the quadratic response surface (e.g. Box and Draper 1987). $E\hat{r}$ can be written in terms of the eigenvectors and eigenvalues of $E\hat{B}$:

$$E\hat{r} = \sum_{i} \lambda_i (v_i'\mathbf{x})^2$$

A useful reduction of the model can be achieved if only a small number of eigenvalues λ_i are significant. In particular, if the leading two eigenvalues have much larger size than others, we anticipate to find clear quadratic patterns in the pHd plot. In fact, the case that there are only two nonzero eignvalues can be characterized by the model

$$Ey = \mu + \frac{1}{2}(\sum_{i} \alpha_i x_i + (\sum_{i\in I_1} c_i x_i)(\sum_{i\in I_2} d_i x_i)$$

where the first group I_1 of factors has no intersection with the second group I_2.

Example 13.1: Fracture Force. This is a small (n = 8) data set. A 2^3 full factorial design was conducted. The response variable y is the force (in pounds) required to fracture a connection joint consisting of a plate, and a nut-bolt assembly. The fracture force was dependent on the following 3 factors: plate surface preparation at the connection point, torque applied to the connecting nut, and size of the connecting nut. The data are given in table 13.2 below.

*** Insert table 13.2 here–Fracture Force Data (the file for this table is in the plot folder) ***

The matrix $\hat{B}$ is

$$\hat{B} = \begin{bmatrix} 0 & 6.75 & 17.75 \\ 6.75 & 0 & 1 \\ 17.75 & 1 & 0 \end{bmatrix}$$

Now applying pHd, we find the first two pHd directions :

$$\hat{v}_1 = (-.655, -.260, -.615)'$$
$$\hat{v}_2 = (-.667, .208, .622)'$$

The absolute values of the three eigenvalues are

$$19.33, 18.67 \text{ and } 0.66$$

The third value is much smaller, an indication that the third direction is not important.

The 3-D pHd plot is shown in Figure 13.3. From the figures, we can easily locate the two perpendicular line-clusters and attribute this dichotomy to factor 1. This shows that factor 1 (surface preparation) is indeed the optimal choice for further splitting. Further confirmation is seen by observing that $\hat{v}_1+\hat{v}_2 (= (-1.322, -.052, .007)')$ is approximately in the direction of $(1, 0, 0)'$ except for an opposite sign.

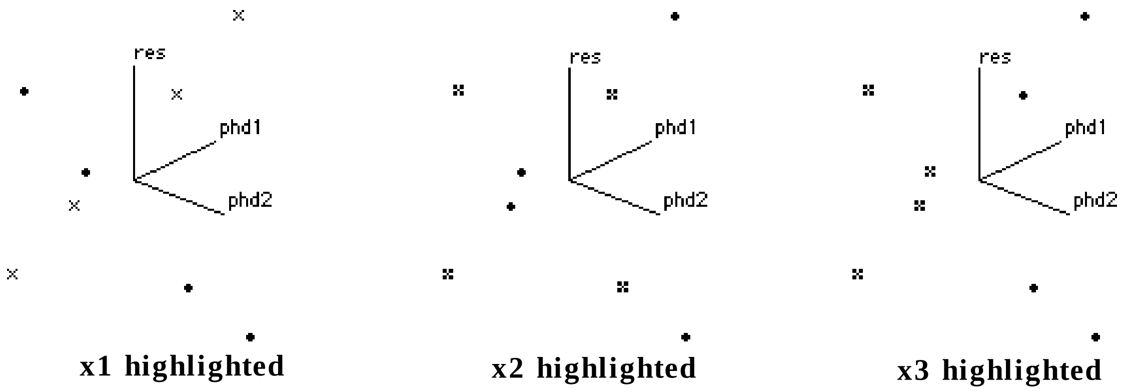


Figure 13.3: 3-D pHd plots(with each factor coded) Automotive fracture force example.

Example 13.2: Chemical Reactor Efficiency.

This example is taken from Box, Hunter, and Hunter (1978). A 2^5 complete factorial design was run. The response variable y is reactor efficiency (percent of chemical reaction completed). The 5 factors are feed rate, catalyst, agitation, temperature, and concentration. The data is from table 12.1a on page 377 of Box, Hunter, and Hunter. Figure 13.4 is the normal plot.

The matrix of two term interactions is given by

$$\hat{B} = \begin{bmatrix} 0 & 1.375 & .75 & .875 & .125 \\ 1.375 & 0 & .875 & 13.25 & 2.0 \\ .75 & .875 & 0 & 2.125 & .875 \\ .875 & 13.25 & 2.125 & 0 & -11.0 \\ .125 & 2.0 & .875 & -11.0 & 0 \end{bmatrix}$$

The pHd eigenvalues, in absolute value, are found to be

$$9.22,\ 8.22,\ 1.49,\ .34,\ .15,$$

Again, the leading two eigenvalues are much larger than others. The first two pHd eigenvectors are given by

$$\hat{v}_1 = (.07, -.54, -.08, .67, .46)'$$
$$\hat{v}_2 = (.01, .53, .10, .71, -.41)'$$

The pHd plot is given in Figure 13.5. Factor 4 is seen to be the source factor for causing the dichotomous grouping and the perpendicular line clusters. Further confirmation may come from the calculation that $\hat{v}_1 + \hat{v}_2$, is approximately in the direction of $(0, 0, 0, 1, 0)$. We use x_4 for splitting.

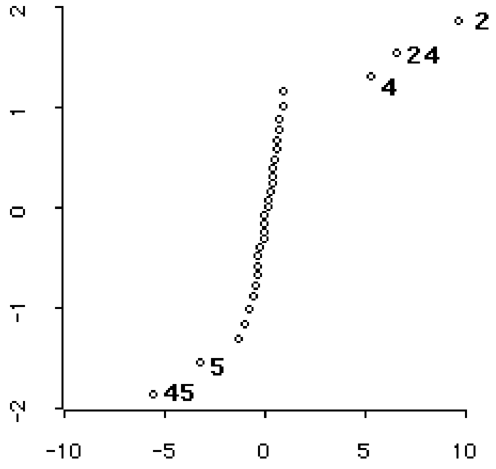


Figure 13.4: Normal plot for estimated effects in Chemical reactor example.

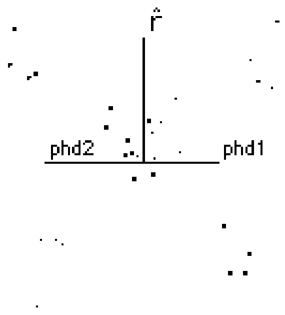


Figure 13.5: Line clusters found by rotating pHd plot in Chemical reactor example.

13.2.2 Underlying models with higher-order interactions.

The method of pHd does not use information from the higher order interaction terms. But this limitation can be somewhat lifted by recursive splitting. For example, suppose the response y can be written in terms of main effects, two-term interactions involving x_1, and three-term interactions involving x_1 and x_2:

$$Ey = \mu + \frac{1}{2}(\sum_{i=1}^{p} \alpha_i x_i + \sum_{i=2}^{p} \alpha_{1i} x_1 x_i + \sum_{i=3}^{p} \alpha_{12i} x_1 x_2 x_i)$$

The pHd method picks up only the directions to find x_1 and $a_1'\mathbf{x}$. But as to be shown in the following, we may ferret out three-term interactions by applying the pHd method twice.

First due to the presence of the third order interactions, the shape of the pHd plot might look different from the perpendicular linear pattern described earlier. But the crucial feature of two clusters due to x_1 can still be expected. This suggests us to split the data into two groups by x_1 as before. After splitting, we get

$$\begin{aligned} E\,y &= (\mu + \frac{1}{2}\alpha_1) + \frac{1}{2}\sum_{i=2}^{p}(\alpha_i + \alpha_{1i})x_i + \frac{1}{2}\sum_{i=3}^{p}\alpha_{12i}x_2x_i \text{ , for } x_1 = +1 \\ &= (\mu - \frac{1}{2}\alpha_1) + \frac{1}{2}\sum_{i=2}^{p}(\alpha_i - \alpha_{1i})x_i - \frac{1}{2}\sum_{i=3}^{p}\alpha_{12i}x_2x_i \text{ , for } x_1 = -1 \end{aligned}$$

Each split sample still forms a two-level factorial design but the number of factors has been decreased by 1. Each also follows the sole-source-interaction model, with $p-1$ main effects and $p-2$ two-term interactions involving only the factor x_2. Therefore, if we apply the analysis with pHd on each split sample, the role of x_2 can be revealed from each pHd plot. More discussion on recursive splitting is to be discussed later.

13.3 A flow chart for the pHd based partition .

The flow chart given by figure 13.6 illustrates our recursive process of analyzing the data and steps involved in growing the partition tree:

Step 1. Start with the top node of the tree which includes all data points. Apply Yates algorithm to the entire data set. Find the linear residuals by removing the mean and the main effects from each observed y. Run a pHd analysis. Locate important effects with normal or half-normal probability plots.

Termination. Compute the standard deviation for the linear residuals (residuals from the main-effect model). If it is smaller than a pre-specified value (an engineering cutoff point, for example), then stop and report the linear equation (based on significant main effects) as the prediction function; otherwise go to Step 2.

Step 2. Compute a crude estimate of σ by temporarily assuming all interactions involving three or more are negligible. Then use this information to evaluate the number of the significant pHd directions.

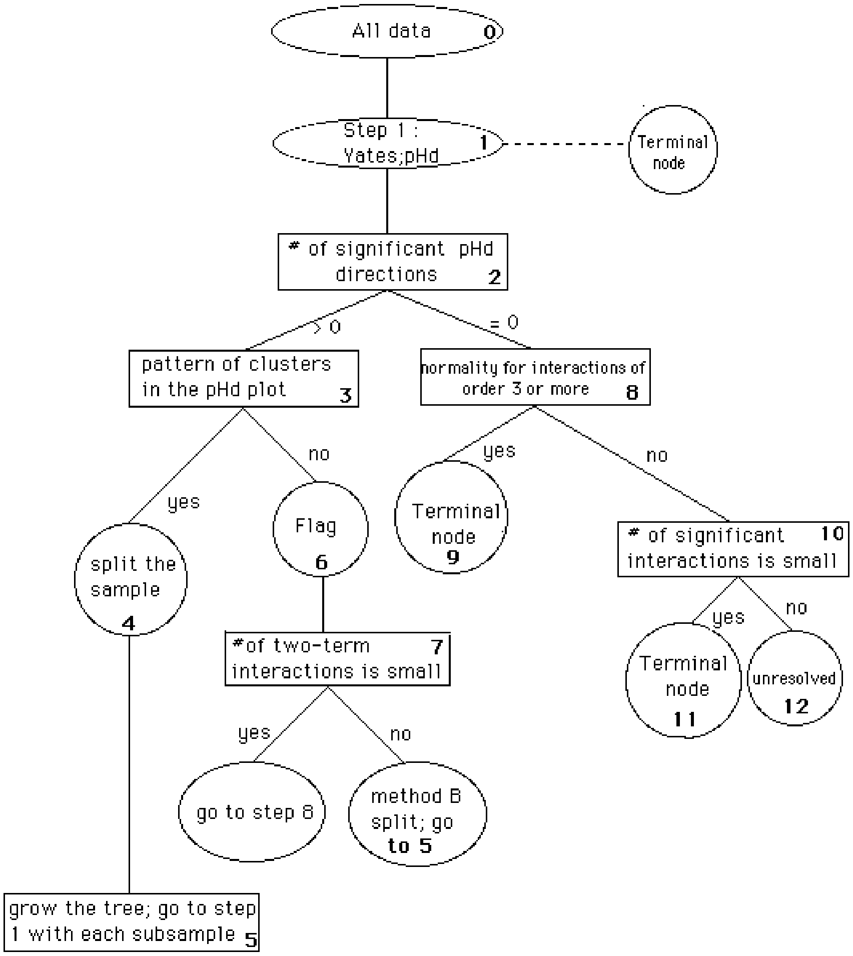

All data 0
Step 1 : Yates;pHd 1
Terminal node
# of significant pHd directions 2
> 0
= 0
pattern of clusters in the pHd plot 3
normality for interactions of order 3 or more 8
yes
no
split the sample 4
Flag 6
Terminal node 9
# of significant interactions is small 10
#of two-term interactions is small 7
Terminal node 11
unresolved 12
go to step 8
method B split; go to 5
grow the tree; go to step 1 with each subsample 5

Figure 13.6: General flow chart

Step 3. Examine the 3-D pHd plot and look for patterns of clustering.

Step 4. Find the factors that cause the clustering. Split the sample according to the factors found. Consult Section 9 for details of splitting.

Step 5. Grow children nodes. Each child node represents one split subsample. Now repeat the same procedure as described in Step 1 for each node. Continue the analysis with Steps 2,3,4, if applicable. When Step 5 is reached again, more children nodes are grown.

Step 6. Since no pattern of clustering is found, the choice on the splitting factor may be not be clear. Give a flag to point out the weakness in this portion of data. Then go to Step 7.

Step 7. Find the number of significant two-term interactions. If the number is large, then split the data by Method 9.4 and go to Step 5. Otherwise go to Step 8.

Step 8. Draw the normal and the half-normal probability plots for all interactions involving three or more factors.

Step 9. A terminal node is reached. If Step 8 is reached via Step 2, then only main effects are significant for this node. If Step 8 is reached via Step 7, then a small number of two-term interactions may also be significant. Report the linear fit (plus a small number of two-term interactions, if necessary) and the residual standard deviation.

Step 10. Find the number of significant interactions involving three or more terms.

Step 11. A terminal node is reached. Report the linear plus the significant interactions and construct the fitted equation based on them. Report also the standard deviation for the residuals after the fit.

Step 12. If there are still a few interactions involving three or more factors remaining, then we have an unresolved node.

13.4 Applications.

In this section, we discuss the results of applying the recursive partition method to earlier examples.

13.4.1 Tree for fracture force example.

For the fracture force example, we only need one splitting. The tree is given in figure 14. For each terminal node, the interaction term in the fitted equation may be treated as the random error.

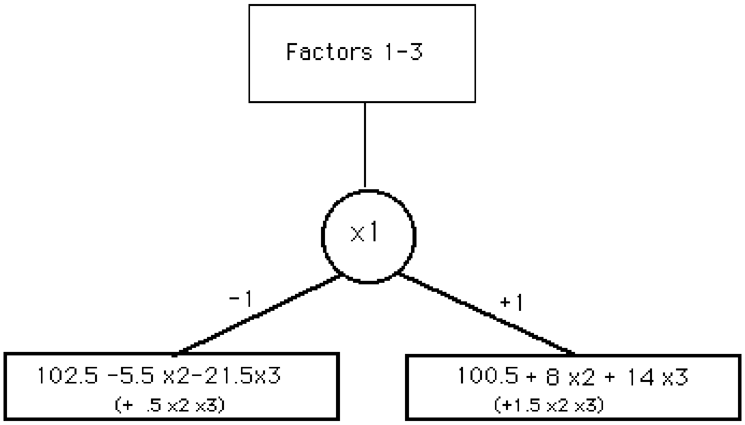


Figure 13.7: Tree for linear domain splitting. Automotive fracture force example.

13.4.2 Tree for chemical reactor efficiency example.

Consult figure 13.8. The starting node shows 5 factors to begin with. The standard deviation for residuals after linear fit is 10.1. Names of the leading interactions and their sizes (in parentheses) are given. The size is measured by the ratio of the effect estimate to the standard deviation of all interactions involving three or more terms. The first split is based on factor x_4, the temperature.

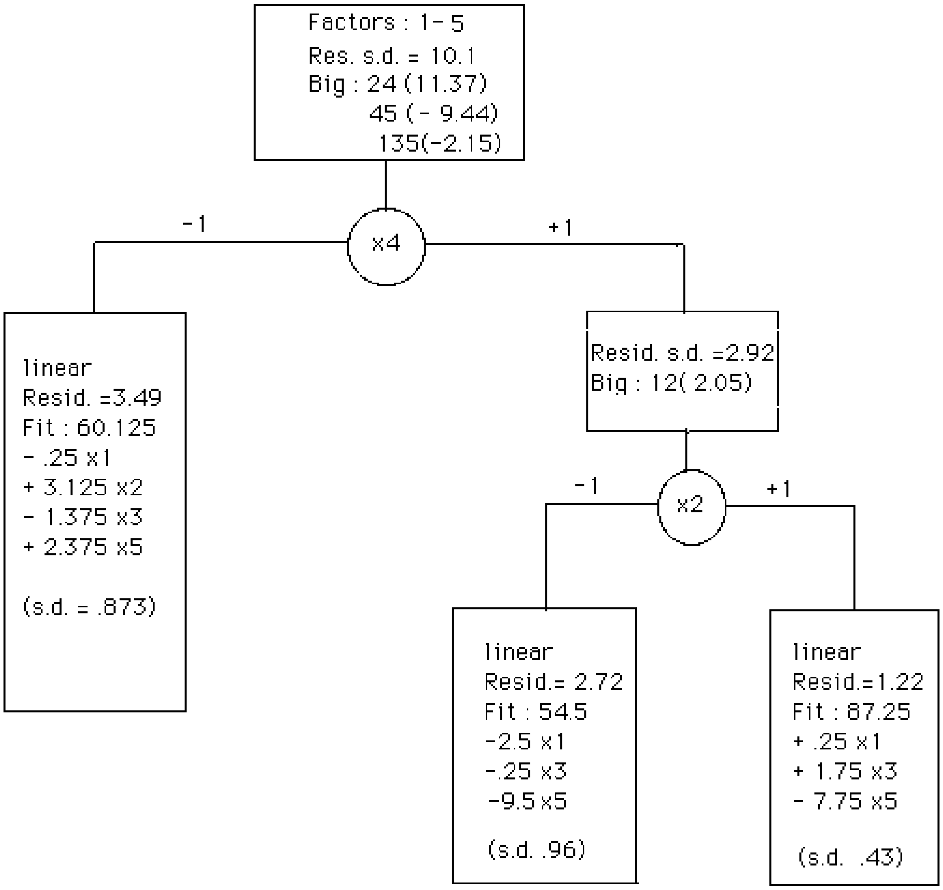


Figure 13.8: Tree for linear domain splitting. Chemical reactor example.

After splitting, consider the branch for $x_4 = -1$. The linear fit is now adequate for this subsample. The residual standard deviation is reduced to 3.49 and the fitted equation is given in the leftmost panel. The standard error for the estimated coefficient, given in parenthesis is .87.

Consider the other branch, $x_4 = 1$. The standard deviation for the ”main effects model” residual is 2.92, slightly smaller than that for $x_4 = -1$. One large interaction term, the interaction of factors 1 and 2, is found. The pHd plot shows line clusters again, suggesting further splitting by factor x_2.

After splitting by x_2, each subsample becomes linear. The residual standard deviation is smaller when x_2 is set at the high level.

We have seen that the output tree given in figure 15 provides us a clearer picture about structure of the data which can not be easily sorted out from a traditional analysis. We find that x_4, temperature, is the most crucial factor, to be followed by the factor x_2, catalyst. We can compare the response equations for different regions. For example, when the temperature is set at the low level, the case for the left-most panel, the mean response (60.125), is a lot smaller than the mean response (87.25) when setting both the temperature and the catalyst at high levels, the region corresponding to the right-most panel. At the latter setting, the effect of factor x_5, concentration is three times as strong as that for the former setting, and in opposite directions. The residual is also a lot smaller in the latter setting. If our goal is to find an optimal setting for maximizing the yield, then we are led to focus the search on the region defined by the right-most panel; namely by $x_4 = 1, x_2 = 1$. We may further set $x_5 = -1$, and $x_3 = 1$. The effect of x_1, feed rate, seems negligible, and can be set at a convenient level.

13.4.3 Tree for DAC error example.

We apply our method to the Digital-to-Analog converter data, which is much more complicated than the first two examples.

Figure 13.9 gives the partition tree. The bottom panels show the fitted equations in each terminal node. All of them are linear except for the first panel which also includes a statistical significant (but practically insignificant by the judgment of the engineer whom we consult with) two-factor interaction term. The order of factors for successive split largely coincides with the engineering intuition on the order of bit importance. The voltage reading in an A-to-D converter is equivalent to the binary coding for the associated bit combination. Thus x_{10} is the most important bit, and x_1 is the least important bit; miss-coding in bit x_{10} would result in an error in the magnitude of 2^9, compared to that of 2^3 for miss-coding in bit x_4, for example.

Splitting starts with x_{10}, as suggested by figure 13.10.

The right branch $x_{10} = 1$ is further split by by x_8 and x_9 simultaneously because pHd plot finds four clusters, figure 13.11.

After that, the response in each node becomes linear and no more splitting is necessary.

The left branch $x_{10} = -1$ is split by x_9. After splitting, nonlinearity still exists and the pHd plots , not reported here, suggest bit x_5, and bit x_8 respectively in each sub-branch. Continue the splitting with these factors, we end up with the four terminal panels on the left.

The role of x_5 is unexpected. It is used to split the leftmost branch of the tree before reaching the terminal nodes. It also has the largest main effect in each of the remaining terminal nodes. In addition, the nearly diminishing role of x_7 in the first two terminal panels also surprises our engineer. Finally, large scale off-line analysis like this may offer important guideline on how to take sample for on-line quality control. which forbids the extensive testing on all code states. For example, highly fractionated factorial design can be used in each panel where only main effects are present.

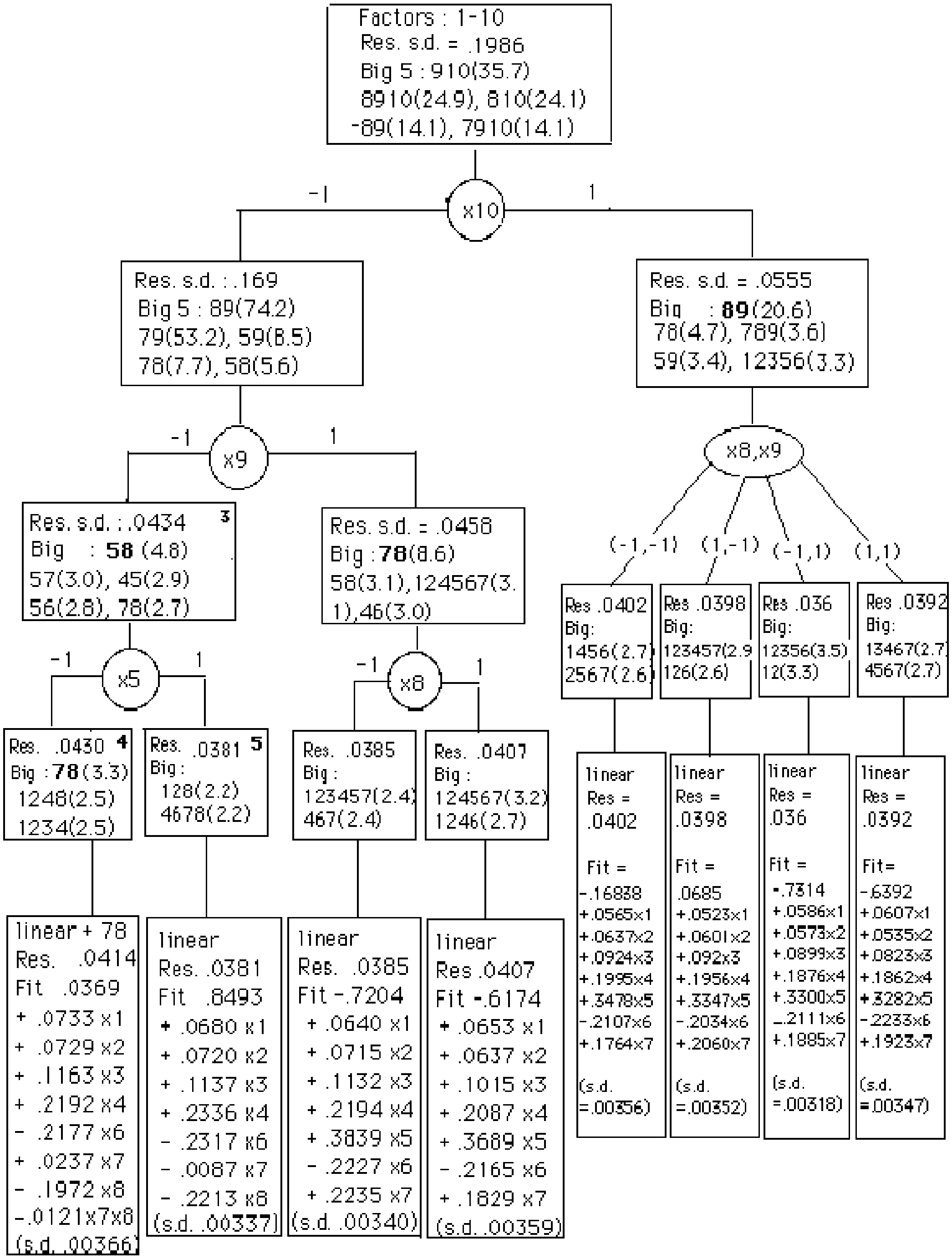


Figure 13.9: Tree for linear domain splitting. Digital-to-analog Converter.

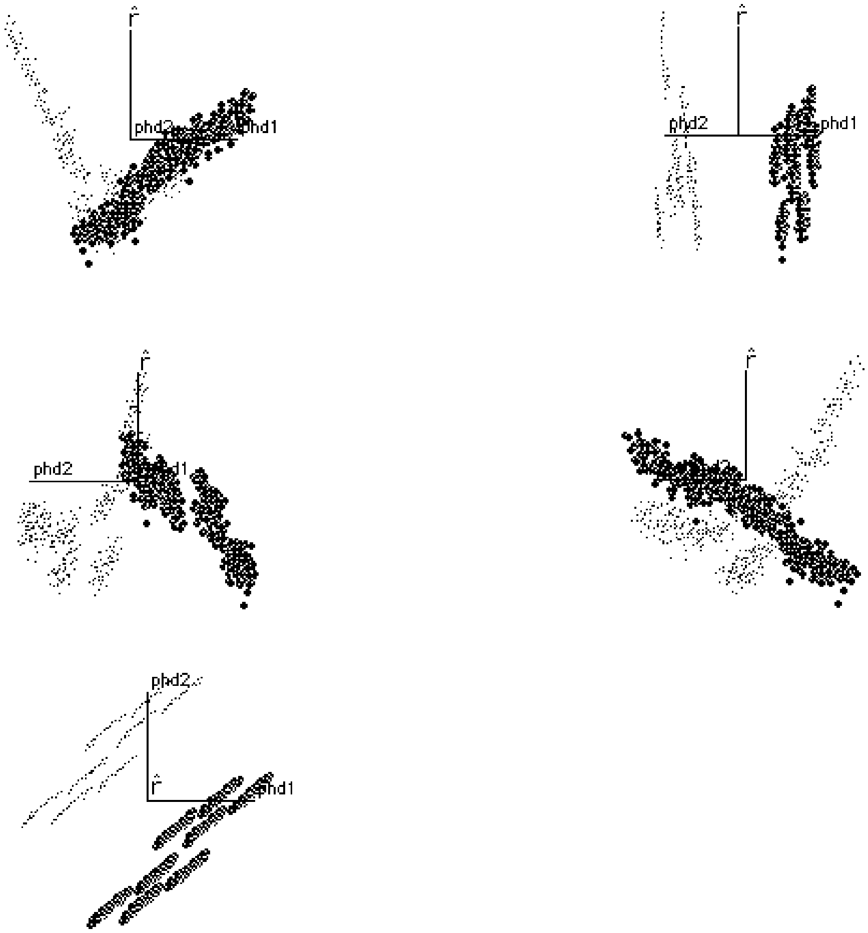


Figure 13.10: 3-D pHd plot(all data)(Coding based on factor 10) Digital-to-analog example.

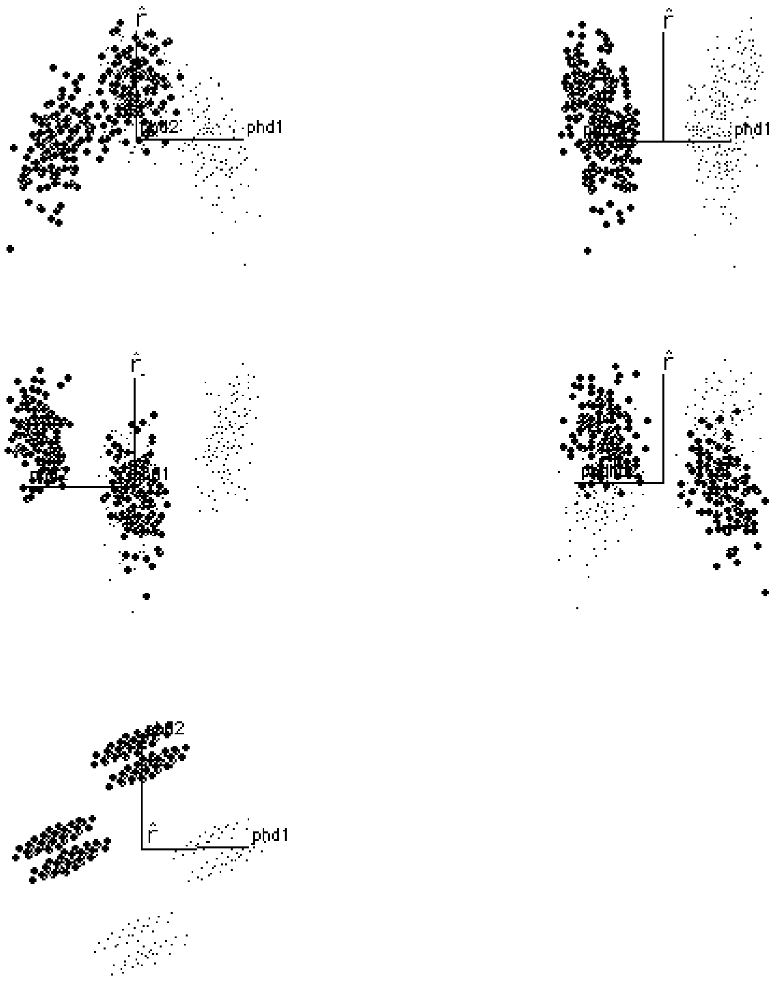


Figure 13.11: 3-D pHd plot(subset based on factor 10 $= +1$) (Coding based on factor 9) Digital-to-analog example.

Part II

Further Development- some unpublished manuscripts

Chapter 14

Generalizing Fisher's linear discriminant analysis via the SIR approach

This chapter is a minor modification of Chen and Li(1998).

Despite of the rich literature in discriminant analysis, this complicated subject remains much to be explored. In this chapter, we study the theoretical foundation that supports Fisher's linear discriminant analysis (LDA) by setting up the classification problem under the dimension reduction model (1.1) of chapter 1. Through the connection between SIR and LDA, our theory helps identify sources of strength and weakness in using CRIMCOORDS(Gnanadesikan 1977) as a graphical tool for displaying group separation patterns. This connection also leads to several ways of generalizing LDA for better exploration and exploitation of nonlinear data patterns.

14.1 Introduction.

Discriminant analysis aims at the classification of an object into one of K given classes based on information from a set of p predictor variables. Among the many available methods, the simplest and most popular approach is linear discriminant analysis (LDA).

A most well-known property for LDA is that LDA is a Bayes rule under a normality condition about the predictor distribution. More precisely, the condition requires that for the ith class, $i = 1, \cdots, K$, the p-dimensional predictor variable $\mathbf{x} = (x_1, \cdots, x_p)'$ follows a multi-variate normal distribution with mean μ_i and a common covariance Σ_c. Together with the prior probability π_i, $i = 1, ..., K$, about the relative occurrence frequency for each class, this basic normality assumption leads to a Bayes discriminant rule which coincides with the rule of LDA.

Another way of deriving LDA originates from the consideration about group separation when there are only two classes, $K = 2$ (Fisher 1936, 1938). The idea is to find a linear combination of the predictors , $z = a_1x_1 + \cdots, a_px_p$, that exhibits the largest difference in the group means relative to the within-group variance. The derived variate z is known as Fisher's discriminant function, or the first canonical variate. Fisher's result is further generalized by Rao(1952, Sec 9c) to the multiple class problem, $K \geq 2$. In general, after finding the first r canonical variates, the $(r+1)$th canonical variate is the next best linear combination z that can be obtained subject to the constraint that z must be uncorrelated to all canonical variates obtained earlier. Canonical variates are also referred to as the discriminant coordinates (CRIMCOORDS) in Gnanadesikan(1977).

Empirical evidence has shown that scatterplots of the first few CRIMCOORDS can reveal interesting clustering patterns. Such graphical displays are helpful in studying the degree and nature of class separation and for detecting possible outliers. However, the nonlinear patterns often observed in such plots also point to the limitation of the commonly-used normality assumption in justifying LDA. The data points within each class do not always appear elliptically distributed. Even if they do appear so, they hardly have the same orientation-violating the equal covariance assumption.

The motivation of our study stems from the concern about the theoretic foundation of LDA. To what extent, can LDA be applied effectively without the normality assumption? In what sense, can the reduction from the original p predictors to the first few CRIMCOORDS be deemed ”effective” ? Are there any other linear combinations more useful than the CRIMCOORDS in providing graphical information about group separation? If so, how can one find them? In this article, we address these issues by formulating the classification problems via the dimension reduction approach of Li(1991). A key notion in that article is the effective dimension reduction (*e.d.r.*) space for general regression problems.

This chapter is organized in the following way. In Section 2, we review the dimension reduction approach and bring out the connection of sliced inverse regression(SIR) with LDA. It turns out that the e.d.r. directions found by SIR are proportional to the vectors **a** used in the canonical variates. Via this connection, the theory of SIR is applied to offer a new theoretical support for using CRIMCOORDS.

Prior information about the occurrence frequency for each class plays a crucial role in discriminant analysis. It is certainly needed in forming a Bayes rule. But how critical is it for dimension reduction? This issue is discussed in Section 3. We argue that dimension reduction can be pursued independent of the specification of a prior distribution.

LDA can be viewed as a two-stage procedure. The first stage is to find the canonical variates for reducing the predictor dimension from p to K or less; the second stage is to split the canonical space linearly into K regions for class-membership prediction via the Mahalanobis distance. While the SIR theory justifies the use of canonical variates at the first stage, the theory itself does not support the use of linear split rules at the second stage. Section 4 discusses this issue. Nonparametric classification rules more effective than LDA can be formed using the first few canonical variates found at the first stage of LDA.

As is known, the first moment based SIR does not always work in finding the entire e.d.r. space. Knowledge about when SIR will fail helps identify sources of potential weakness in using CRIMCOORDS. An important special case is when there are only $K = 2$ classes. There is only one CRIMCOORD available now, no matter how complex the true dimension reduction model is. This may not be enough for locating the entire e.d.r. space because the e.d.r. space can have more than one dimension. In section 5, more general methods will be considered to help find more e.d.r. directions that cannot be found by SIR. There are two types of generalization. The first one follows the thoughts of Principal Hessian directions (PHD) (Li 1992a). It amounts to the comparison of the second moments of the predictors between classes. The second type of generalization explores an idea of double-slicing mentioned. Several simulation examples are provided and an application to a real data set is given.

Further discussion and some concluding remarks are given in Section 6.

14.2 SIR and Fisher's canonical variates.

In this section, the relationship between SIR and canonical variates is established first. Then the assumptions used to guarantee the success of SIR are discussed in the context of classification. These assumptions provide more general theoretical support for the use of canonical variates than the well-known normality assumption underlying LDA.

14.2.1 Connection.

Recall the dimension reduction model (1.1) from Chapter 1. For ease of presentation, let's rewrite it here:

$$Y = g(\beta_1'\mathbf{x}, \cdots, \beta_d'\mathbf{x}, \epsilon). \tag{2.1}$$

Here Y is the response variable, and g is an unknown function with $(d+1)$ arguments. Notice that we have changed the notation a little bit : d is used to replace K as the dimension of the e.d.r. space. This change is because K is reserved for the number of classes.

Recall the population version of SIR from Chapter 2 first. Denote the covariance matrix of $\mathbf{x}$ by $\Sigma_\mathbf{x}$. The central idea of SIR is to reverse the roles of $\mathbf{x}$ and Y. Instead of regressing Y on $\mathbf{x}$, we may consider the inverse regression curve $E(\mathbf{x}|Y) = (E(x_1|Y), \cdots, E(x_p|Y))'$. In general, this curve is in the p dimensional space. However, Theorem 3.1 of Li(1991) shows that under (2.1) and another condition to be discussed later, the inverse regression curve indeed falls into a d dimensional subspace. This subspace is determined only by the e.d.r. directions and $\Sigma_\mathbf{x}$. Denote the covariance matrix of the random vector $\eta = E(\mathbf{x}|Y)$ by $\Sigma_\eta = cov(\eta) = cov(E(\mathbf{x}|Y))$. We are led to the following eigenvalue decomposition for finding e.d.r. directions:

$$\begin{aligned} &\Sigma_\eta b_i = \lambda_i \Sigma_\mathbf{x} b_i \\ &\lambda_1 \geq \cdots \geq \lambda_p, \end{aligned} \tag{2.2}$$

Li's theorem implies that all but the first K eigenvalues must be zero and that the eigenvectors associated with nonzero eigenvalues are the e.d.r. directions.

The sample version of SIR is to substitute Σ_η and $\Sigma_\mathbf{x}$ in (2.2) by their estimates from an i.i.d. sample $(Y_i, \mathbf{x}_i), i = 1, \cdots, n$. The estimate of $\Sigma_\mathbf{x}$ is just the sample covariance $\hat{\Sigma}_\mathbf{x} = n^{-1}\sum_{i=1}^n(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$. Here $\bar{\mathbf{x}}$ denotes the sample mean. The estimate of Σ_η can be formed by first partitioning the response variable Y into H intervals, $I_h, h = 1, \cdots, H$. Within each slice, compute the mean of $\mathbf{x}$, $\hat{\mathbf{m}}_h = n_h^{-1}\sum_{Y_i \in I_h} \mathbf{x}_i$, where n_h is the number of cases in slice h. These slice means constitute a simple estimate of $E(\mathbf{x}|Y)$ and they can be combined to give a weighted covariance matrix, $\hat{\Sigma}_\eta = n^{-1}\sum_{j=1}^{H} n_j(\hat{\mathbf{m}}_j - \bar{\mathbf{x}})(\hat{\mathbf{m}}_j - \bar{\mathbf{x}})'$, for estimating Σ_η. The eigenvectors $\hat{b}_i$'s are the SIR directions and we shall call $\hat{b}_i'\mathbf{x}$ the SIR variates.

The examples and discussion in earlier chapters focuse on the case where the response variable Y is continuous. But the continuity of Y is not required in (2.1). In fact, when Y is discrete and can take only K distinct values, the slicing step of SIR is automatic for $H = K$. This special circumstance fits well into our classification problem. We can regard each $(\mathbf{x}_i, Y_i)$ as one case in the training set and the response variable Y_i is just the class label for that case. The slice mean $\hat{\mathbf{m}}_j$ corresponds to the vector of the predictor's mean for the jth group. The matrix $\hat{\Sigma}_\eta$ coincides with the between group variance-covariance matrix in one-way multivariate analysis of variance (MANOVA).

To elucidate how canonical variates are related to the e.d.r. directions found by SIR, recall that the first canonical variate is derived by maximizing the ratio of the between-group variance to the within-group variance. In our notation, for a linear combination $z = \mathbf{a}'\mathbf{x}$, the group means are just $\mathbf{a}'\hat{\mathbf{m}}_j$, $j = 1, \cdots, K$. The between-group variance, $n^{-1}\sum n_j(\mathbf{a}'\hat{\mathbf{m}}_j - \mathbf{a}'\bar{\mathbf{x}})^2$, can be written as $\mathbf{a}'\hat{\Sigma}_\eta\mathbf{a}$. On the other hand, the within-group variance can be written as $n^{-1}\sum_{i=1}^{n}(\mathbf{a}'\mathbf{x}_i - \mathbf{a}'\hat{\mathbf{m}}_{j(i)})^2 = \mathbf{a}'\hat{\Sigma}_e\mathbf{a}$, where the class membership for the i-th case is denoted by $j(i)$ and $\hat{\Sigma}_e$ is the within-group variance-covariance matrix. The first canonical variate is the linear combination of $\mathbf{x}$ formed by the vector $\mathbf{a}$ which solves the following maximization problem:

$$\max_{\mathbf{a}} \frac{\mathbf{a}'\hat{\Sigma}_\eta\mathbf{a}}{\mathbf{a}'\hat{\Sigma}_e\mathbf{a}}, \tag{2.3}$$

The solution of (2.3) is the largest eigenvector of the following eigenvalue decomposition:

$$\begin{aligned} \hat{\Sigma}_\eta\mathbf{a}_i &= \hat{\gamma}_i\hat{\Sigma}_e\mathbf{a}_i, \\ \hat{\gamma}_1 &\geq \hat{\gamma}_2 \geq \cdots \geq \hat{\gamma}_p \end{aligned} \tag{2.4}$$

To see the connection with SIR, we can rearrange the above eigenvalue decomposition equation by adding $\hat{\gamma}_i\hat{\Sigma}_\eta\mathbf{a}_i$ on both sides :

$$(1 + \hat{\gamma}_i)\hat{\Sigma}_\eta\mathbf{a}_i = \hat{\gamma}_i(\hat{\Sigma}_\eta + \hat{\Sigma}_e)\mathbf{a}_i$$

Now we can use the identity that the sum of the between-group variance and within-group variance equals the total variance, $\hat{\Sigma}_{\mathbf{x}} = \hat{\Sigma}_\eta + \hat{\Sigma}_e$, to obtain :

$$\hat{\Sigma}_\eta\mathbf{a}_i = \frac{\hat{\gamma}_i}{1 + \hat{\gamma}_i}\hat{\Sigma}_{\mathbf{x}}\mathbf{a}_i$$

Comparing this equation with the sample version of (2.2), we see that $\hat{\lambda}_i = \hat{\gamma}_i/(1 + \hat{\gamma}_i)$, and $\mathbf{a}_i \propto \hat{b}_i$. We now reach the following observation.

Observation I : *The SIR variates are the same as the canonical variates except for possible differences in scaling.*

Canonical variates are often associated with LDA, which can only be theoretically justified under the normality assumption :

$$\mathbf{x}|Y = j \sim N(\mu_j, \Sigma_c). \tag{2.6}$$

If we further assume that

$$\text{the vectors } \mu_j - \mu_1,\ \ j = 2, \cdots, K,\ \text{ spans a } d \text{ dimensional space,} \tag{2.7}$$

then the Bayes discriminant rule will depend on **x** only through the first d canonical variates. This is the traditional support for using only the first few significant canonical variates in applying LDA. But (2.6) is apparently too stringent. In fact, one can even argue that if the predictors' distribution is normal, then there won't be any interesting patterns to see in the CRIMCOORDS plots. Thus to fully justify the merit of CRIMCOORDS, we need to consider a different situation where CRIMCOORDS can serve as an effective way of conveying the importance informance in the predictors.

By relating the cannocial variates with SIR variates, Observation I brings in a very broad context for using CRIMCOORDS to reduce the dimension of the predictors. This is because SIR can be justified under much weaker conditions. We shall discuss these conditions next.

14.2.2 Condition (2.1).

SIR is founded on two assumptions. One of them is the dimension reduction model (2.1). A general comparison of (2.1) to (2.6)-(2.7) can be made more clear by re-formulating (2.1) from the inverse regression point of view. Put $B = (\beta_1, \cdots, \beta_k)$. (2.1) implies that the conditional density of Y given **x**, $f(Y|\mathbf{x})$ depends only on $B'\mathbf{x}$; $f(Y|\mathbf{x}) = f(Y|B'\mathbf{x})$. Thus the conditional density of **x** given Y can be written as

$$\begin{aligned} f(\mathbf{x}|Y) &= \frac{f(Y|\mathbf{x})\, f_{\mathbf{x}}(\mathbf{x})}{f_Y(Y)} = \frac{f(Y|B'\mathbf{x})\, f_{\mathbf{x}}(\mathbf{x})}{f_Y(Y)} \\ &= \frac{f(Y, B'\mathbf{x})\, f_{\mathbf{x}}(\mathbf{x})}{f_Y(Y)\, f_{B'\mathbf{x}}(B'\mathbf{x})} = f(B'\mathbf{x}|Y)\frac{f_{\mathbf{x}}(\mathbf{x})}{f_{B'\mathbf{x}}(B'\mathbf{x})} \end{aligned} \tag{2.8}$$

Here all f with subscripts are marginal density functions.

For classification problems, the rightmost side in the expression (2.8) gives a useful factorization for comparing the predictor distributions in different classes. This can be summarized by the following statement:

Observation II. *For classification problems, (2.1) is equivalent to the condition that for any two classes, j and j', the ratio of their density functions of* **x** *depends only on $B'\mathbf{x}$:*

$$\frac{f(\mathbf{x}|Y = j)}{f(\mathbf{x}|Y = j') = \frac{f(B'\mathbf{x}|Y=j)}{f(B'\mathbf{x}|Y=j')}} \tag{2.9}$$

It is straightfoward to verify that (2.6) and (2.7) imply (2.9) if we take $\beta_1, \cdots, \beta_d$ to be any basis of the space spanned by the differences in μ_i's.

14.2.3 Condition on the predictor distribution.

Recall that in addition to (2.1) (or equivalently (2.9) for classification problems), SIR requires another condition on the distribution of $\mathbf{x}$: **(L.D.C)**: for any $b \in R^p$,

$$\text{the conditional expectation } E(b'\mathbf{x}|\beta_1'\mathbf{x}, \cdots, \beta_d'\mathbf{x}) \text{ is linear} \tag{2.10}$$

(2.10) is the same as the condition that for any variate $\mathbf{a}'\mathbf{x}$,

$$cov(\mathbf{a}'\mathbf{x}, B'\mathbf{x}) = 0 \text{ implies } E(\mathbf{a}'\mathbf{x}|B'\mathbf{x}) = \mathbf{a}'E\mathbf{x}, \tag{2.11}$$

(2.11) is much weaker than (2.6)-(2.7). Normality assumption is not needed here. Within group-covariances also need not be entirely the same.

As we have known before, one sufficient condition for (2.10) (or equivalently (2.11)) to hold is that

$$\mathbf{x} \text{ follows an elliptically-contoured distribution.} \tag{2.12}$$

But this often leads to the impression that (2.12) is equivalent to (2.10). A counter-example to this impression is indeed the normal model, (2.6) and (2.7). As a mixture of normal distributions, the marginal distribution of $\mathbf{x}$ certainly cannot be elliptically symmetric. As we have seen in Chapter 8, this predictor distribution condition is not too restrictive.

Remark 2.1. SIR variates are scaled to have unitary variance but canonical variates are usually scaled to have unitary *within-group* variance. Since the covariance is no longer the same for every group, we prefer the way SIR variates are scaled.

14.3 Prior distribution and dimension reduction.

The discussion in Section 2 assumes that the training set consists of $i.i.d$ observations from the same population as the target population where the test set will come from. This may not be the case in some applications. This section discusses the case that the training sample is obtained by stratified sampling. More specifically, a pre-specified number n_j of cases are drawn independently from each class j. The sampling allocation n_j/n does not always match the prior $\pi_j(= P\{Y = j\})$, the probability that a random test case from the target population falls into group j. Recall that under the 0-1 loss, the Bayes rule classifies a future observation by

$$\max_y \pi_y f_{\mathbf{x}|Y}(\mathbf{x}|y). \tag{3.1}$$

Now suppose the target population follows a dimension reduction model (2.1), or equivalently (2.9). We can translate (3.1) into

$$\max_y \pi_y f(B'\mathbf{x}|y). \tag{3.2}$$

This shows that in order to find the Bayes rule, we only have to focus on the $e.d.r.$ variates.

The next question is whether SIR is still applicable for finding the e.d.r. space under stratified sampling. To answer this question, we study the population version of SIR by letting n_j tend to the infinity; while fixing $p_j = n_j/n$. We notice that SIR takes the same form as (2.2) but with a slightly different interpretation about the two covariance operators. By fixing $p_j = n_j/n$ and Σ_η is still the between group variance-covariance matrix as in the one-way MANOVA with the weight for group j being p_j instead of π_j. Similarly, $\Sigma_{\mathbf{x}}$ is the overall sample covariance of $\mathbf{x}$.

Theorem 3.1. *Suppose the sample is drawn by stratified sampling. Then under (2.9) and (2.11), the eigenvectors with nonzero eigenvalues in the eigenvalue decomposition(2.2) fall into the e.d.r. space.*

Proof. *From (2.11), we see that for any* $\mathbf{a}$ *such that* $\mathbf{a}'\Sigma_{\mathbf{x}}B = 0$*, we must have* $\mathbf{a}'\Sigma_\eta\mathbf{a} = 0$*, or equivalently* $\Sigma_\eta\mathbf{a} = 0$*. This shows that the eigenspace for (2.2) associated with the zero eigenvalue must contain any such vector* $\mathbf{a}$*. Since all non-zero eigenvectors* b_j *must be orthogonal to* $\mathbf{a}$ *with respect to* $\Sigma_{\mathbf{x}}$*, they must fall into the column space of* B*. The theorem is now proved.*

14.4 Nonparametric regression after SIR.

Observation I, Observation II and Theorem 3.1 provide a general theoretical foundation for LDA. But this only justifies the first stage of LDA, namely using the canonical covariates to reduce the dimension. The further use of linear split rule can only be justified under normality assumption on the distributions for the e.d.r. variates are completely arbitrary. Without the normality assumption, it is only natural to apply nonparametric density estimation techniques after dimension reduction. For illustration, we shall discuss only the standard kernel estimation here. Other nonparametric procedures can similarly be applied.

Let $\mathbf{x}_{yi}, i = 1, \cdots, n_y$ be the sample drawn from class $Y = y$. The SIR directions, $\hat{b}_1, \cdots, \hat{b}_d$, converge to $b_1, \cdots, b_d$ respectively at the usual root n rate, provided that all d nonzero eigenvalues are distinct. The kernel estimate of the density function of $B'\mathbf{x}$ for class $Y = y$ takes the following form:

$$\hat{f}_{B'\mathbf{x}}(t_1, \cdots, t_d) = \frac{1}{nh^d}\sum_{i=1}^{n_y}\Pi_{j=1}^{d}\mathcal{K}(\frac{\hat{b}_j'\mathbf{x}_{yi} - t_j}{h}), \tag{4.1}$$

where the kernel $\mathcal{K}(\cdot)$ is a one-dimensional density function. The bandwidth h has to converge to 0 at an appropriate rate.

(4.1) can be compared to the ”theoretical” kernel density estimate, should we be given B exactly:

$$\tilde{f}_{B'\mathbf{x}}(t_1, \cdots, t_d) = \frac{1}{nh^p}\sum_{i=1}^{n_y}\Pi_{j=1}^{k}\mathcal{K}(\frac{b_j'\mathbf{x}_{yi} - t_j}{h}). \tag{4.2}$$

The consistency of (4.2) for estimating $f_{B'\mathbf{x}}(t_1, \cdots, t_d)$ is the subject of standard kernel density estimation. This allows us to conclude that the discriminant rule obtained by substituting $f(B'\mathbf{x}|y)$ in (3.2) by the kernel estimate (4.1) is asymptotically Bayes.

Example 4.1 Wave recognition. This example is taken from Breiman et al. (1984, pp 49-55); see also Loh and Vanichsetakul (1988). There are three classes and 21 variables.

Three triangular basic waveforms $w_1(\cdot)$, $w_2(\cdot)$, $w_3(\cdot)$, are involved: for $j = 1, \cdots, 21$,

$$w_1(j) = max(6 - |j - 11|, 0); \quad w_2(i) = w_1(j - 4); \quad w_3(j) = w_1(j + 4). \tag{4.3}$$

Each class is a random convex combination of two basic waveforms with noise added. Let $\mathbf{w}_i = (w_i(1), \cdots, w_i(21))', i = 1, 2, 3,$ and u_1, u_2, u_3 be independent random variables uniformly distributed on [0, 1]. The predictor $\mathbf{x}$ is generated by

$$\begin{aligned} \mathbf{x} &= u_1\mathbf{w}_1 + (1 - u_1)\mathbf{w}_2 + \epsilon, \ \text{for}\, Y = 1 \\ &= u_2\mathbf{w}_2 + (1 - u_2)\mathbf{w}_3 + \epsilon, \ \text{for}\, Y = 2 \\ &= u_3\mathbf{w}_3 + (1 - u_3)\mathbf{w}_1 + \epsilon, \ \text{for}\, Y = 3, \end{aligned} \tag{4.4}$$

where ϵ follows the standard normal distribution.

The vector space parallel to the three-dimensional hyperplane spanned by $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ is the e.d.r. space. This can be seen by verifying (2.9).

Now generate 200 cases from each group. Then SIR is applied. The eigenvalues are (0.651, 0.546, 0, $\cdots$). Kernel estimation is applied. Figures 4.1(a)-(b) show the Bayes rules with $\pi_y = 1/3$ and $\pi_y = y/6$ respectively. Classification boundaries are seen to be approximately linear. This is as expected. In fact, SIR variates for the population version can be represented by mixtures of normals with means being on a equilateral triangular, Figure 4.1(c). By a geometric argument, we can show that the contours for the likelihood ratios must be straight lines.

Another interesting feature about this example is that the e.d.r. space does not depend on the distribution of $u_y, \quad y = 1, 2, 3$. We generate another 200 cases from each group but with u_i from the density $f(u) = 3u^2$ for $u \in [0, 1]$. Apply SIR and kernel estimation again. For equal prior $\pi_y = 1/3$, the result is shown in Figure 4.1(d). Now the Bayes rules are nonlinear.

14.5 Other SIR type methods for dimension reduction.

SIR may only recover part of the e.d.r. space if the dimension of the hyperplane spanned by the group means $E(\mathbf{x}|y)$ is less than the dimension of the e.d.r. space d. When this happens, other SIR type methods can help find more e.d.r. directions that cannot be found by using CRIMCOORDS.

14.5.1 SIR-II.

In our context, SIR-II explores the variation in the group covariance matrices. Let $\Sigma_a = E[Cov(\mathbf{x}|Y)]$ be the average of the group covariance matrices. Define

$$\Sigma_{II} = E\{[Cov(\mathbf{x}|Y) - \Sigma_a]\Sigma_{\mathbf{x}}^{-1}[Cov(\mathbf{x}|Y) - \Sigma_a]\}. \tag{5.1}$$

Then the eigenvalue decomposition for SIR-II is

$$\begin{aligned} \Sigma_{II} c_i &= \gamma_i \Sigma_{\mathbf{x}} c_i \\ \gamma_1 &\geq \cdots \geq \gamma_p. \end{aligned}$$

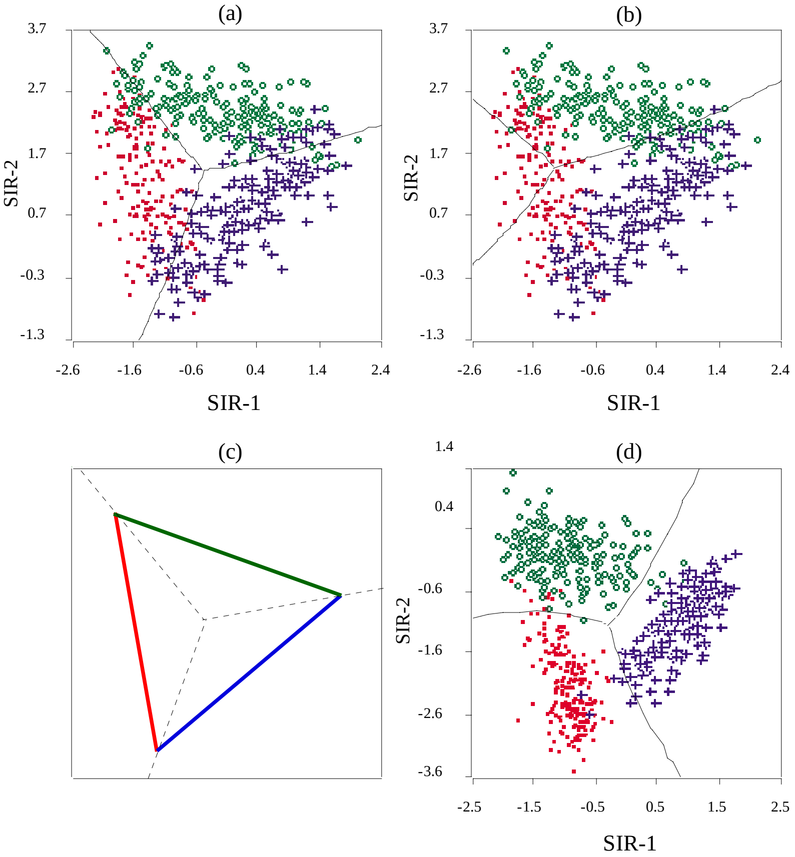


Figure 4.1: Wave Recognition Problem:
(a). SIR's View with Equal Contour Boundary, $\pi_y = \frac{1}{3}$;
(b). SIR's View with Equal Contour Boundary, $\pi_y = \frac{y}{6}$;
(c). SIR Variates for the Population Version;
(d). SIR's View with Equal Contour Boundary, $(\pi_y = \frac{1}{3}, u_i - f(u) = 3u, u \in [0, 1])$

The insertion of the matrix $\Sigma_{\mathbf{x}}^{-1}$ in the construction of Σ_{II} is to assure the affine invariance of the SIR-II procedure.

Compared with SIR, a condition stronger than (2.11) is required for SIR-II to find e.d.r. directions: for any variable $\mathbf{a}'\mathbf{x}$,

$$cov(\mathbf{a}'\mathbf{x}, B'\mathbf{x}) = 0, \text{ implies that } \mathbf{a}'\mathbf{x} \text{ is independent of } B'\mathbf{x}. \tag{5.2}$$

Thus the covariance of $(B'\mathbf{x}, \mathbf{a}'\mathbf{x})$ for each group $Y = y$ takes a diagonal partition: $cov[(B'\mathbf{x}, \mathbf{a}'\mathbf{x})|Y = y] = 0$. The first diagonal $cov(B'\mathbf{x}|Y = y)$ depends on y, but the second one does not: $cov(\mathbf{a}'\mathbf{x}|Y = y) = cov(\mathbf{a}'\mathbf{x})$. This implies that $\Sigma_{II}\mathbf{a} = 0$. The **a**must be in the eigenspace with zero eigenvalue. Now it is clear that like SIR, SIR-II can find e.d.r. directions.

Theorem 5.1. *Under dimension reduction framework, (2.10) (or (2.1)), if (5.2) holds, then there are at most a nonzero eigenvalues in (5.1) and the corresponding eigenvectors are in the e.d.r. space.*

However, under the weaker condition (2.11), we can only conclude that some of the eigenvectors with nonzero eigenvalues might be in the e.d.r. space. If none of them are in the e.d.r. space, then the e.d.r. space must be contained in the eigenspace with zero eigenvalue.

Example 5.1 Spherical Distribution Problem

This example comes from Friedman (1977). There are two classes and ten variables with the following distributions:

Group $Y = 1$: (1) $x_1, \cdots, x_d$ distributed uniformly within a d-dimensional spherical slab centered at the origin, with inner radius r_1 and outer radius r_2; (2) $x_{d+1}, \cdots, x_{10}$ are independent $N(0, 1)$.

Group $Y = 2$: $(x_1, \cdots, x_{10})$ is a 10-dimensional multivariate normal centered at the origin, with identity covariance matrix.

The last $10 - d$ variables are just noise. Because of the perfect symmetric pattern, SIR fails to find the $e.d.r.$ directions, but SIRII does a good job. For $d = 2$, $r_1 = 3.5$, $r_2 = 4.0$ and $n_1 = n_2 = 200$ cases, best view of x_2 against x_1 shows that the first class almost completely surrounds the second; Figure 5.1(a). Figure 5.1(b) is the SIRII view of the first two directions being found, which also illustrates the equal-contour line of the two densities which can be used as the boundary classifier for classification of future observations. The eigenvalues for this example are (0.613, 0.526, 0.063, 0.031, $\cdots$).

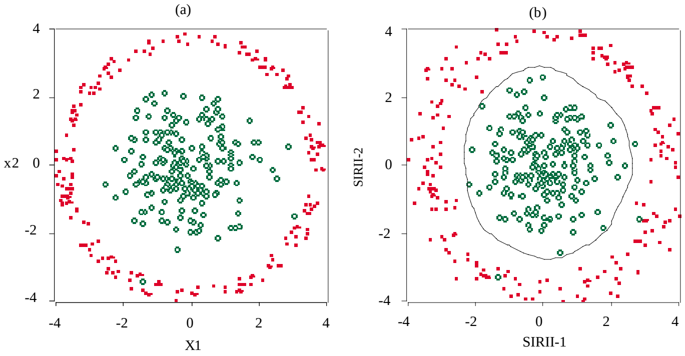


Figure 5.1: Spherical Distribution Problem: (a)Best View; (b)SIRII's View with Equal-Contour Boundary.

14.5.2 Double slicing.

We begin with the binary case K=2. SIR-I can only find at most one e.d.r. direction. The rest of them may have to rely on the 2nd moment based methods to recover. But as the following example shows, this may not be enough.

Example 5.2 Tai Chi. Consider Figure 5.2a, the well-known Tai Chi figure in the Asian culture. The regions are in black and in white called Ying and Yang respectively. The concepts of Yin and Yang and the Five Agents provide the intellectual framework for much of ancient Chinese scientific development especially in fields like biology and medicine (Ebrey 1993). Illustrated by the Tai Chi figure, the exercise of Yin and Yang, which is originated from Tai Chi, is the foundation of the entire universe. Yang, the element of light and life (in white), and Yin, the element of darkness and death (in black), are the most natural and original type of classes. Originated from each other, they represent all possible kinds of opposite forces and creatures, yet Yin and Yang work with each other in proper harmony. This makes the understanding and separation of them a difficult binary discriminant problem in real life applications.

The basic structure of Tai Chi is formed by drawing one large circle, two medium half circles and two small circles. The two small Yin and Yang circles located at the centers of the Yang and Yin half circles which are tangent to each other and also to the large circle.

We set up the model as follows:

(1). x_1 and x_2 distributed uniformly from within the large circle. We then assign the class label $Y = 1$ and $Y = 2$ to the points located in the Yin region and the Yang region accordingly.

(2). $x_3, \cdots, x_p$ are independent $N(0, 1)$.

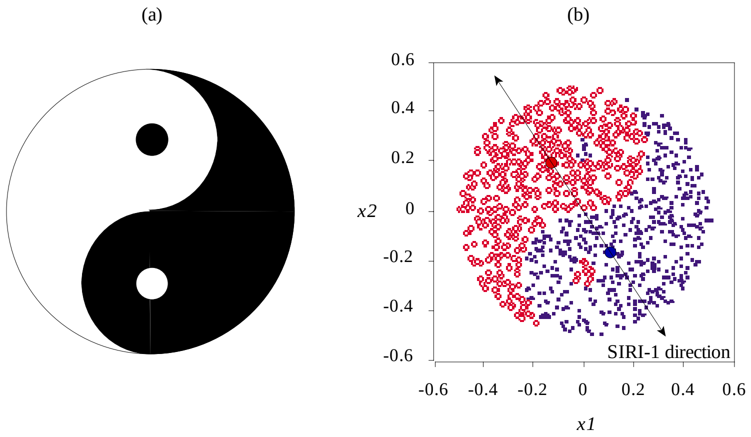


Figure 5.2: Tai Chi Example:
(a). The Tai Chi Model with Yin and Yang Classed;
(b). Simulation of Tai Chi Model with 1000 Observations and the SIRI direction.

For this model, SIR-I can only find a single e.d.r. direction which pass through the mass centers of Yin and Yang regions (Figure 5.2b). But SIR-II can not identify any e.d.r. direction. This is because the Yin and the Yang regions are anti-symmetry to each other implying that covariance matrix of (x_1, x_2) for Y=1 is the same as that for Y=2. One simple way to find the second direction necessary for completing the e.d.r. space is to slice the joint space of the direction identified by SIR-I and the class label Y.

In general, suppose that an e.d.r. direction b_0 is already obtained. We can take $\Sigma_\eta = cov(E(\mathrm{x}|b_0^{'}\mathrm{x}, Y))$ and conduct the eigenvalue decomposition (2.1). Under the same condition as SIR, the eigenvectors with nonzero eigenvalues can be shown to fall into the e.d.r. space. This suggests a double slicing procedure:

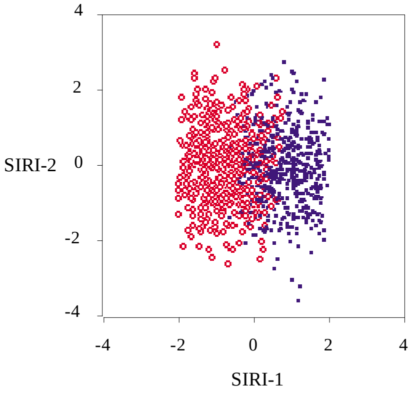


Figure 5.3: SIR's View for the Tai Chi problem without Double Slicing.

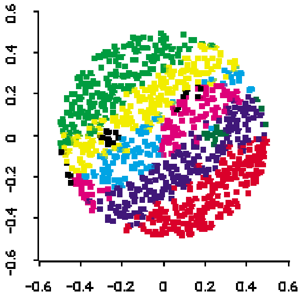


Figure 5.4: Four possible combinations of double slicing for the Tai Chi example:
(a). Y as First Direction and SIRI-1 as Second Direction with Equal Number of Observations
(b). SIRI-1 as First Direction and Y as Second Direction with Equal Number of Observations
(c). Y as First Direction and SIRI-1 as Second Direction with Equal Length of Intervals
(d). SIRI-1 as First Direction and Y as Second Direction with Equal Length of Intervals

Example 5.2 (continued) For $p = 6$, we simulated 1000 i.i.d. cases of (x, Y) using the model specified. Result of SIR-I and SIR-II without double slicing is in Figure 5.3 and Table 5.1. Figure 5.4 shows four different possible combinations of double slicing using SIR-I and the class label Y. In Figure 5.5 and Table 5.2, the double-sliced SIR-I and SIR-II clearly have recovered the complete e.d.r space for the Tai Chi structure.

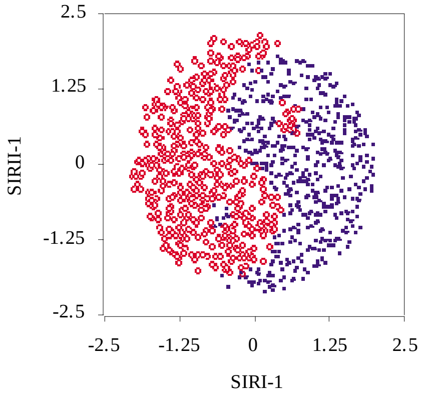


Figure 5.5: SIR's View for the Tai Chi problem with Double Slicing.

Table 5.1: Eigenvalues of SIRI and SIRII for the Tai Chi problem without Double Slicing.

SIRI	(.607 .000 .000 .000 .000 .000)
SIRII	(.016 .006 .002 .001 .000 .000)

Example 5.3 The Twist Problem. This example was first introduced as a clustering problem by Koontz and Fukunaga (1972), see also Koontz et. al. (1975), and Fukunaga (1990). There are two C-shaped trigonometric curves with random normal noise tangle with each other in a two-dimensional space (Figure 5.6a). There are two classes, one for each curve:

Group $Y = 1$:

(1). $x_1 = 20\cos\theta + z_1,$

$x_2 = 20\sin\theta + z_2,$

where, z_1, z_2 are independent $N(0, 1)$ and θ is $N(\pi, (0.25\pi)^2)$.

(2). $x_3, \cdots, x_p$ are independent $N(0, 1)$.

Group $Y = 2$:

(1). $x_1 = 20\cos\theta + z_1,$

$x_2 = 20\sin\theta - 20 + z_2,$

where, z_1, z_2 are independent $N(0, 1)$ and θ is $N(0, (0.25\pi)^2)$.

(2). $x_3, \cdots, x_p$ are independent $N(0, 1)$.

Since the first two variables are correlated through the structure of θ, SIR-I direction will not pass through the mass centers of these two curves, see Figure 5.6a. It can be verified that $E(x_1|Y = 1) = -14.6921$ and $E(x_1|Y = 2) = 14.6921$. It is easy to see that $E(x_2|Y = 1) = 0$ and $E(x_2|Y = 2) = 20$. Can SIR-II help in identifying the second e.d.r. direction necessary for obtaining the complete e.d.r. space?

For $p = 10$, we generate 300 cases from each class. Since the scale of the first two variables are too large relative to the rest of the noise variables, we further standardize these

Table 5.2: Eigenvalues and eigenvectors of SIRI (a) and SIRII (b) for the Tai Chi problem with Double Slicing.

(a) SIRI

eigenvalues	(.987 .027 .019 .013 .006 .001)
1st eigenvector	(-2.299 3.269 -0.005 -0.011 -0.021 -0.034)
2nd eigenvector	(3.088 2.260 0.039 0.123 0.024 0.312)

(a) SIRII

eigenvalues	(.225 .133 .109 .081 .073 .001)
1st eigenvector	(-3.360 -2.287 -0.009 0.038 0.033 0.060)
2nd eigenvector	(0.200 -0.004 0.386 0.745 0.520 0.102)

two variables to have unit variance. The result for SIR-I and II with double slicing is in Figure 5.6b and Table 5.3.

Table 5.3: Eigenvalues and eigenvectors of SIRI (a) and SIRII (b) for the Twist problem with Double Slicing.

(a) SIRI

eigenvalues	(.973 .129 .041 .018 .008 .001 .000 .000 .000 .000)
1st eigenvector	(-0.861 0.200 0.034 -0.003 0.049 0.007 0.037 0.014 -0.038 0.023)
2nd eigenvector	(0.158 0.089 0.630 0.115 0.379 0.113 0.480 0.055 -0.365 0.283)

(a) SIRII

eigenvalues	(.694 .130 .118 .095 .089 .079 .071 .049 .041 .004)
1st eigenvector	(-0.926 -1.219 -0.108 0.107 0.010 -0.013 -0.161 -0.001 -0.055 0.024)
2nd eigenvector	(-0.051 -0.211 0.386 0.310 -0.300 0.209 0.630 0.170 -0.089 0.416)

Example 5.4 Sonar data.

This data set can be found in Gorman and Sejnowski (1988). Sonar signals bounced off a metal cylinder (class 1) or off a roughly cylinder rock (class 2) are recorded in 60 channels. The training set consists of 111 cases from class 1 and 97 cases from class 2. Each case is viewed as a curve $x(t)$, $t = 1, \cdots, 60$. The direct application of LDA or any generalization to 60 predictors with a sample of only 208 training cases is questionable because of the instability in estimating the covariance matrices (see Appendix C). An alternative is to approximate each curve with a small number of basis functions. Four basis functions, $\phi_1(t), \cdots, \phi_4(t)$, are selected according to a scheme which we describe in detail in Appendix. We reduce the predictor dimension to 6. The first 5 predictors are the coefficients $\alpha, \beta_1, \cdots, \beta_4$, obtained

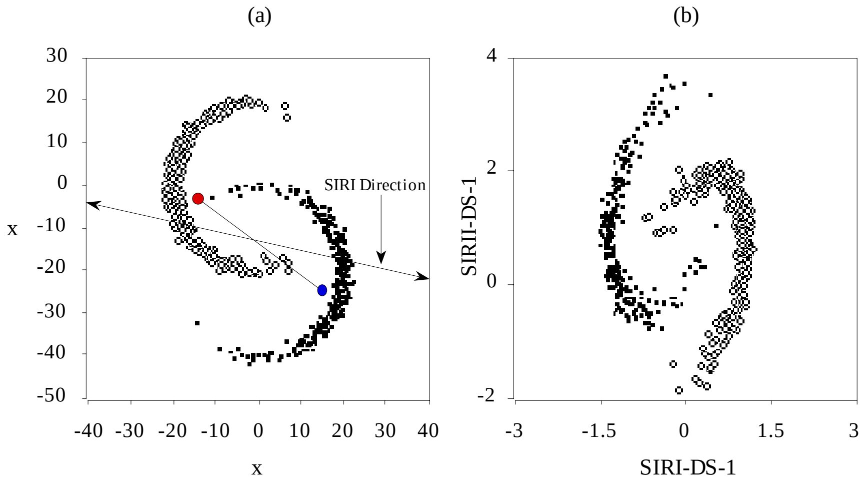


Figure 5.6: Twist Problem: (a). Best View with SIRI direction; (b). SIR's View Double Slicing.

by least squares fitting:

$$x(t) = \alpha + \beta_1\phi_1(t) + \cdots + \beta_4\phi_4(t) + \epsilon_t, t = 1, \cdots, 60.$$

The sixth predictor is $\log(r^2/(1-r^2))$, where r^2 is the R-squared value from the least squares fit.

SIR is then applied to this set of six predictor variables. The first direction found by SIR and the class label Y are used as the directions for running double-sliced SIR (SIR_{ds}). Two directions are found by SIRDS, which we denote as SIRDS-11 and SIRDS-12 (Figure 5.7). The eigenvalues and first two eigenvectors for SIRDS are displayed in Table 5.4. We observe that the correlation coefficients between r^* and SIRDS-11 with SIRDS-12 are -0.07 and 0.6826 respectively. Thus although SIR (or equivalently LDA) does not use r^*, the information contained in r^* is used in SIRDS.

After reducing to the two SIRDS variates, a k-nearest-neighbor classifier is applied. For k=1,3, . . . ,15, the resubstitution error rates are (24.52%, 22.60%, 23.56%, 22.12%, 19.23%, 20.19%, 23.56%, 24.04%) respectively with a minimum resubstitution error-rate of 19.23% at k=9.

14.6 Conclusion.

LDA is a popular method for classification. We re-investigate its theoretical property from the dimension reduction point of view. The canonical variates are found to be the same as the SIR variates except for the scaling. We examine in detail the assumption underlying SIR and apply them to the classification problems. This helps justify the use of CRIMCOORDS for informative graphical display of separation patterns between different classes. However the theory of SIR does not justify the use of linear rules. We illustrate that nonparametric density estimation following dimension reduction can be more informative then LDA. As in known, SIR may not be able to find “all” e.d.r. directions. We investigate two types of generalizations for finding more e.d.r directions. One of them is the second moment based method. This method explores differences between group covariance. Compared to SIR, one drawback of this method is the uncertainty introduced by covariance estimation. Another method, double-slicing, is less sensitive to covariance estimation. These methods extend the power of LDA and can be used to reveal more complicated data pattern.

Appendices: More on Sonar data.

A. Description of the procedure in choosing the basis functions.

The first two basis functions $\phi_1(t)$, $\phi_2(t)$ are taken as the average of all curves $x(t)$ from the first class and the second class, respectively (Figure A.1). Each curve is tentatively fitted with these two basis functions. Then an LDA is conducted on the three predictors, α, β_1, β_2. Figure A.2 of first CRIMCOORD shows a good portions of cases in the middle part cannot be distinguished well. This portion is extracted out, which has 141 cases. Our third and fourth basis functions $\phi_3(t)$ and $\phi_4(t)$ are just the average of all curves in this portion that came from class 1 and class 2 respectively (Figure A.3).

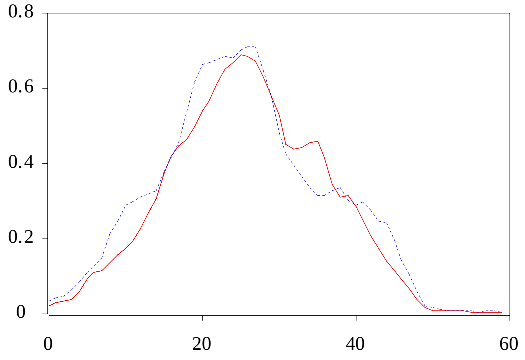


Figure A.1: First two basis functions, $\phi_1(f_i)$, and $\phi_2(f_i)$, $i = 1, L, 60$. The solid(red) and dashed(blue) curves represent the mean signals of the two groups of metal($y = 1$) and rock($y = 2$).

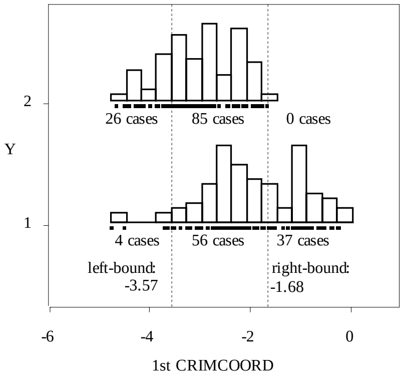


Figure A.2: Fisher's linear discriminant analysis for the three new predictors with 141 ambiguous signals.

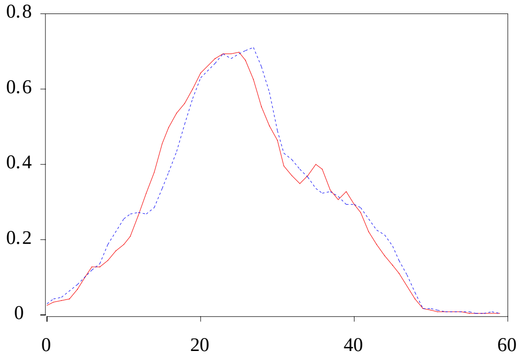


Figure A.3: Additional Basis functions, $\phi_3(f_i)$, and $\phi_4(f_i)$, $i = 1, L, 60$. The solid(red) and dashed(blue) curves represent the mean signals of the two groups of metal($y = 1$) and rock($y = 2$) for the 141 ambiguous cases.

B. Stability.

To see how stable the proposed procedure is, we proceed with the following simulations.

Each time we split all 208 cases into a training set and a test set with probabilities equal to 0.75 and 0.25 respectively. From a training set we first identify the 4 basis functions using the procedure described in Appendix A. Then we find two SIRDS directions. Signals in the test set are then projected to the obtained SIRDS directions. 1000 simulations with *k*=1,3, . . . ,15 are carried out, the result, Figure B.1.c. The lowest average error rate of 22.68% for test set is reached at *k*=15 with an average standard deviation of 0.056. For comparison, the same simulation data is also used to carry out the LDA analysis and k-NN analysis for the original 60 variables. The average error for test set for LDA is 26.63% (Figure B.1.a), much higher than the k-NN results for the 6 base functions. The k-NN analysis for the original 60 variables are also consistently worse than that of the 6 base functions one, Figure B.1.b.

LDA with 60 predictors.

The resubstitution error rate of Fisher's linear discriminant analysis with the original sonar data is 9.615%, which corresponds to 20 misclassified signals (12 metal signals and 8 rock signals each), Figure C.1. We suspect that this 9.615% resubstitution error rate is too low to be true. Since there are only 208 subjects in total with 60 variables, the estimation of the covariance matrix will be unstable with this configuration which may result in a problem of overfitting in the training procedure. We carry out the following leave-one-out procedure to justify this suspicion. Each time we use one of the 208 signals as a test signal and use the other 207 signal to find the Fisher's linear discriminant function for predicting the class label of that selected test signal. Among all 208 runs, 51 signals are misclassified which corresponds to a leave-one-out error rate of 24.52%. This leave-one-out error rate of 24.52% is much higher than the resubstitution error rate of 9.615% by the single run linear discriminant analysis for the full 208 cases as we expected.

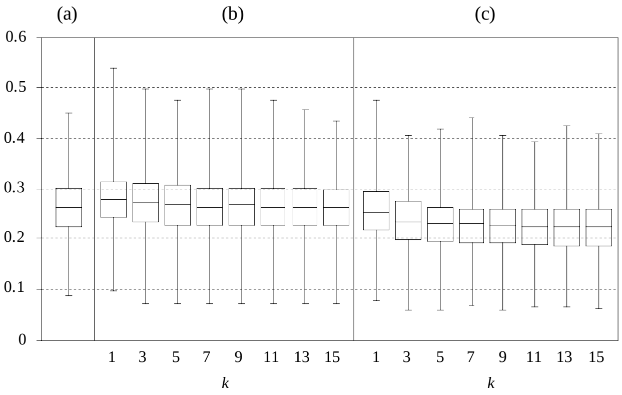


Figure B.1: Box-Plots for Simulation Result with 1000 runs:
(a). LDA with 60 original variables;
(b). k-nearest-neighbor classifier with 60 original variables;
(c). k-nearest-neighbor classifier with 6 basis functions.

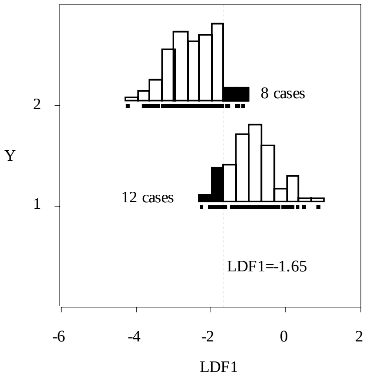


Figure B.2: Fisher's linear discriminant analysis for the full sonar data set with 20 misclassified signals. The upper and lower histograms represent the distributions of observations projected onto the Fisher's linear discriminant function(LDF) of the original 60 variables from the metal(y=1) and rock(y=2) groups respectively. The dashed line is the cutting point for prediction from LDF. The black bars represent the cases that are misclassified by LDF.

References

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), *Classification and Regression Trees*. CA: Wadsworth.

Brillinger, D. R. (1991), Comments on "Sliced Inverse Regression for Dimension Reduction", by K. C. Li, *J. Amer. Stat. Assoc.*, **86**, 333-333.

Chen, C. H., and Li, K. C. (1998), "A three-way subclassification approach to multiple-class discriminant analysis", *Academia Sinica, Ser.*

Cook, R. D. (1994), "On the Interpretation of Regression Plots," *J. Amer. Stat. Assoc.*, **89**, 177-189.

Cook, R. D., and Nachtsheim, J. C. (1994), "Reweighting to Achieve Elliptically Contoured Covariates in Regression," *J. Amer. Stat. Assoc.*, **89**, 592-599.

Cook, R. D., and Weisberg, S. (1991), Comments on "Sliced Inverse Regression for Dimension Reduction", by K. C. Li. *J. Amer. Stat. Assoc.*, **86**, 328-333.

Ebrey, P. (1993), *Chinese Civilization : A Sourcebook*, (2nd ed.), New York: Free Press, 77-79.

Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Ann. Eugen.*, **7**, 179-188.

Fisher, R. A. (1938), "The Statistical Utilization of Multiple Measurements," *Ann. Eugen.*, **8**, 376-386.

Friedman, J. H. (1977), "A Recursive Partitioning Decision Rule for Nonparametric Classification," *IEEE Transactions on Computers*, **26**, 404-408.

Fukunaga, Keinosuke (1990), *Introduction to Statistical Pattern Recognition*.

Gorman, R. P. and Sejnowski, T. J. (1988), "Analysis of hidden units in a layered network trained to classify sonar targets." *Neural Networks*, **1**, 75-89.

Gnanadesikan, R. (1977), *Methods for statistical data analysis of multivariate observations*, New York: John Wiley & Sons.

Hall, P., and Li, K. C. (1993), "On Almost Linearity of Low Dimensional Projection from High Dimensional Data," *Ann. Stat.*, **21**, 867-889.

Hsing, T and Carroll, R. J. (1992), "An Asymptotic Theory for Sliced Inverse Regression," *Ann. Stat.*, **20**, 1040-1061.

Koontz, W. and Fukunaga, K. (1972), "A Nonparametric Valley-Seeking Technique for Cluster Analysis," *IEEE Transactions on Computers*, **21**, 171-178.

Koontz, W., Narendra, P. and Fukunaga, K. (1975), "A Graphic-Theoretic Approach to Nonparametric Cluster Analysis," *IEEE Transactions on Computers*, **25**, 936-944.

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction," (with discussion), *J. Amer. Stat. Assoc.*, **86**, 316-342.

Li, K. C. (1992a), "Uncertainty Analysis for Mathematical Models with SIR", in *Probability and Statistics*, eds. Z. P. Jiang, S. H. Yan, P. Cheng, and R. Wu, Singapore: World Scientific, pp. 138-162.

Li, K. C. (1992b), "On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma," *J. Amer. Stat. Assoc.*, **87**, 1025-1039.

Loh, W. Y. and Vanichsetakul, N. (1988), "Tree-Structured Classification via Generalized Discriminant Analysis," (with discussion), *J. Amer. Stat. Assoc.*, **83**, 715-728.

Rao, C. R. (1952), *Advanced Statistical Methods in Biometric Research*, New York: John Wiley & Sons.

Schott, J. R. (1994), "Determining the Dimensionality in Sliced Inverse Regression," *J. Amer. Stat. Assoc.*, **89**, 141-148.

Van Ness, J. W. and Simpson, C. (1976), "On the Effects of Dimension in Discriminant Analysis," *Technometrics,* **18**, 175-187.

Zhu, L. X., and Fang, K. T. (1996), "Asymptotics for Kernel Estimate of Sliced Inverse Regression," *Ann. Stat.* **24**, 1053-1068.

Zhu, L. X., and Ng. (1995), "Asymptotics of Sliced Inverse Regression," *Statistica Sinica*, **5**, 727-736.

Chapter 15

Three-way subclassification with application in handwritten digit recognition

This chapter is a minor modification of Li and Chen(1998). In the previous chapter, we have encountered the problem of failing to find more e.d.r. directions when there are only two classes. We shall now discuss the opposite situation: the case that number of classes k is large. In this case, the significant SIR directions are likely to be too many to handle simulteneous.

To handle this situation, we propose a strategy based on three-way subclassification. The original k-class problem is first decomposed into $k(k-1)/2$ smaller problems, each involving a three-way classification between a pair of classes A and B, and a combined third class which is just the union of all other classes. Then the results from each subclassification will be synthesized under a conditional error rate analysis. Our formulation of three-way subclassification is ideal for exploring the degree of unanimity among different decisions made in subclassification. This advantage is illustrated well in the context of hand-written digit recognition, using the data set of LeCun et al. (1989). A large portion of high quality images in the test set can be easily filtered out by three-way subclassification - they are predicted with an error rate less than 1%.

15.1 Introduction.

Discriminant analysis problems are also called classification problems. They are concerned with the class membership prediction of an item from a population which is partitioned into k classes. One well-known example is the iris data studied by R. A. Fisher. Three kinds of iris are under consideration and the problem is how to distinguish them based on data from four physical measurements. In search of a solution for this problem, Fisher introduced the celebrated linear discriminant analysis (LDA), which we have given a more general justification through SIR theory in the previous chapter. Another classification problem which has recently received a great deal of attention is the recognition of handwritten digits such as zip codes from envelopes. The number of classes in this problem is $k = 10$ and each class represents one of the ten digits, $0, 1, \cdots, 9$. Other classification problems may come from different scientific areas such as speech recognition, medical diagnosis, remote sensing, and so on. It is also called supervised machine learning in pattern recognition.

As expected, the task of classification becomes harder when the number of classes k increases. This is in part due to the difficulty of finding main feature variables suitable for all k classes. In generic terms, features are just any real-valued functions of the original variables manufactured from the training data. They are often needed for reducing the data dimensionality in complex applications. At this juncture, one may wonder why not apply SIR or its generalization directly to the original variables. The reason is that the feature attraction procedure typically involves special nonlinear transformation on the original input variables. Such transformation, although subjective, does reflect expertise knowledge on the subject matter. It is difficult to approximate such transformation by linear combinations as we do in our general-purpose dimension reduction scheme.

Suppose that we can restrict attention to a given set of important feature variables. There is another difficulty for large k. The significant SIR directions may already be too many. It would be difficult to visualize joint distribution of the SIR variates directly. Under such circumstances, further steps are needed to simply the result.

Since binary classification is easier, an attractive strategy to solve the $k-$group problem is to consider each of the $k(k-1)/2$ two-class subclassification problems separately first (Friedman 1996). In this approach, only the items from the two classes in the learning sample are used to construct a classification rule for each two-class problem. After that, a final decision is made by combining results from each problem together. A simple way to do so is to use the majority rule which assigns the class membership of a new item in a way similar to a tournament with k players, each player representing one class. A two-class classification can be thought of as a match between two players and the winner is the one representing the class into which the new item is classified under the rule obtained from the learning sample. The tournament champion, the player with the best winning record (if without ties), is then assigned to be the item's final class membership. Another interesting way of using two-class subclassification is proposed in Hastie and Tibshirani (1996).

In this chapter, a different approach is taken that also breaks the original k-class problems into $k(k-1)/2$ smaller problems. But instead of binary classification, each smaller problem is formulated in terms of a three-way subclassification. The three classes are a pair of classes A and B from the original k classes, together with a combined third class which is just the union of all other classes. The combined class will be referred to as "Others" later on.

An advantage of 3-way subclassification over the binary subclassification is immediate. We can take full advantage of the two-dimensional CRIMMCOORD plot.

Our formulation of three-way classification aims at the exploration about the degree of agreement among the different decision rules constructed from the simplified subclassification problems. In assigning the final membership of an item, this strategy allows us to observe if the classification is reached by an unanimous decision or not. For a class to be an unanimous winner, not only it must win in each subclassification problem where the class actively participates as class A or class B, but also in all other problems where the class is passively absorbed into the combined ”Others” , an "Others" decision must be obtained.

There is an important issue which is not much discussed in most discriminant analysis literature. Quite often due to economic reasons, the tolerable error rate in industrial applications can be very stringent. This rate might not be easy to achieve when applied to all items to be tested. However, it can be expected that a good portion of them may be easier to classify than others. Thus it is useful to find a simple tool that can help locate items in such a higher-grade subpopulation. Intuitively, the group of items which are classified under an unanimous decision should have a lower rate of prediction error. Thus they are the natural candidates to be in the higher-grade subpopulation.

We shall use hand-written digit recognition as an example to illustrate the advantages of our approach. The data base consists of digitized images of zip codes on envelopes passing through Buffalo, NY. (LeCun et al., 1989). Using a conditional error rate analysis, we can identify a large portion of sample images in the test set that can be classified under simple linear classification rules at a misclassification rate less than 1%.

In section 2, we first give a brief account of the zip code data. A centre-of-mass based method for feature extraction is introduced, to be followed by a preliminary analysis using linear discriminant rules. Section 3 describes details of the three-way subclassification settings and discuss how the results can be combined using conditional error analysis. Further discussions are given in Section 4. Here we compare the three-way subclassification with binary subclassification. We introduce another way of partitioning for extracting feature variables from the zip code data. With this new feature space, we show how to combine the two-way and the three-way methods together to get better results. Section 5 gives some concluding remarks.

15.2 Data, features, and LDA.

In this section, we first describe the handwritten data set which will be used throughout this article. Then we introduce a method for extracting features that reduce the dimensionality to a level easier to manage. A preliminary analysis involving linear discriminant analysis (LDA) is also reported.

15.2.1 The zip code data.

Our data base comes from handwritten zip codes that appeared on some envelopes of U.S. mail passing through the Buffalo, NY post office. The digits were written by many different people with a great variety of writing styles and instruments. Each digit is converted into a 16 by 16 pixel image after some preprocessing as described in LeCun et al. (1989). Figure 2.1 shows some of these normalized images.

The seminal work of LeCun et al. uses a neural network with three hidden layers- 768, 192, and 30 hidden units respectively for each layer. A misclassification rate of 0.14% on the training data and 5.0% for the test set were reported. This remarkably low rate of error cannot be achieved without clever ideas and deliberated efforts on setting up clever connection architecture and contrains on weight constraints. For example, the same authors also reported that a fully connected network with one hidden layer of 40 nodes yields 8.1% error rate on the test data. It is also worth mentioning that since backpropagation is used, the

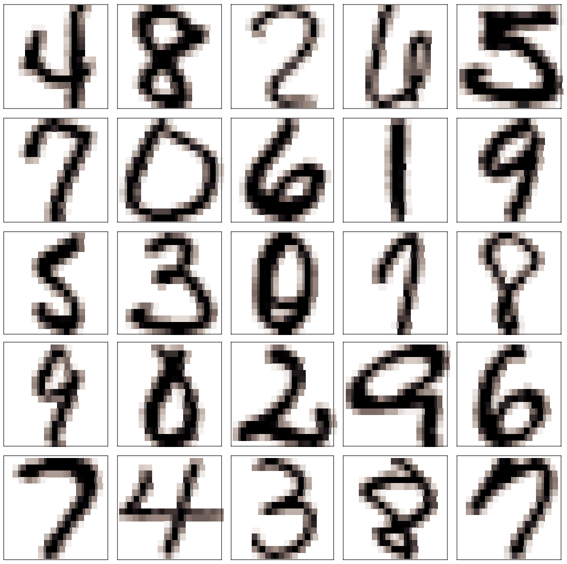


Figure 2.1: Sample Hand-Written Digits

connection coefficients are updated sequentially as each image pattern in the training set is presented. A total of m passes through the training set means 7291m pattern presentations, and the coefficients are updated accordingly. The 5% rate for test set is obtained when the number of passes is 23 (which means 23 times 7291 = 167,693 updates). Figure 2 of LeCun et al. shows how this error rate depends on the number of passes. Although there is no clear stopping rule, it is noticed that for a wide range, between 5 and 30, the error rates are falling between 5% and 6% - reaching about 5.5% at $m = 30$.

This data set has since received a great deal of academic attention. Their error rates have become the bench-marks for comparison. There are several attempts aiming directly at improving the error rates for optical digit recognition. The best improvement reported in the literature seems to be 2.5% by Simard, LeCun and Denker (1993) who use a transformation based nearest neighbor method. The transformations intend to incorporate various kinds of distortion due to factors such as shifting, scaling, rotating, and so on. Unfortunately, to implement such a discriminant rule, it requires a heavy amount of computation.

This data set is also often used as an illustration for new all-purpose classification methods which may not be specially tailored for digit recognition. In such cases, the rates are usually poorer than the 5% to 6% range. For example, using a penalized discriminant rule, Hastie, Buja, and Tibshirani (1995), a rate of 8.2% is obtained, which is an improvement over the rate of about 10% by the usual LDA.

15.2.2 Centre-of-mass-based partition.

Due to the spatial arrangement, the 256 variables representing the 16 by 16 image for each sample character are highly correlated. Such redundancy among the input variables allows a certain degree of flexibility in designing strategies for feature extraction.

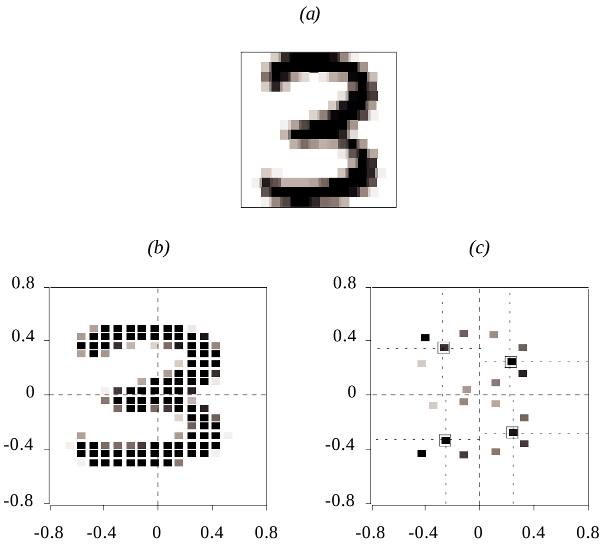


Figure 2.2: Locations of mass Centers for Digit 3: (a). Orogonal Image; (b). Transformed Digit; (c). Locations of mass Centers with Weights.

The method we use treats each character image as a piece of object with mass at each pixel (i, j) equals to its grey level w_{ij}. For each object, we first compute the centre of mass (C_{01}, C_{02})= $(\sum w^*_{ij} i, \sum w^*_{ij} j)$ where $w^*_{ij} = w_{ij}/\sum w_{ij}$ is the total mass of the object. This mass center is then used to normalize the image by shifting (C_{01}, C_{02}) to the origin $(0, 0)$. The range of the two coordinate axes are scaled to take values within -1 and 1 , using grey levels greater than a prespecified threshold value as a common denominator. Figure 2.2(b) gives a transformed digit 3 whose original image is in Figure 2.2(a). The entire image is now partitioned into four pieces, one piece from each quadrant.

The next step is to calculate the mass center for the image in each quadrant. This generates 4 pairs of locational variables. After that, each quadrant is again partitioned into four quadrants, using the mass center calculated earlier as the origin. This yields a total of 16 rectangle regions and the mass center of the image in each rectangle is computed. The 20 locations of the mass centers for the digit 3 in Figure 2.2(b) are shown in Figure 2.2(c). Our feature space consists of these 40 locational variables and the 16 weight variables for the total mass in each of the final 16 rectangles. Thus the original 256 grey-level variables is reduced to 56 feature variables.

Quadrants are certainly not the only choices we can use in our centre of mass based partition. Later on a system of 8 radial lines as described in Section 4.2. will be used, which leads to better results.

In LeCun et al. (1989), there are 7291 cases in the training set and 2007 cases in the test set. We use only 7188 and 1991 respectively as our training and test sets - this will be referred to as 7188/1991 data. Other cases are deleted for the moment because they have no mass in some of the final 16 rectangles. This problem can be corrected if we use the radial partition system - details to be discussed later.

It is not clear how the original 2007 test cases were selected in LeCun et al.. Typically, if the test set comes from the same population as the the training set. then it should exhibit characteristics similar to any randomly selected subsample of equal size from the training set. However, this is not the case here. Several authors have reported an unexpected increase in error when applied to LeCun et al.'s test set. For comparison, these authors often generate two randomly selected subsets of size 2000 each from the 7291 training set, one as the new training set and the other as the new test set. They find that the error rates are smaller for the new test set than the original size-2007 test set. We shall follow this practice and generate a 2000/2000 data - 2000 cases for the training set and 2000 cases for the test, both being randomly selected from the 7188 cases.

15.2.3 A preliminary analysis.

We first apply linear discriminant analysis (LDA) to the 2000/2000 data. The error rate is 5.75% for the training set and 8.3% for the test set. We then apply LDA to the 1991/7188 data. As expected, the result is even worse, 7.2% for the training set and 10.7% for the test set.

We take a closer look at the training set of 2000/2000 data. The first three canonical variates from LDA are shown in Figure 2.3. Noticeable clustering patterns can be found. For example, a good portion of digit 0 visually separable from other digits is found in the lower left corner of Figure 2.3(a). In the same figure, digit 1 appears to occupy in another corner.

We are led to the suggestion of isolating these two clusters from the rest of data because they appear easier to classify than others. To do so, we can formulate a three-way classification problem by considering ”0” as one class, ”1” as another class, and pooling all other digits together as a third class called ”Others”. We first reduce the dimension of the feature space from 56 to 2 using the canonical variates from LDA. Figure 2.4 shows the scatterplot of the two canonical variates. The clustering pattern is more clear-cut than what is seen in Figure 2.3. Since we are interested in isolating cases that can be predicted with higher precision, the boundaries of discriminant regions for each class are pushed a bit toward the centers of 0 and 1 - this is done by setting the prior probability of 0 and 1 to be .0005 each. After we isolate these two clusters with mostly 0 and 1, we proceed by considering another three-way classification for the class ”Others”. This time we choose digits 6 and 7 as two classes and pooling all other digits (including 0 and 1) together as ”Others”. We continue to partition the left-over ”Others” till we obtain a three-way classification tree as shown in Figure 2.5.

From the tree we obtained and the scatterplots we generated along the analysis, we see that a good portion of data forms distinctive clusters which can be isolated in an iterative fashion. Out of the 2000 test cases, about a half of them can be classified with 1% of error. What should we do with the left-over cases? One simple suggestion is to apply the linear discriminant analysis. Another possibility is to repeat the same procedure again before linear discriminant analysis is applied. The result is given in Figure 2.6.

One problem with this three-way tree approach is how to decide the ordering of partition.

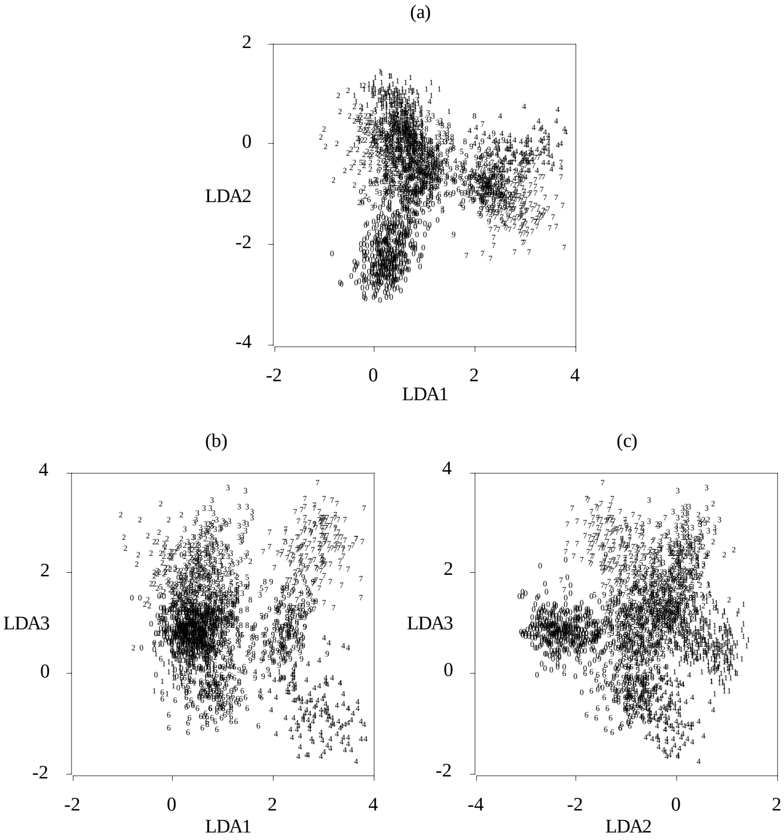


Figure 2.3: LDA Cannonical Variates: (a). LDA1 versus LDA2; (b). LDA1 versus LDA3; (c). LDA2 versus LDA3.

Why 0 and 1 are chosen first? There are certainly many criteria that can be used. We simply choose the pair which gives the smallest error rate for the training data. Another problem is the choice of prior in determining how much the lines should be pushed away from the center of the ”Others”. This is another optimization problem, which needs to be resolved.

While the above line of thoughts may be worth pursuing further, in this article we shall alter our strategy a bit in order to bypass such difficult optimization decisions. There is no need to generate three-way trees. Instead, we shall synthesize results from all three-way classifications in a way to be discussed in the next section.

15.3 Three-way subclassification.

We begin with a general description of our three-way subclassification method. Suppose there are in total k groups under consideration. For each pair of groups, i and j, we formulate a three-group problem - group i , group j , and a third group which is the union of all other groups. Suppose from the training data, a classifier denoted by T_{ij}, is obtained. For any

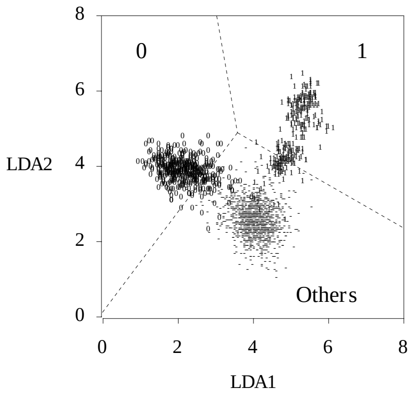
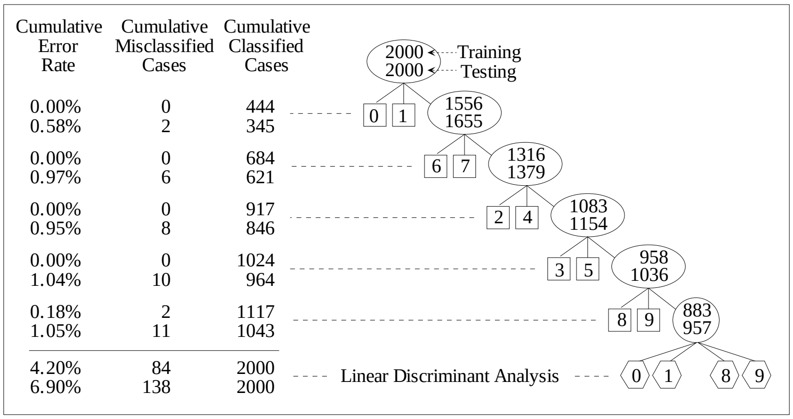


Figure 2.4: LDA Cannonical Variates for Group 0, 1 and Others(-).

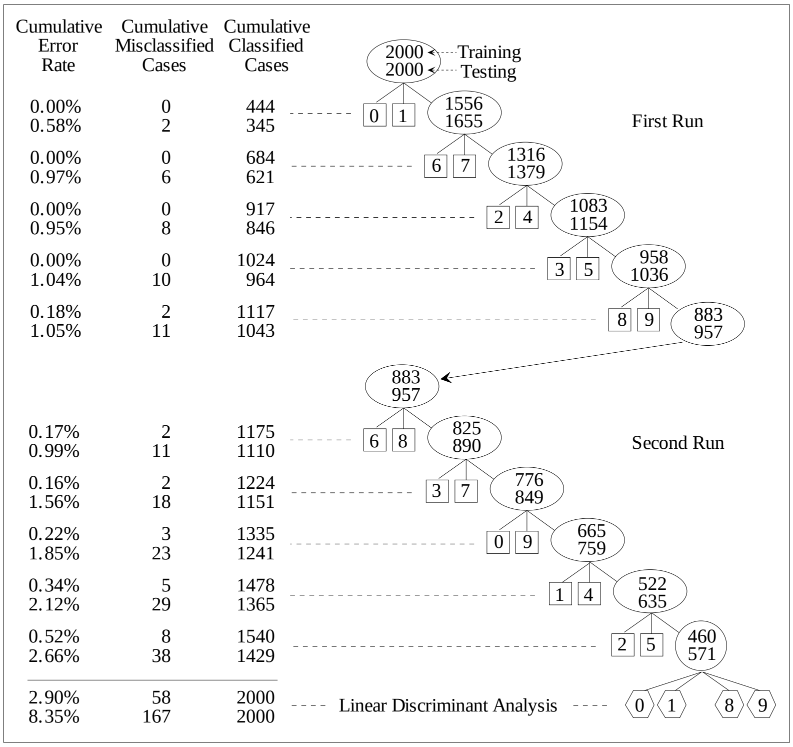


Cumulative Error Rate	Cumulative Misclassified Cases	Cumulative Classified Cases
0.00% 0.58%	0 2	444 345
0.00% 0.97%	0 6	684 621
0.00% 0.95%	0 8	917 846
0.00% 1.04%	0 10	1024 964
0.18% 1.05%	2 11	1117 1043
4.20% 6.90%	84 138	2000 2000

Figure 2.5: Three-Way Classification Tree for the Hand-Written Digit Recognition Problem.

unit with feature **x**, the classifier assigns it a membership $T_{ij}(\mathbf{x}) = i, j$ or $Others$ with $Others$ standing for the third group. Subclassification is carried out for all of the $k(k-1)/2$ three-group problems. At the end, for each **x**, we can have a total of $k(k-1)/2$ values, $T_{ij}(x), 1 \le i < j \le k$.

For the digit data, there are $k = 10$ classes, yielding a total of 45 subclassification problems. For each problem, we consider a linear partition rule with a simple tree structure (Figure 3.1). The tree consists of two levels of partitions - the first one is to divide the training set into two nodes $i/Others$ and $j/Others$ along the projection that best separates class i and class j. Node $i/Others$ is expected to contain most members from digit i and a good portion of members from digits other than i and j. This node is further divided into two children nodes, i and $Others$ again by a best linear partition rule. The other node $j/Others$ is similarly partitioned into two nodes, j and $Others$. There are certainly quite a few alternative partition rules worth trying. But since our main focus in this article is on the strategy



Cumulative Error Rate	Cumulative Misclassified Cases	Cumulative Classified Cases
0.00% 0.58%	0 2	444 345
0.00% 0.97%	0 6	684 621
0.00% 0.95%	0 8	917 846
0.00% 1.04%	0 10	1024 964
0.18% 1.05%	2 11	1117 1043
0.17% 0.99%	2 11	1175 1110
0.16% 1.56%	2 18	1224 1151
0.22% 1.85%	3 23	1335 1241
0.34% 2.12%	5 29	1478 1365
0.52% 2.66%	8 38	1540 1429
2.90% 8.35%	58 167	2000 2000

Figure 2.6: Repeated Three-Way Classification Tree for the Hand-Written Digit Recognition Problem.

of three-way partition itself, to avoid further distraction, we shall not make such attempts here.

15.3.1 Majority rule.

To combine the results from each subclassification, we first count the frequency that each class is assigned. For any fixed unit $\mathbf{x}$, we obtain a k-dimensional vector $(c_1(\mathbf{x}), \cdots, c_k(\mathbf{x}))'$ where for each l between 1 and k, $c_l(\mathbf{x})$ equals the total number of times that $T_{ij}(\mathbf{x})$, $1 \leq i < j \leq k$, equals l. We can think of each classifier T_{ij} as assigning the winner for a match between players i and j, but with possibility for a tie in which $T_{ij}(\mathbf{x})$ equals Others. The vector $(c_1(\mathbf{x}), \cdots, c_k(\mathbf{x}))'$ is just the score board showing the winning record for each player. Then it should be clear that the largest value that $c_l(\mathbf{x})$ can take is $k-1$ because that is the total number of matches that class l has participated. We can rearrange the score board in a non-increasing order, $M_1(\mathbf{x}) \geq M_2(\mathbf{x}) \geq \cdots \geq M_k(\mathbf{x})$.

Consider the majority rule for the final class membership assignment- the class winning the most is assigned. We may classify unit $\mathbf{x}$ into the class l achieving $c_l(x) = M_1(\mathbf{x})$. To keep the procedure simple, if necessary, randomization can be used as a tie-breaker.

We apply the majority rule to the 2000/2000 data and the 7188/1991 data. The results are

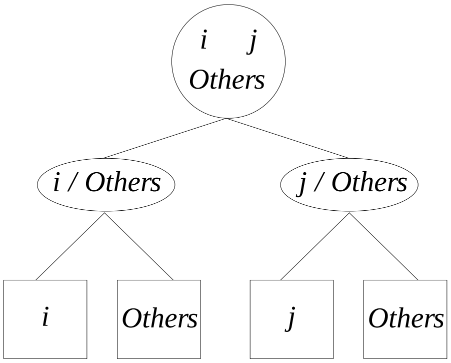


Figure 3.1: Tree-Structure for Three-Way Subclassification.

summarized by two misclassification matrices in each case - one for the training set and the other for the test set; Tables 3.1(a)-(b) and 3.2(a)-(b). A cell in each matrix shows the number of times a digit in the beginning of a row is misclassified as another digit on the top of the corresponding column. For example, take a look at the row with the digit 2 in the training set of 7188/1991 data. We see that 2 is misclassified as 0 two times, as 1 three times, and so on. It is also unclassified 16 times as the last column ”Other” indicates. For the 2000/2000 data, the overall error rate is 2.6% for the training and 4.85% for the test set; the unclassified rate is 1.45% for the training and 1.65% for the test set. For the 7188/1991 data, we have error rates 3.7% (training set) and 6.9% (test set) and unclassified rates 2.0% (training set) and 2.5% (test set). In general, these numbers are not impressive. However, there are better ways of using three-way subclassification than the straightforward application of the majority rule. This is to be explored next.

15.3.2 Conditional error analysis and unanimity.

The performance of a classification rule is usually evaluated in terms of two misclassification rates, one for the training data and one for the test data. They are simply the proportion of cases being incorrectly classified in the respective data set. We shall examine these rates more closely.

Our conditional error analysis exploits the largest two values on the score board for each unit $\mathbf{x}$, $M_1(\mathbf{x})$ and $M_2(\mathbf{x})$.

The most interesting condition is when $M_1(\mathbf{x}) = k - 1$ and $M_2(\mathbf{x}) = 0$. Suppose there is an ideal unit $\mathbf{x}$ from say class l, which is very easy to classify. Then for this unit the maximum score $M_1(\mathbf{x})$ is very likely to be $k - 1$ and it is expected to be achieved by class l, $c_l(\mathbf{x}) = M_1(\mathbf{x}) = k - 1$. This is because when class l competes with any other group ($k - 1$ times in total) , the classifier should return l. On other hand, when the competition is between any two classes i, j other than class l, the classifier should return $Others$ because class l is contained in “Others”. Thus $(M_1, M_2) = (k-1, 0)$ represents the situation where an unanimous decision is reached. We anticipate that the final classification to be most accurate

Table 3.1: Misclassification Matrices for Three-way Subclassification Using Majority Rule with the 55-Features-2000/2000 data: (a). Training Set; (b). Test Set.

(a)

$y\backslash\hat{y}$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Other*
0	0	0	0	1	0	0	2	0	0	1	6
1	0	0	2	0	0	0	0	0	0	1	6
2	0	1	0	1	1	1	3	0	2	1	2
3	0	0	3	0	0	0	0	0	0	1	4
4	0	1	0	0	0	0	1	0	0	11	0
5	2	0	0	1	0	0	0	0	1	0	0
6	3	1	0	0	0	1	0	0	0	0	2
7	0	0	1	0	0	0	0	0	0	3	0
8	2	0	0	0	0	0	0	0	0	1	4
9	0	0	0	0	0	1	0	1	1	0	5

(b)

$y\backslash\hat{y}$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Other*
0	0	0	0	1	0	0	0	0	0	0	5
1	0	0	0	0	0	0	0	0	0	1	7
2	1	0	0	2	1	0	0	2	1	2	2
3	4	0	4	0	0	0	0	0	0	1	4
4	0	1	1	1	0	0	8	1	0	8	0
5	6	0	3	6	0	0	2	0	1	2	3
6	1	0	1	0	5	3	0	0	0	0	2
7	0	0	0	0	0	0	0	0	0	6	1
8	1	0	0	0	1	1	1	0	0	1	8
9	1	2	0	0	1	0	1	11	0	0	1

under this condition.

For the 2000/2000 data, we find that there are 1604 cases in the training set with $(M_1, M_2) = (9, 0)$. From this subset, there are only 4 misclassification cases, representing an error rate of 0.25% which is much smaller than the overall error rate 2.6% given earlier. In the test set, there are 1460 cases with $(M_1, M_2) = (9, 0)$, and 10 out of them are misclassified. This amounts to 0.68% of conditional error, which are again much smaller than the overall error 4.85%. Substantial reduction in conditional error rate also occurs for the 7188/1991 data - 31 out of 5592 (= 0.55%) for the training set and 20 out of 1407 (= 1.4%) for the test set, as compared to the overall rates of 3.7% and 6.9% respectively.

For each fixed value of M_1, we can also anticipate the quality of final classification to go down as M_2 increases because this reflects that the leader faces a stronger challenge from the runner-up and the degree of unanimity goes down. Similarly, for a fixed M_2, the classification quality also degrades as M_1 decreases. Such trends can be found from Tables 3.3(a)-(b)

Table 3.2: Misclassification Matrices for Three-way Subclassification Using Majority Rule with the 55-Features-7188/1991 data: (a). Training Set; (b). Test Set.

(a)

$y\backslash\hat{y}$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Other*
0	0	0	0	2	0	1	5	0	0	2	16
1	0	0	3	0	0	0	0	0	0	2	28
2	2	3	0	13	1	3	5	2	6	6	16
3	4	0	5	0	0	6	0	1	2	2	20
4	0	7	2	2	0	0	24	2	3	34	0
5	13	0	4	6	0	0	5	0	2	2	15
6	8	3	0	0	2	7	0	0	1	0	10
7	1	0	2	0	0	0	0	0	1	20	3
8	4	0	0	2	2	6	1	0	0	4	24
9	0	2	0	1	1	1	1	14	1	0	12

(b)

$y\backslash\hat{y}$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Other*
0	0	1	0	0	0	0	1	2	1	1	14
1	0	0	3	1	0	0	2	1	2	2	5
2	3	2	0	2	1	4	2	1	2	1	5
3	3	0	2	0	0	7	0	1	2	4	7
4	0	4	3	0	0	0	4	2	1	18	1
5	4	0	0	5	0	0	3	1	0	1	5
6	3	0	1	0	3	2	0	0	0	1	1
7	0	1	0	1	2	1	0	0	2	5	1
8	1	1	3	4	0	1	0	0	0	2	8
9	0	0	0	0	1	0	1	2	0	0	3

and 3.4(a)-(b). For example, in Table 3.3(a) and (b), combining numbers from 4 cells corresponding to $M_1 = 5, 6, 7, 8$, $M_2 = 0$ the error rates are $3/90$ (training) and $8/87$ (test) which are lower than the corresponding error rates for $M_1 = 5, 6, 7, 8$, $M_2 =$ 1- 5/20(training) and 8/ 36(test). For $M_1 = 9$, $M_2 = 1, \cdots, 5$, the error rates are 1/124(training), 5/ 199(test) which are lower than the corresponding error rates for $M_1 = 8$, $M_2 = 1, \cdots, 5$ - 4/9(training) and 6/ 26 (test).

15.4 Further considerations.

In this section, we discuss possible ways of enhancing the three-way subclassification approach.

Table 3.3: Conditional Error Matrices for Three-way Subclassification with the 55-Features-2000/2000 data (Number of Misclassification Cases / Number of Total Cases): (a). Training Set; (b). Test Set.

(a)

$M_1\backslash M_2$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
0	29/29								
1	8/29	2/2							
2	0/8	4/8	0/1						
3	3/7	0/5	3/4	0/1					
4	2/10	2/3	1/2	0/1	1/1				
5	0/16	0/3	1/3	0/1	2/3	0/1			
6	0/12	0/3	1/3	2/5	0/0	1/1	1/1		
7	0/19	3/10	1/2	0/0	0/1	0/0	2/4	0/1	
8	3/43	2/4	1/2	1/1	0/2	0/0	0/0	0/2	1/1
9	4/1604	0/64	0/20	1/15	0/13	0/12	0/6	0/8	0/3

(b)

$M_1\backslash M_2$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
0	33/33								
1	14/32	6/7							
2	3/19	4/5	1/1						
3	1/14	3/6	4/6	0/1					
4	1/9	1/3	0/2	1/2	1/2				
5	3/19	1/4	1/3	1/1	2/2	0/0			
6	2/21	1/9	1/2	2/3	0/1	2/3	0/0		
7	1/18	4/11	1/2	0/2	4/6	2/3	1/2	1/2	
8	2/29	2/12	1/6	0/1	3/7	0/3	0/2	1/4	0/2
9	10/1460	2/105	1/44	1/21	1/17	0/12	2/8	0/6	1/5

15.4.1 Conditional error analysis for binary classification.

Unlike three-way subclassification, binary subclassification does not have a straightforward apparatus for unanimity assessment. One possibility is to follow a similar conditional analysis as in the three-way approach. Let $M_1(\mathbf{x})$ and $M_2(\mathbf{x})$ be the highest two scores again , but now obtained from the scoreboard by binary-subclassifications. The ideal condition for most accurate prediction requires $M_1(\mathbf{x})$ to be as large as possible and $M_2(\mathbf{x})$ be as small as possible. The larger the gap between them, the less competitive the runner up is, thus reflecting certain degree of unanimity.

For the digital problem, since there are 10 classes, it is easy to argue that if $M_1 = 9$, then M_2 cannot be smaller than 5. The condition $M_1 = 9, M_2 = 5$ represents the most favorable situation for better classification. We shall anticipate the error rate to increase

Table 3.4: Conditional Error Matrices for Three-way Subclassification with the 55-Features-7188/1991 data (Number of Misclassification Cases / Number of Total Cases): (a). Training Set; (b). Test Set.

(a)

$M_1\backslash M_2$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
0	144/144								
1	23/93	9/16							
2	15/72	14/28	2/2						
3	8/52	6/15	3/10	3/9					
4	6/63	5/13	1/3	1/4	1/1				
5	6/53	4/10	4/7	2/3	2/3	1/2			
6	9/63	2/9	2/10	2/3	2/6	3/6	2/3		
7	6/67	7/15	2/7	3/5	7/13	3/3	2/4	4/5	
8	10/114	12/29	2/12	5/12	4/4	2/4	3/8	8/10	3/4
9	31/5592	4/257	1/98	2/51	0/55	2/32	2/35	2/30	2/19

(b)

$M_1\backslash M_2$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
0	50/50								
1	14/42	5/10							
2	8/32	4/9	3/3						
3	1/17	0/1	4/6	2/3					
4	4/15	0/6	2/2	4/4	0/0				
5	1/18	2/2	2/2	2/2	1/2	1/1			
6	0/19	4/6	1/3	2/3	2/4	1/2	0/0		
7	2/20	1/4	1/4	0/1	1/2	1/1	0/0	2/2	
8	5/49	1/5	2/3	0/3	3/4	0/0	1/2	5/7	0/3
9	20/1407	5/100	5/33	2/19	1/11	1/19	2/10	3/8	4/10

as M_2 increases. This is indeed what we can find from Tables 4.1(a)-(b) and 4.2(a)-(b). For example, in the test set for 2000/2000 data, when fixing M_1 at 9, the error rates are seen to increase : - 0/3, 0/96, 15/506(=2.96%), 50/1337(=3.7%), respectively for $M_2 = 5, 6, 7, 8$. However, an undesirable pattern is that the most favorable condition $M_2 = 5$ is only satisfied by three cases, while the least favorable condition $M_2 = 8$ has 1337 cases. Thus the conditional error analysis on binary classification does not lead to a useful way of finding a large portion of cases which can be classified with very high precision. The error rate has already reached $15/(3 + 96 + 506) = 2.5\%$ when conditioning on $M_1 = 9, 5 \leq M_2 \leq 7$ as compared to $10/1460 = 0.68\%$ for $M_1 = 9$, $M_2 = 0$ from three-way subclassification reported earlier.

However, a positive note for binary classification is that it can be used to improve the non-(9,0) group from three-way subclassification. The error rate is reduced from $120/540 =$

Table 4.1: Conditional Error Matrices for Binary Classification with the 55-Features-2000/2000 data (Number of Misclassification Cases / Number of Total Cases): (a). Training Set; (b). Test Set.

(*a*)

$M_1 \backslash M_2$	*5*	*6*	*7*	*8*
5	0/0			
6	0/0	0/0		
7	0/0	0/0	0/0	
8	0/0	0/0	0/3	1/8
9	0/4	0/113	4/571	15/1301

(*b*)

$M_1 \backslash M_2$	*5*	*6*	*7*	*8*
5	0/0			
6	0/0	0/0		
7	0/0	1/1	4/5	
8	0/0	1/1	12/18	21/33
9	0/3	0/96	15/506	50/1337

22.22% (among which $33/540 = 6.11\%$ were unclassified) to $94/540 = 17.41\%$.

15.4.2 Radial partition.

Encouraged by the promising results from three-way analysis, we apply the same strategy to another set of feature variables. This new feature space has 69 variables. Just like the old feature space, they are also constructed using the centre of mass to guide the partition. The difference comes from the geometric configuration of the partitioning lines. We use an 8-region radial partition system.

We begin with the centre of mass of the whole digit. Instead of quadrants, 8 regions are obtained by further partitioning each quadrant diagonally into two equal pieces. The location of the mass center for each of these 8 regions is computed. After that, the whole digit is horizontally divided into two halves - one half is above the x-axis and the other half is below the x-axis. For each half, we then compute the new centre of mass and apply the 8 region radial partition system to get another 8 locations of mass centers. Figure 4.1(a) shows how the same digit 3 is partitioned. A total of 24 mass centers are located in Figure 4.1(b). Our new feature space consists of these 48 locational variables and together with 21 weight variables. Each weight variable represents the total mass from one of the 24 regions obtained before. Note that due to colinearity among each of the three 8-region partitions, we cannot use all 24 weight variables.

Using this new set of feature variables, the error rate from LDA is 5.8 % = 425 / 7291 for

Table 4.2: Conditional Error Matrices for Binary Classification with the 55-Features-7188/1991 data (Number of Misclassification Cases / Number of Total Cases): (a). Training Set; (b). Test Set.

(*a*)

$M_1\backslash M_2$	*5*	*6*	*7*	*8*
5	0/0			
6	0/0	0/0		
7	0/0	0/0	0/0	
8	0/0	0/0	12/20	41/62
9	0/4	1/117	15/1456	119/5529

(*b*)

$M_1\backslash M_2$	*5*	*6*	*7*	*8*
5	0/0			
6	0/0	0/0		
7	0/0	1/1	2/2	
8	0/0	0/0	7/11	19/23
9	0/2	1/22	7/409	107/1521

the training set and 9.9%= 199 / 2007 for the test set. They are not much different from the LDA results by 56 features. Can the three-way method help filter out a group of high quality cases?

The result is shown in Table 4.3. Here we find that for the unanimous winners, $M_1 = 9, M_2 = 0$, the error rate in the test set is reduced to below 1% (13 out of 1535 cases).

15.4.3 A combined use of different feature spaces.

As pointed out before, three-way subclassification is especially effective in isolating high quality cases- cases that are easier to predict their membership. This is achieved by a conditional error analysis which exploits the degree of unanimity among different classification decisions from subclassification. After locating the very high quality cases, we can then focus on the rest of cases and try to find better classification, perhaps even with drastically different classification methods. For example, consider the nearest neighbor classifier used by LeCun et al.. As mentioned before, it has an error rate of 2.5% for the 2007 test cases, but is computationally very demanding. If we can use it only for the non-($M_1 = 9, M_2 = 0$) cases, then the overall error rate would be still be at most around 3%. But in this way, we have allocated the total computation time more effectively without sacrificing much of the overall classification quality.

Perhaps an easier way of improvement is to combine the results from different feature spaces. We try the following path.

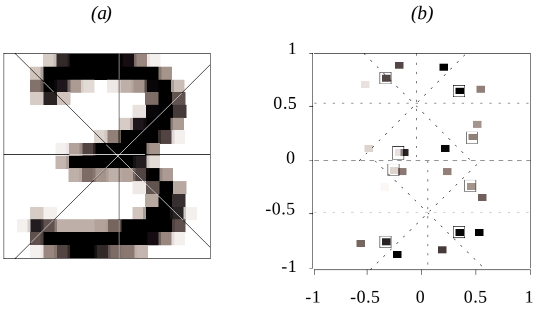


Figure 4.1: Radial Partition for Digit 3: (a). First Level Radial Partition; (b). Locations of Mass Centers with Weights for three radial partitions.

Table 4.3: Misclassification Matrix for Three-way Subclassification Using Majority Rule with the Feature69-7291/2007 data.

$M_1\backslash M_2$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
0	35/35								
1	9/26	7/10							
2	3/13	6/10	2/3						
3	7/17	4/6	6/9	4/4					
4	2/14	1/2	0/2	0/3	1/1				
5	4/12	2/3	3/4	0/1	0/1	1/1			
6	1/11	1/4	2/3	2/2	2/2	1/2	0/0		
7	3/18	3/5	2/4	1/2	2/4	2/4	0/1	1/1	
8	3/28	2/11	4/7	0/3	1/2	3/5	2/2	3/6	1/1
9	13/1535	5/69	1/22	1/15	1/15	0/12	1/10	3/12	4/12

(1). Apply three-way subclassification trained by 69-features to all test cases and locate the $M_1 = 9$, $M_2 = 0$ group.

(2) Apply three-way subclassification trained by 56-features to the non-$(9, 0)$ cases and locate the $M_1 = 9$, $M_2 = 0$ cases.

(3) Apply binary classification trained by 69-features to all other left-over cases.

A breakdown of the error rate is given in Table 4.4. The overall error rate is now about 5.7%, which is compatible with the result (between 5% and 6%) by the neural network approach. It is certainly a great improvement over the original LDA result which is about 10% for either feature. It is also significantly better than the 8.2% error rate obtained by Hastie, Buja, and Tibshirani (1995).

It is interesting to observe that there are a good number of tied scores in binary subclassification. In order to keep the procedure simple, we resolve these ties essentially by a random choice. Further investigation on how to handle these cases seems worthwhile.

Table 4.4: Breakdown of Error Rate for Combining Use of Different Feature Spaces.

Group	Cases	Misclassified Cases	Error Rate
(1). $M_1 = 9$, $M_2 = 0$/3-way/69-features	1535	13	0.85%
(2). $M_1 = 9$, $M_2 = 0$/3-way/55-features	160	15	9.38%
(3). Left-over/binary/69-features	312	87	27.88%
Total	2007	115	5.73%

15.5 Conclusion.

Discriminant analysis is relatively easier when the number of classes is small. In view of this tendency, subclassification appears to be a promising strategy for alleviating the complexity of many classes. In this article, we propose a three-way subclassification method and show that it can be fruitfully applied to complement binary subclassification.

Three-way subclassification is designed to exploit the degree of unanimity among various decisions during subclassification. Thus, unlike the binary situation, the majority rule by itself is not appropriate for the three-way subclassification. Instead, the information gathered from the conditional error rate table in the training set is used for grading the discriminant quality of an incoming unit **x** to be classified. If it falls into the highest grade group $(M_1(\mathbf{x}), M_2(\mathbf{x})) = (k-1, 0)$, then we have the best confidence about the classification accuracy. On the contrary, if it falls into very low grade group with small $M_1(\mathbf{x})$ or large $M_2(x)$, then we had better send it to other more powerful classifiers and hope for a better result.

In many industrial applications, tolerable error rates are usually set up by practical concerns from the economic aspects. Thus it is important to identify sub-populations that are easier to classify than others. Our method has the appeal of being able to identify such higher-grade subpopulations. For the remaining lower-grade subpopulation, we can then rely on more complicated methods to carry out the task. The ability to filter out cases that are harder to classify is a very important consideration in quality management because this helps engineers identify where the quality improvement should be focused on. Further investigation along this line of thoughts is worth pursuing.

References

Hastie and Tibshirani, R. (1996). Classification by pairwise coupling. *Technical Report.*

Friedman, J.(1996). another approach to polychotomous classification. *Technical Report.*

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. *Ann. Stat.* **23**, 73-102.

Li, K.C. (1991) Sliced inverse regression.

LeCun, Y. Boser, B., Denker, J.S., Henderson, D, Howard, R.E., Hubbard, W., and Jackel, L.D.(1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation.* **1,** 541-551.

Simard, P.Y., LeCun, Y. and Denker, J.(1993). Efficient pattern recognition using a new transformation distance, in " *Advances in Neural Information Processing Systems*," Morgan Kaufman, San Mateo, CA, pp 50-58.

Chapter 16

Tree-structured Regression via PHD

This chapter is a minor modification of Li and Lue(1998).

We introduce a new approach to tree-structured regression. The distinctive feature is the use of geometric shape information about the regression surface for guiding the choice of splitting directions. The procedure begins with finding a direction along which the regression surface bends the most. This direction is used for splitting the data into two regions. Then within each region, another direction is found in the same manner and partitioning is carried out recursively. The process continues until the entire regressor domain is decomposed into smaller regions wherein the surface no longer bends significantly and linear regression fit becomes appropriate. For implementing the direction search, PHD is applied. Several simulation and empirical results are reported. Comparison with three methods CART, SUPPORT, and MARS, is made. It highlights the benefit of using geometric information which is hard to retrieve directly by other methods.

16.1 Introduction.

Tree-structured regression is a promising approach to high dimensional data analysis. It estimates the regression function by recursively partitioning the entire data into several homogeneous regions. Since one can split the data in many possible ways, this leaves a great deal of flexibility in model building. AID (Sonquist, 1970) and CART (Breiman, Friedman, Olshen and Stone, 1984) are among the most popular methods in this area.

CART searches over all possible cut-points for a univariate split at each intermediate node by optimizing a cost function. A large initial tree is generated. Then it uses a cross-validation procedure to prune down the size of the tree. A simple average of the response variable is computed in each terminal node, yielding a piece-wise constant prediction equation. Another tree-structured method, SUPPORT, is proposed in Chaudhuri, et. al. (1994). SUPPORT also uses univariate splits. To avoid extensive searching in optimizing the cost function, the cut-point/variable selection is determined via significance testing procedures for some two-sample problems. It then uses a looking-ahead cross-validation procedure to control the tree size. Multiple linear regression is applied in each terminal node for prediction. Alexander and Grimshaw (1996) proposed yet another splitting strategy by a different kind of cost consideration. Simple linear regression is used.

Despite of their differences in the splitting rule selection, tree construction has always been driven by the concern of cost optimization. This is because one of the primary goals in regression is to find rules that have as small prediction errors as possible and cost optimization appears to be a direct way to meet this need. Indeed, such a rationale also extends to other nonparametric regression methods. See the reference in Friedman(1991) and its discussion for a long list of smoothing techniques. In our study, we shall only include Friedman's MARS for comparison. MARS is not a partition-based procedure. It uses splines to fit the data. The selection of spline basis function is guided by a variant of generalized cross validation.

In this article, we shall introduce a new method of constructing regression trees, which is based on a totally different rationale. Instead of being overwhelmed by cost consideration, our primary concern is how to explore the geometric information about the regression surface. Intuitively, such information could be best utilized if we can physically "see" the surface. For example, if the response surface appears like a plane, then splitting is unnecessary. The regression function can be well approximated by the standard multiple linear regression method. On the other hand, if the surface appears to bend in some global direction, then it is advantageous to split the data along that projection. We can anticipate that after recursive splitting, the surface in each node should become flatter and flatter. In the end, this could allow us to fit the data linearly.

The weakest link in this geometrically-motivated splitting strategy is how to detect the bending directions in the regression surface without having to estimate the surface in the first place. For this purpose, we use the dimension reduction tool, PHD. We have early discussed the ability of PHD to find curvature in Chapter 7. More discussion on our splitting direction and cut-point selection is given in Section 2. Section 3 summarizes the theoretical foundation of PHD. In Section 4, some examples are given to illustrate the merits of our method. First, two simulation studies are given in section 4.1, highlighting how our method can outperform the three methods included for comparison - CART, SUPPORT and MARS. Then in Section 4.2, our method is used to analyze three real data sets which have been studied before by others using CART, SUPPORT and MARS. In each case, we found that our method yields a simpler tree-structure without compromising the quality of fit.

Our method has the advantage of utilizing the graphical power to explore nonlinear data structure. At each node of a tree, we can focus on the data points in the node and project them along the directions found by PHD. This provides an invaluable opportunity for finding more subtle data patterns. For example, in the Baseball Hitters' Salary Data, a few outliers become visible from our graphical output produced by PHD. As we shall demonstrate, these points also seriously affect the performance of all three methods we have included for comparison. Outliers in high dimensional data are not easy to be ”seen”. Unlike our graphics-oriented procedure, CART, SUPPORT and MARS do not provide outputs that could be effectively used for outlier detection.

In the literature of smoothing, comparison of different methods has placed a good deal of emphasis on the issue of accuracy in prediction. But the fact is that in complex situations, no single method can dominate others in each application in this regard. Simulation studies are useful in promoting new techniques because they at least provide examples where the old methods can be outperformed. The simple simulation settings of Section 4.1 serves this purpose. However, if the intent were to imply the overall superiority of a new method over its rivals, then the simulation should be conducted in a more comprehensive way to avoid the criticism of being “unfair” to some methods. This kind of universal superiority is not what we can demonstrate for our method. Besides, it is certainly not an easy task to decide which set of examples to use in order to achieve a genuine sense of fairness. Nevertheless, the cross-examination strategy taken in Section 5 can be viewed as an attempt in conducting a more extensive simulation study.

Cross-examination is based on the anticipation that a good method should perform better than others for the data simulated according to a model constructed from the method itself. To implement this idea, we first take a real data set and apply method A to build a model. After that, we simulate a new data set from this model and then fit it with both method A and method B to see which method provides better prediction. Of course, we anticipate method A to predict better than method B. But if this turns out not the case, the adequacy of applying method A to the original real data should become doubtful. We call this kind of comparison “a credibility check”. This strategy can be generalized easily to comparison for more than two methods. Our findings on cross-examination are reported in Section 5. For the Ozone example, as expected, each method manages to outperform others for data simulated from its own model. But for the Hitters’ Salary example, there is a somewhat unexpected finding in favor of PHDRT - even for data simulated from the SUPPORT model, PHDRT has a slim edge over SUPPORT.

We have written a package called PHDRT using Xlisp-Stat (Tierney 1990), a user-friendly object-oriented environment. Users can have very easy access to summary information (both numerical and graphical) contained in each node of the tree by pointing to the computer screen. PHDRT offers a rich framework for visualizing, learning, and modeling high dimensional regression data. Further discussion and some concluding remarks are given in Section 6.

16.2 Shape of regression surface and a basic splitting strategy.

The shape of a nonlinear multi-dimensional regression surface is usually hard to describe. This complexity comes mainly from the variation in the gradient vectors for the regression surface at various locations. If all the gradient vectors are the same, then the surface is a hyperplane. This is the case in which multiple linear regression is applicable. For the nonlinear case, gradient vectors do vary and the surface begins to bend or twist. Plotting is a common way to learn about the shape of a two-dimensional surface. But this is obviously hard to do when the dimension gets larger.

Our domain splitting method offers a way to gradually simplify the shape of the surface. The main idea is to find a projection direction where the surface bends the most first and then use this projection to split the data. It is hoped that after recursive splitting, the gradient vectors within each region will become similar so that the regression surface can be adequately approximated by linear regression.

It is not easy to implement this idea in a straightforward manner. The main difficulties come from the estimation of high dimensional gradient vectors. To overcome this problem, we use the method of PHD which does not rely on direct gradient estimation. PHD offers a simple way to find projection directions with bending patterns.

16.2.1 An Example.

Consider a simple case where the regression function consists of two hyperplanes continuously joined together:

$$\begin{aligned} y &= x_5, \text{ if } x_1 + x_2 + x_3 + x_4 \geq .5, \\ &= x_5 - x_1 - x_2 - x_3 - x_4 + .5, \text{ if } x_1 + x_2 + x_3 + x_4 < .5 \end{aligned}$$

where $x_1, \cdots, x_5$ are $i.i.d.$ standard normal random variables. For the sake of illustration, the random error term is purposively ignored. A sample of size 400 is generated. There are only two distinct gradient vectors in this case, $(0, 0, 0, 0, 1)'$ and $(-1, -1, -1, -1, 1)'$. Let $\beta = (1, 1, 1, 1, 0)'$ and denote $(x_1, \cdots, x_5)'$ by $\mathbf{x}$. The best way to describe the regression surface is to project it onto the plane spanned by the axis of $\beta'\mathbf{x}$ and the axis of x_5; see Figure 1(a). It looks like a hinge and β is the direction of bending. We apply the r-based PHD method described in Section 3. The first PHD direction is found to be $\hat{b}_1 = (.51, .50, .48, .49, .04)'$ which is almost in the same direction as β. Figure 1(b) is a graphical output of PHD which reveals the bending pattern very well. The first PHD variate, denoted by $\hat{b}_1'\mathbf{x}$, is in the horizontal axis $phd1$. The vertical axis displays the residuals after the global linear trend obtained from multiple linear regression is removed. We could use y for the vertical axis and in fact this produces a pattern almost identical to Figure 1(a). However, bending patterns are usually better revealed by plotting the residuals. This observation is consistent with a recent study by Cook(1998). PHD can work well for modest sample sizes. For this example, a sample of size 100 is large enough to see a clear bending pattern.

Figure 1(b) strongly suggests that the data be split into two groups at the point where the two line segments meet. The cut-point is estimated to be $\hat{c} = .248$. All points with $\hat{b}_1'\mathbf{x} \geq \hat{c}$ fall into the first group and all other points go to the second group. Since $\hat{b}_1$ is about a half of β, this splitting is quite consistent with what the inequality in the true model would suggest. After splitting, we then apply linear regression in each group. As expected, the fit is very good.

16.2.2 Binary Regression Tree.

Due to its simplicity, binary splitting will be our primary choice in generating a regression tree and this is what we shall describe in this section. But our approach is flexible enough to allow multiple splitting if necessary and we shall encounter such applications later.

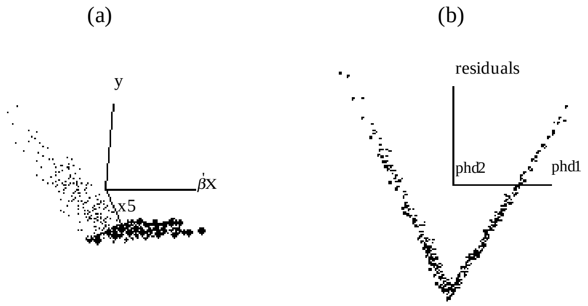


Figure 1: For the simulated data in section 2.1, the 3-D plot (a) shows two planes joined together like a hinge; PHD plot (b) finds the hinge direction.

16.2.3 Splitting.

A binary split of a node s takes the form $b'\mathbf{x} \geq c$, or $< \ c$. Two distinct steps are needed:

(a) Split direction selection. An ideal split direction b is one that clearly reveals a bending surface pattern. We shall use the first r-based PHD direction $\hat{b}_1$ (see (3.1) in section 3) which is obtained by applying PHD to the data contained in the current node. The output PHD plot should reveal some pattern of bending; otherwise splitting will not be fruitful (see stopping rule (s.2) below).

(b) Cut-point. Our procedure seeks for a place in the PHD plot to break the bending pattern so that after splitting, each part in the plot will become more linear. Details of our procedure for selecting the cut-point c are described in Appendix A.

16.2.4 Stopping.

Node s is declared to be a terminal node if one of the following conditions is satisfied:

(s.1) The sample size in s is less than a user-specified value.

(s.2) The PHD direction is insignificant: the first eigenvalue (or the corresponding p-value) is smaller (larger) than a user-specified value.

16.2.5 Prediction.

After successive splitting, the change in the gradient of the regression function within each node is gradually reduced. We shall fit a multiple linear regression in each terminal node.

Suppose the region has been broken into several terminal nodes R_t, $t = 0, \cdots, T$. Let $B_t(\mathbf{x}) = \mathrm{I}(\mathbf{x} \in R_t)$ be the indicator function for node R_t. We represent the final prediction function as

$$\hat{g}(\mathbf{x}) = \sum_{t=0}^{T} (\hat{a}_t + \hat{b}_t'\mathbf{x}) B_t(\mathbf{x})$$

where $\hat{a}_t$ and $\hat{b}_t$ are the least squares linear regression estimates for node R_t; i.e.

$$\min_{a_t, b_t} \sum_{\mathbf{x}_i \in R_t} (y_i - (a_t + b_t'\mathbf{x}_i))^2.$$

16.3 3. Variable selection.

In the literature of tree-structured regression and classification, opinion has been divided regarding the relative merits of univariate split against the linear combination split; see the Breiman and Friedman's discussion of FACT by Loh and Vanichsetakul (1988) and the authors' Rejoinder. From our observation, what is missing from either side's argument is the role of geometry. Since our tree construction is geometrically motivated, splitting directions must be p-dimensional vectors in order to allow enough flexibility in viewing the regression surface from various angles. But a realistic concern is that for large p, these vectors may not be easy to interpret. To address this issue, we use a variable selection strategy similar to the one discussed in Chapter 3 for SIR. For a given PHD direction $\hat{b}$ obtained when all regressor variables are used, we want to find a smaller subset of regressor variables that have a high R-squared value when regressed against $\hat{b}'\mathbf{x}$. Our default is to use the forward variable selection method. We stop recruiting more variables when the R-squared value first reaches 90% or more. After variable selection, we rerun PHD on the selected variables. Since the new split direction now involves fewer regressors, it is easier to interpret. Our program also checks the correlation between the old and the new split variables to make sure that they are nearly the same.

16.4 Examples.

We apply our PHD based splitting method for generating regression trees (abbreviated by PHDRT) to several simulations and real data sets. The performance is then compared with CART, SUPPORT, and MARS. These results illustrate that our method often produces a much simpler tree while maintaining compatible (if not better) predictive accuracy. All work was done on the SUN (SPARC) workstation in the Xlisp-Stat environment (Tierney, 1990). More examples and other details not reported here can be found in the thesis of Lue (1994).

16.4.1 Simulations.

In our simulation study, we shall generate data from models of the form $y = g(\mathbf{x}) + \epsilon$. To measure the predictive accuracy, we can compute the true average squared error of the fit defined by

$$\text{ASE} = \frac{1}{n} \sum_{i=1}^{n} (g(\mathbf{x}_i) - \hat{g}(\mathbf{x}_i))^2$$

where $\hat{g}$ denotes any fit of g. The average of squared residuals is denoted by $\hat{\sigma}^2 = \frac{1}{n} \sum (y_i - \hat{g}(\mathbf{x}_i))^2$. These examples are constructed in order to illustrate situations where PHDRT can dramatically outperform other established methods. It is intended to highlight the value of studying the shape of surface - which is the fundamental difference between the PHDRT method and others. It is not intended for supporting any claim about the total domination of one method over another.

Simulation I.

The regression surface in this example consists of the three sides of a triangle pyramid. The model is

$$\begin{aligned} y &= -\beta_1'\mathbf{x} - \sqrt{3}\beta_2'\mathbf{x} + 1 + .5\epsilon, \text{ if } \beta_2'\mathbf{x} \geq 0, \text{ and} \sqrt{3}\beta_1'\mathbf{x} + \beta_2'\mathbf{x} \geq 0 \\ &= -\beta_1'\mathbf{x} + \sqrt{3}\beta_2'\mathbf{x} + 1 + .5\epsilon, \text{ if } \beta_2'\mathbf{x} < 0, \text{ and} \sqrt{3}\beta_1'\mathbf{x} - \beta_2'\mathbf{x} \geq 0 \\ &= 2\beta_1'\mathbf{x} + 1 + .5\epsilon, \text{ if } \sqrt{3}\beta_1'\mathbf{x} + \beta_2'\mathbf{x} < 0, \text{ and} \sqrt{3}\beta_1'\mathbf{x} - \beta_2'\mathbf{x} < 0 \end{aligned}$$

where $\mathbf{x} = (x_1, \cdots, x_{10})'$, all ten coordinates of $\mathbf{x}$ and ϵ are $i.i.d.$ standard normal random variables, $\beta_1 = (1, 1, 1, 1, 1, 0, \cdots, 0)'$, $\beta_2 = (0, \cdots, 0, 1, 1, 1, 1, 1)'$. Apparently, the multiple linear regression does not fit the data well. The lack of fit can be seen from the scatterplot of the linear residuals against the first two components found by PHD, given in Figure 2.

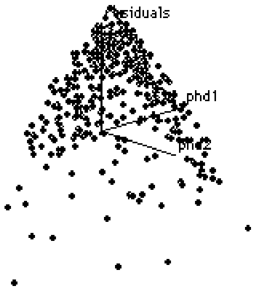


Figure 2: Simulation I. Linear residuals are plotted against the first two PHD component.

A summary of PHDRT analysis is given in Table 1. Reading from the top, this table is divided into several sections. The first section, “linear”, gives the results of the linear regression applied to the entire data set. The R-squared value is only .03 and residual standard deviation is large, $\hat{\sigma} = 2.617$. The application of PHD to the root node (the entire data set, which has 400 cases) leads to the next section, “root 400”. Both the p-value sequence, 0, 2e-15, .69,$\cdots$, and the eigenvalue sequence, 2.3, 2.1, .63, $\cdots$, indicate that two directions are significant. Here the chi-squared test given in Li (1992) is used. The first PHD direction $\hat{b}_1$ should be a vector of 10 dimensions, but to save space, only the coordinates for the 5 selected variables, $x_9, x_{10}, x_6, x_8, x_7$ are given there. Other coordinates are quite small and are omitted. By default, PHDRT uses the first PHD direction $\hat{b}_1$ and the cut-point $\hat{c} = -.01$ to split the data into two subnodes of size 207 and 193 respectively. The cases with $\hat{b}_1'\mathbf{x} = .48x_9 + .48x_{10} + .50x_6 + .42x_8 + .37x_7 \geq \hat{c} = -.01$ are assigned to the first subnode. All other cases go to the second subnode.

The application of PHD to the first subnode leads to the third section, "subnode 207". Both p-values (6e-9, .91,$\cdots$) and eigenvalues (2, .39,$\cdots$) indicate that only one direction is significant. The coordinates of the first PHD direction $\hat{b}_1$ with the 5 selected variables are given there. Using this direction and the cut-point $\hat{c} = -.57$, this subnode is further split into two children nodes of size 150 and 57 respectively. Again PHD is applied to each of these two nodes. But this time, no more split is recommended because no significant PHD directions are found. Therefore they become terminal nodes and are called "T150" and "T57" respectively. The linear models are used to fit the data in each terminal node. The R-squared values are .91 and .94 respectively, and the residual standard deviations are about .84 and .65 respectively. These are recorded in sections "T150" and "T57".

The PHD analysis for the second subnode (with 193 cases), is quite similar. The results are given under the section "subnode 193". This node is split into two children nodes, "T124" and "T69" and then the process is terminated. The R-squared value and the residual standard deviation for fitting a linear model in each terminal node are also given there. The last section, "Overall", gives (1) the residual standard deviation for the entire data set using PHDRT, and (2) the performance measure ASE as defined in the beginning of this section.

We also apply CART, SUPPORT, and MARS to see how well they perform. The trees generated by CART, SUPPORT and PHDRT are shown in Figure 3. The main findings are:

1. CART doesn't make any split. The ASE obtained by CART is 6.528.

2. SUPPORT chooses five variables, x_1, x_4, x_6, x_7 and x_9, and makes seven splits; that is, after first splitting by x_1 at .478e-1, then one subnode is split by x_7 at .132 and the other subnode is split by x_4 at .23e-1, etc.. Note that in applying SUPPORT, we have followed the guidance of Chaudhuri et al (1994) and select the user-specified parameters by trying several values. The best choice is found at $f = .1$, $\eta = .2$, MinDat = 40, V = 10. The ASE obtained by SUPPORT is 3.492, which represents an 85% (= $6.528/3.492 - 1$) improvement over CART.

3. MARS needs twelve basis functions to fit the surface. To save space, the results are omitted. We note however that the ASE obtained by MARS is 4.359 which is between CART and SUPPORT.

4. As described earlier, PHDRT finds four terminal nodes. The ASE obtained by PHDRT is only .371. The improvement over SUPPORT is 840%(=3.492/.371 -1).

Simulation II.

Our second simulation uses the following model:

$$y = |\beta_1'\mathbf{x}| + .5(|\beta_2'\mathbf{x}| - \mu) \cdot sign(\beta_1'\mathbf{x}) + .5\epsilon$$

Table 1: **A summary output of PHDRT for simulation I.** The root node is split into two subnodes of size 207 and 193 respectively. The node with 207 cases is further split into two terminal nodes, T150 and T57, while the node with 193 cases is split into terminal nodes, T124 and T69.

linear	R-squared	.03	$\hat{\sigma}$	2.617
root	p-values	(0, 2e-15, .69,$\cdots$)	eigenvalues	(2.3, 2.1, .63,$\cdots$)
	$\hat{b}_1$	(.48, .48, .50, .42, .37)	$\hat{c}$	-.01
400	variables	$x_9, x_{10}, x_6, x_8, x_7$	subnode sizes	(207, 193)
subnode	p-values	(6e-9, .91,$\cdots$)	eigenvalues	(2, .39,$\cdots$)
	$\hat{b}_1$	(.51, .48, .42, .36, .31)	$\hat{c}$	-.57
207	variables	x_1, x_4, x_3, x_5, x_2	subnode sizes	(150, 57)
subnode	p-values	(5e-10, .68, $\cdots$)	eigenvalues	(2, .44,$\cdots$)
	$\hat{b}_1$	(.56, .52, .46, .46, .33)	$\hat{c}$	-.28
193	variables	x_3, x_2, x_5, x_4, x_1	subnode sizes	(124, 69)
T150	R-squared	.91	$\hat{\sigma}$	.8406
T57	R-squared	.94	$\hat{\sigma}$	.6497
T124	R-squared	.95	$\hat{\sigma}$	.5736
T69	R-squared	.82	$\hat{\sigma}$	1.266
Overall	$\hat{\sigma}$	.8389	ASE	.371

where $\beta_1 = (1, 1, 0, \cdots, 0)'$, $\beta_2 = (0, 0, 1, \cdots, 1)'$, $\mu = E|\beta_2'\mathbf{x}|$ and $\mathbf{x} = (x_1, \cdots, x_{10})'$, all ten coordinates of **x** and ϵ are $i.i.d.$ standard normal random variables; sign(z) = 1 if z $\geq$ 0, otherwise -1. The plot of the linear residuals against the first two components found by PHD is given in Figure 4. Again, a bending pattern is found. Table 2 summarizes the PHDRT analysis. As in Table 1, it is divided into several sections, each representing results from one node. It is interesting to observe that although this model involves two directions, the PHD analysis on the root node only gives one significant direction. However, the other direction is recovered after splitting.

The results of trees generated by CART, SUPPORT, and PHDRT are shown in Figure 5. MARS is also applied. Here is a summary of the relative performance.

1. CART chooses x_1 and makes six consecutive splits; that is, after first splitting by x_1 at 1.120, one subnode is split again by x_2 at -1.232, etc. The ASE found by CART is 1.011.

2. SUPPORT chooses four variables, x_1, x_2, x_3 and x_6, and makes six splits; that is, after first splitting by x_1 at .683e-1, one subnode is split by x_2 at .69e-2 and the other subnode is split by x_2 at -.651e-1, etc. (Here again after several passes, we take f = .1, $\eta = .1$, MinDat = 40, V = 10 as the user-specified values). The ASE obtained by SUPPORT is .681, which is about 50% better than CART.

3. MARS uses 12 basis functions. The ASE is .727, which is about the same as SUPPORT.

4. Four terminal nodes are found by PHDRT. The ASE obtained by PHDRT is only .098, which is about 600% better than SUPPORT or MARS.

Again for this example, the surface has a nice geometric shape. The dramatic improve-

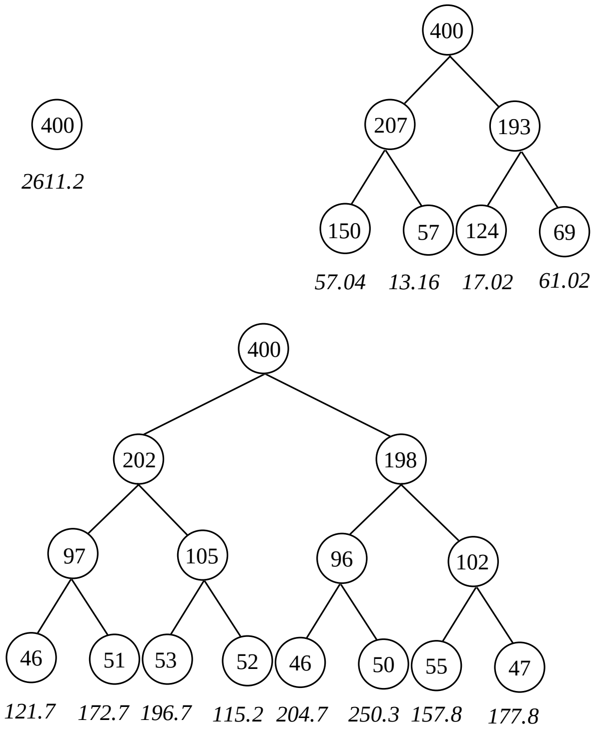


Figure 3: CART(upperleft), SUPPORT(lower) and PHDRT trees for simulation I, the value of SAE are 6.528, 3.492 and 3.71, respectively. The number in each node is the sample size, and sum of squared error loss is given in italics under each terminal node.

ment of PHDRT over other methods is simply because PHDRT is the only one that uses geometric information.

16.4.2 Real Data.

We apply CART, SUPPORT, MARS and our method to three popular data sets. The first data set comes from a study on the atmospheric ozone concentration in the Los Angeles basin; see Breiman and Friedman (1985). The second one concerns the salary of 263 baseball hitters; it is originally given in 1988 ASA Graphics Poster Session (c.f. Chaudhuri et al., 1994). The last one studies the fuel efficiency for automobiles; it comes from the ASA Data Exposition data set (1983) collected by Ernesto Ramos and David Donoho. The names of the variables involved are given in Tables 3, 4, 5. The numbers of regressors are 8, 16, and 6 respectively.

The Ozone data.

For the ozone data, PHDRT makes only one split with the splitting variable $.004x_3+.0005x_2+.03x_6$, and the cut-point is at 3.46. The fitted equation in each node is given in Table 6. The standard errors of the regression coefficients are also given. It is interesting to observe that

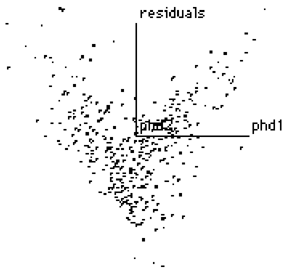


Figure 4: Simulation II. Plot of linear residuals against the first two PHD components.

Table 2: PHDRT's results for simulation II.

linear	R-squared	.04	$\hat{\sigma}$	1.329
root	p-values	(2e-15, .74, .91, $\cdots$)	eigenvalues	(1.1, .34, .27, $\cdots$)
	$\hat{b}_1$	(.73, .76)	$\hat{c}$	.05
400	variables	x_1, x_2	subnode sizes	(205, 195)
subnode	p-values	(2e-11, .23, $\cdots$)	eigenvalues	(1.2, .45, $\cdots$)
	$\hat{b}_1$	(.38, .44, .33, .42, .38, .33, .21)	$\hat{c}$	.09
205	variables	$x_5, x_9, x_6, x_4, x_7, x_3, x_{10}$	sizes	(88, 117)
subnode	p-values	(1e-6, .80,$\cdots$)	eigenvalues	(1.1, .32,$\cdots$)
	$\hat{b}_1$	(.45, .29, .37, .48, .35, .28, .21)	$\hat{c}$	.06
195	variables	$x_3, x_8, x_9, x_{10}, x_7, x_4, x_6$	subnode sizes	(97, 98)
T88	R-squared	.81	$\hat{\sigma}$	.6182
T117	R-squared	.79	$\hat{\sigma}$	.6505
T97	R-squared	.87	$\hat{\sigma}$	.4773
T98	R-squared	.86	$\hat{\sigma}$	.5596
Overall	$\hat{\sigma}$	.5829	ASE	.098

the regression coefficient for x_6 (humidity) is positive in one node but changes to a negative value in the other node.

To see how PHDRT explores the nonlinearity of regression surface, Figure 6 provides the view of log ozone from the split direction found by PHDRT. We see that the data points form a clear pattern of bending. The data points on the right of the cut-point belong to the first node, the node with 114 cases. The rest of the data points go to the second node, the node with 216 cases. After splitting, no more bending patterns are seen and each node is fitted adequately with a linear model. The overall standard deviation for the residuals $\hat{\sigma}$ is about .37.

For a comparison, we run CART, SUPPORT and MARS. It turns out that CART chooses four variables, x_1, x_2, x_4 and x_3, to make four splits, and yields an overall standard deviation $\hat{\sigma} = .43$. The first split is made by x_1 at 65.5; then one subnode is split by x_2 at 3669, and so

Table 3: Variables for Ozone data

y	Logarithm of daily maximum one-hour-average ozone reading at Upland, CA.
x_1	Temperature (degrees F) measured at Sandburg, CA.
x_2	Inversion base height (feet) at LAX
x_3	Pressure gradient (mm Hg) from LAX to Daggett, CA.
x_4	Visibility (miles) measured at LAX
x_5	500 millibar pressure height (m) measured at Vandenberg AFB
x_6	Humidity (%) at LAX
x_7	Inversion base temperature (degrees F) at LAX
x_8	Wind speed (mph) at Los Angeles International Airport (LAX)

Table 4: **Hitters' salary data in 1986.** The variables with the symbol * refer to the entire career up to 1986. Those variables without * refer to the year of 1986 only.

y	the logarithm of 1987 annual salary ($1000)		
hline x_1	# of times at bat	x_9	# of hits *
x_2	# of hits	x_{10}	# of home runs *
x_3	# of home runs	x_{11}	# of runs *
x_4	# of runs	x_{12}	# of runs batted in *
x_5	# of runs batted in	x_{13}	# of walks *
x_6	# of walks	x_{14}	# of put outs
x_7	# of years in major leagues	x_{15}	# of assists
x_8	# of times at bat *	x_{16}	# of errors

on. SUPPORT chooses two variables, Pressure (x_3) and Humidity (x_6), to make two splits and yields an overall standard deviation $\hat{\sigma} = .38$. After the first split by Pressure at 17.2, one subnode is split by Humidity at 46.9. The user-specified parameters are optimally set at f = .2, $\eta = .4$, MinDat = 30, V = 10. MARS uses 10 basis functions and produces a $\hat{\sigma}$ of .35. The trees obtained by CART, SUPPORT and PHDRT are given in Figure 7.

It appears that from the value of $\hat{\sigma}$, there is not much difference between the fits produced by MARS, SUPPORT, and PHDRT. But PHDRT is the only one which provides us with the additional geometric insight about the nonlinear data pattern.

Remark 4.1 Variable selection as discussed in Remark 3.1 has been used in all three real data examples. For the ozone data, the three variables x_3, x_2 and x_6 used in defining the splitting variable have an R-square value of about 92% when regressed against the first PHD component obtained from the initial run of PHDRT on all 8 variables.

Table 5: Regressors description for Cars data

y	miles per gallon		
x_1	number of cylinders	x_4	vehicle weight
x_2	engine displacement	x_5	acceleration
x_3	horsepower	x_6	model year

Table 6: **Piecewise linear fit for Ozone data.** PHDRT makes one split, generating two terminal nodes. This table gives the fitted equation for each node, along with the standard errors of the regression coefficients, the residual standard deviation, and the R-squared value.

Node sizes	Fitted equations	Standard errors	$\hat{\sigma}$	R-squared
114	$2.42 + .032x_1 - .00026x_2 - .019x_6$	.44, .0037, .000049, .0034	.378	.73
216	$-.17 + .027x_1 + .015x_6$	.12 , .0021, .0014	.373	.72

Hitters' salary data.

Real data often contain outliers that could affect model building and any related statistical inference activities. Earlier study on the Hitters' salary data (Chaudhuri et al. 1994) paid no attention to this problem. This is in part due to the fact that outliers are hard to detect in large dimension problems. Unlike CART, SUPPORT, and MARS, outlier detection is a natural byproduct of the graphical feature of PHDRT. This is what we shall demonstrate in this example.

We take the logarithm of the salary as the y variable, as recommended by Chaudhuri et al. (1994). We first run PHDRT on the entire data set. Seven outliers are clearly seen from various angles of the 3-D PHD plot; Figure 8 (a)-(c). We remove these points and then run PHDRT again on the remaining 256 cases. A clear bending pattern is now revealed in the PHD plot; Figure 8 (d).

Table 7: **Piecewise linear fit for Hitters data**

Node sizes	Fitted equations	Standard errors	$\hat{\sigma}$	R-squared
144	$3.95 + .00540x_6 +$ $.0860x_7 + .000877x_8$	.080, .0015 .019, .000057	.318	.838
112	$6.23 + .0104x_6 -$ $.104x_7 + .000225x_8$	.17 , .0019 .019, .000037	.397	.53

The final output of PHDRT is summarized in Table 7. Only one split is made. The node with $.121x_7 + .00081x_9 < 1.29$ has 144 cases. According to Table 4, x_7 is the number of years in major leagues, and x_9 is the total number of hits accumulated in the entire career(up

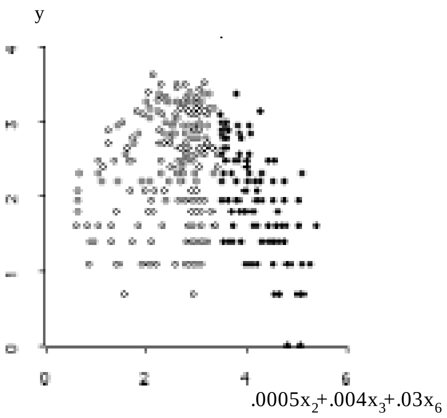


Figure 6: View of log ozone from the split direction. The cut-point is at 3.46. The data form an obvious bending pattern which is the key geometric information explored by PHDRT.

to year 1986). Thus players in this group are either relatively junior or are not hitting well. The other node, $.121x_7 + .00081x_9 \geq 1.29$ has 112 cases. Veteran players with better hitting record should fall into this node. The predictor variables in the final fitted equations for both nodes happen to be the same $-x_7$, x_8 (the total number of times at bat during the career) and x_6 (the number of walks in 1986). An interesting finding related to "aging effect" can be observed in the second node. The negative sign in the coefficient of x_7 is highly significant. For this group(senior/better hitters), x_7 is working against players' salary. But for the other group(junior/poor hitters), x_7 is a positive factor. Figure 9 (a) gives the plot of log salary (y) against the split variable. The two groups are seen to be markedly different. Those points on the left of the cut point have a sharper linear trend. This is consistent with the R-squared values reported in the rightmost column of Table 7. The R-squared value for the first node is about 84% and it goes down to about 53% for the second node.

Some further analysis details are described in Remark 4.2. We also apply CART, SUPPORT, and MARS to the entire data set. The trees generated by CART, SUPPORT and PHDRT are given in Figure 10. In summary, CART chooses four variables, x_8, x_9, x_2 and x_{12}, to make five splits, and yields the overall standard deviation for $\hat{\sigma} = .422$. The first split is made by x_8 at the value 1452, then one subnode is split by x_9 at 182 and so on. SUPPORT chooses two variables, x_7 and x_1, to make two splits, and yields the overall standard deviation for $\hat{\sigma} = .440$. The first split is made by x_7 at 7.31, then one subnode is split by x_1 at 406. Here the user specified parameters are optimally set at f = .1, $\eta = .4$, MinDat = 40, V = 10 after several passes. MARS uses 13 basis functions and yields a $\hat{\sigma}$ of .33.

By comparing the residual standard deviation $\hat{\sigma}$, it appears that MARS is doing better than CART and SUPPORT. It also beats PHDRT slightly($\hat{\sigma}$ for PHDRT is .35). This is done even without having to pay attention to the 7 outliers detected by PHDRT. Of course, a smaller $\hat{\sigma}$ does not necessarily guarantee the superiority in prediction. However, this raises the question of whether PHDRT does find real outliers or not. It is certainly possible that

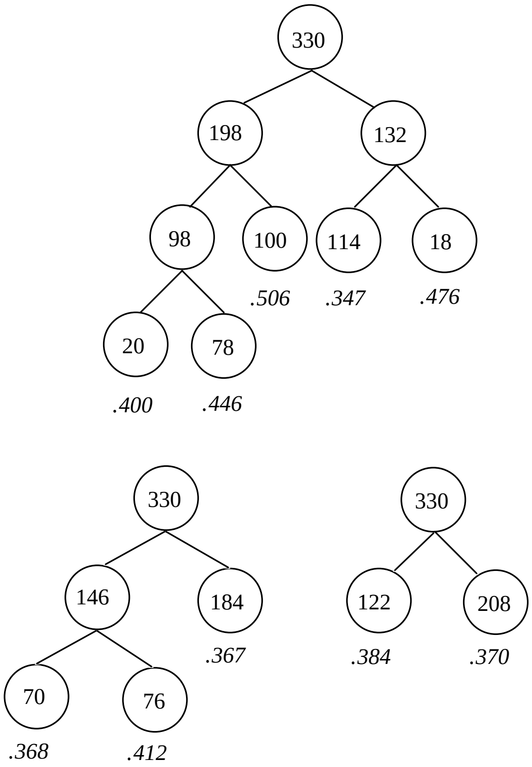


Figure 7: CART(upper), SUPPORT(lower left) and PHDRT trees for atmospheric Ozone concertration in the Los Angeles basin data, the values of the overall residual standard deviation are .434, .378 and .375, respectively. The number in each node is the sample size.

these points may have no effects on procedures other than the PHDRT itself.

To investigate this issue, we leave the 7 cases aside for the moment and randomly partition the remaining 256 cases into two groups – a training group of 200 cases and a testing group of 56 cases. Then we apply each of the four methods to the training group in order to produce the prediction rules. After obtaining the prediction rules, we then use each of them to make prediction for cases in the testing group and the 7 cases considered to be outliers by PHDRT. To evaluate the performance in prediction, the average of the squared prediction errors (ASPE) is computed for the testing group and separately for the 7-outlier group. If there were nothing special about the 7-outlier group, then we expect the two ASPE values to be about the same.

We repeat this random partition and prediction process 10 times and a summary is given in Table 8. Cross the board, the ASPE for the outlier group is one magnitude bigger than the ASPE for the testing group for all methods considered here. This indicates well that the outliers found by PHDRT also affect the behavior of MARS, CART, and SUPPORT. This table also shows that once the outliers are deleted, the prediction errors are about the same for PHDRT and MARS. This is rather interesting because the fitted equation for MARS takes a complicated expression which in no way resembles the simplicity of the fit by PHDRT.

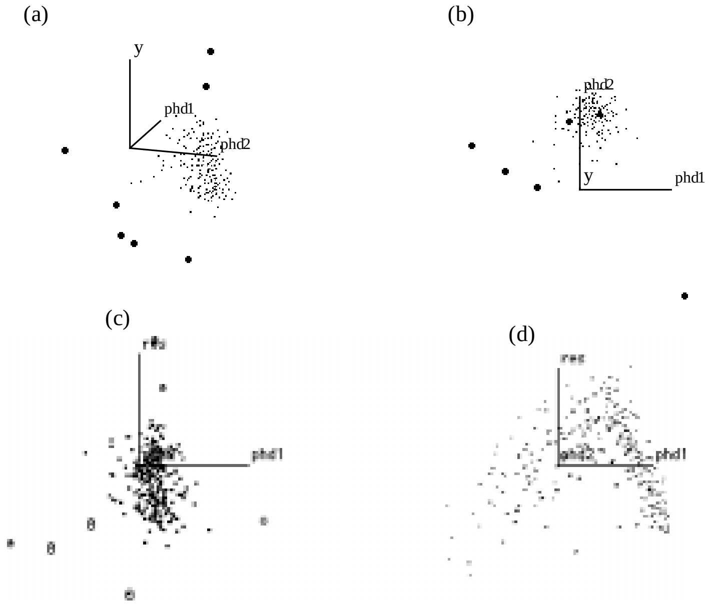


Figure 8: Baseball salary data. Seven outliers are clearly seen from various angles of the 3-D PHD plot (a)-(c). A clear bending pattern is revealed in the PHD plot (d) once they are removed.

Another point worth mentioning is that for these ten different runs, PHDRT always leads to the same tree structure - only one split is needed. On the other hand, the basis functions yielded by MARS vary substantially from run to run. For example, the first run yields a fitted equation:

$$\begin{aligned}\hat{y} \;=\; & 7.00 - .00501(245 - x_{11})_+ - .127(x_7 - 7)_+ - .121(7 - x_7)_+ \\ & + .000000798(x_{11} - 245)_+(x_{14})_+ + .00257(7 - x_7)_+(x_{12} - 274)_+ \\ & + .00000354(x_{11} - 245)_+(x_{10} - 43)_+ - .000324(245 - x_{11})_+(x_{10} - 45)_+ \\ & + .00529(x_6 - 3)_+ - .000513(1906 - x_9)_+ \qquad (4.1)\end{aligned}$$

while for the second run, the fit is

$$\begin{aligned}\hat{y} \;=\; & 4.73 + .00207(x_9 - 489)_+ - .00331(489 - x_9)_+ - .0885(x_7 - 6)_+ - .173(6 - x_7)_+ \\ & + .0126(x_6 - 3)_+ - .209(489 - x_9)_+(x_{10} - 45)_+ - .0263(489 - x_9)_+(45 - x_{10})_+ \\ & + .0194(x_{10} - 221)_+ - .000583(x_8 - 5347)_+ + .000416(5347 - x_8)_+ \\ & - .000236(x_6 - 3)_+(x_{10} - 221)_+ - .0000323(x_6 - 3)_+(x_{10} - 221)_+ \\ & + .00904(6 - x_7)_+(x_{10} - 49)_+. \qquad (4.2)\end{aligned}$$

The output of MARS also provides a relative variable importance index (RVII). For the first run, the most important variable is X_7 (RVII=100), followed by x_{11}(RVII= 79), x_{10}(RVII=41), and so on. But in the second run, the most important variable is x_9(RVII=100), followed by

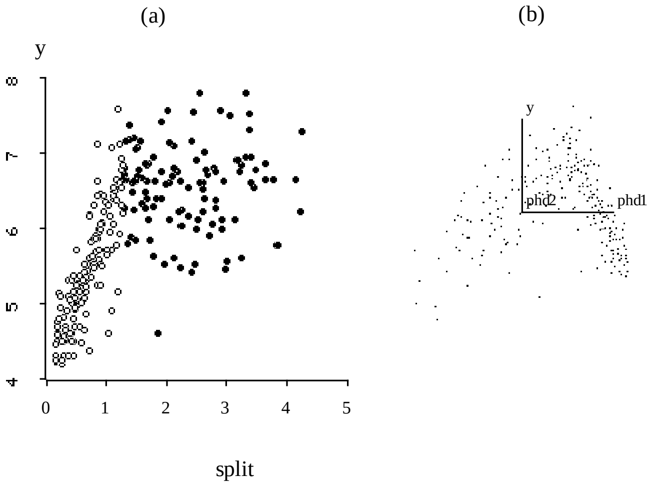


Figure 9: (a) The plot of log salary against the split variable shows a bending pattern. The cut-point is at 1.29. (b). The corresponding PHD plot-the horizontal azis is the same as in (a).

x_7(RVII=90), x_6(RVII=64), and so on. This kind of variation adds further difficulties in interpreting the MARS output.

Remark 4.2. Variable selection as described in Remark 3.1 is repeated twice in simplifying the splitting direction. At first time, seven out of the 16 regressors can explain more than 92% of the variation in $\hat{b}_1'\mathbf{x}$. We use the seven regressors found as the new regressors to rerun PHDRT. Then variable selection is applied again to the new splitting variable. This time we need only two regressors, x_7 (# of years in major leagues) and x_9 (# of hits during his career), could explain more than 94% of the variation. The final round of PHD analysis is based using only x_7 and x_9. This gives the split direction of the PHD tree. The PHD plot is given in Figure 9 (b), which exhibits the bending pattern very well.

Cars.

The default of PHDRT is to use a binary split. However, other possibilities can sometimes be more effective. This example illustrates how the graphical output of PHD can be used for this purpose.

There are 392 cars used in the analysis. The scatter plot of linear residuals against the first two components found by PHD (with x_3, x_6, x_4 and x_1 as regressors), shown in Figure 11, reveals three clusters. This pattern can be attributed to the factor x_1 (the number of cylinders). We split the data accordingly. The first group consists of cases with $x_1 = 3, 4, 5$, the second group with $x_1 = 6$, and the third group with $x_1 = 8$. After splitting, linear models fit reasonably well for each group.

A comparison with other methods is outlined here.

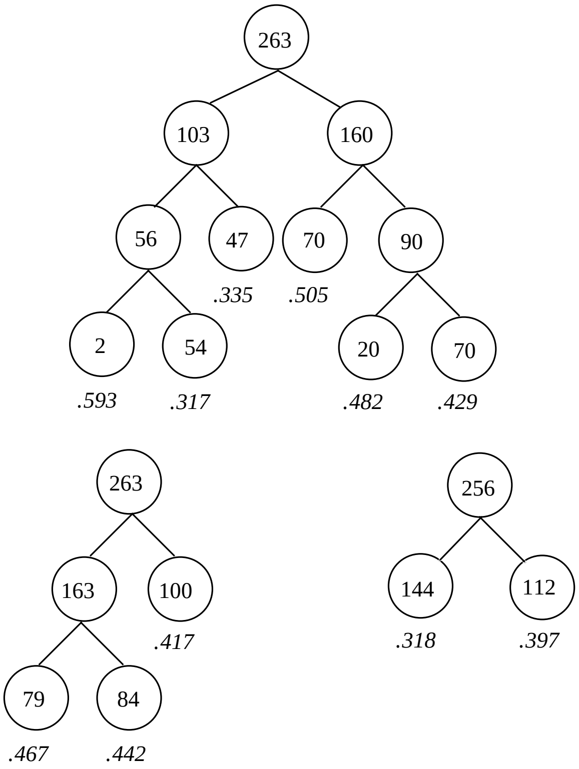


Figure 10: CART(upper), SUPPORT(lower left) and PHDRT trees for Baseball Player Salary data, the values of the overall residual standard deviation are .422, .440 and .375, respectively. The number in each node is the sample size.

1. CART chooses four variables, x_2, x_3, x_6 and x_4, and makes six splits, yielding an overall standard deviation $\hat{\sigma} = 3.24$. After the first split by x_2 at 190.5, one subnode is split by x_3 at 70.5, and subsequently one of the daughter nodes is again split by x_3 at 127, etc.

2. SUPPORT chooses four variables, x_4, x_1, x_6 and x_5, and makes five splits, yielding an overall standard deviation $\hat{\sigma} = 2.64$. After the first split by x_4 at .298e4, one subnode is split by x_1 at 4.21. The one of the daughter nodes is split again by x_6 at 77, etc. (User specified values: f = .2, $\eta = .4$, MinDat = 20, V = 10)

3. The $\hat{\sigma}$ obtained by MARS is 2.47 with twelve basis functions.

16.5 Cross-Examination.

The purpose of cross-examination is to see how one method performs when the data are simulated from models constructed by other methods.

Table 8: **Prediction errors for Hitters data.** Setting aside the 7 outliers found by PHDRT, the rest of cases are randomly partitioned into a training set (200 cases) and a testing set (56 cases). The ASPE for the testing group is compared to the outlier group. The results of 10 random partitions are given here.

Procedure	-	ASPE for 10 runs	Mean of ASPE	SD of ASPE
CART	test(56)	.212 .219 .265 .205 .235	.2197	.0361
		.255 .260 .219 .162 .165		
	outliers(7)	2.42 2.19 2.31 2.46 2.51	2.414	.129
		2.57 2.59 2.45 2.29 2.35		
SUPPORT	test(56)	.162 .128 .236 .112 .145	.1618	.0355
		.154 .149 .170 .160 .202		
	outliers(7)	4.46 4.21 4.70 4.23 4.14	4.489	.364
		4.48 4.09 4.59 5.31 4.68		
MARS	test(56)	.106 .127 .191 .085 .141	.1344	.0306
		.134 .106 .156 .139 .159		
	outliers(7)	4.53 4.33 3.86 5.97 4.78	4.709	.364
		4.26 4.90 5.07 5.43 3.96		
PHDRT	test(56)	.137 .120 .188 .093 .143	.1452	.0319
		.150 .119 .198 .139 .165		
	outliers(7)	4.32 3.96 4.40 4.36 4.42	4.406	.380
		4.14 3.96 4.80 5.22 4.48		

We take the ozone data to begin with. In Section 4.2.1, we have fitted this data set with four methods, CART, SUPPORT, MARS, and PHDRT. Denote the fitted function by $\hat{g}_j(\mathbf{x})$, $j = 1, \cdots, 4$, respectively. We use each of these functions to generate a data set. More precisely, the jth data set is generated according to

$$y_i^* = \hat{g}_j(\mathbf{x}_i) + \epsilon_{ji}, \ \ i = 1, \cdots, 330.$$

The noise ϵ_{ji} is generated from a normal distribution. For the MARS model ($j = 3$), errors are assumed homogeneous with the common standard deviation equal to .35 (this is the noise level as obtained by MARS in Section 4.2.1). For the tree-based methods, $j = 1, 2, 4$, the variance of ϵ_{ji} is set to the residual variance found from each terminal node of the constructed tree. After these data sets are simulated, we then apply CART, SUPPORT, MARS, and PHDRT to see how well the fit is. As in Section 4.1, the average of squared differences between the fitted values and the true values (ASE) is computed for each method. The results are given in Table 10. As expected, each method has smallest ASE for the data set generated under its own model. No method is completely dominant.

We repeat the same cross-examination process for the hitters data (256 cases). The result is given in Table 11. This time we are a bit surprised to find that PHDRT outperforms SUPPORT for the data set generated by SUPPORT's model. To see if this finding is purely

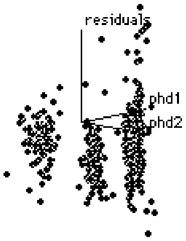


Figure 11: The Sactterplot of linear residuals against the first two PHD components shows three groups for the CARS data.

Table 9: Piecewise linear fit for Cars data

Node sizes	Fitted equations	Standard errors	residuals	R-squared
206	$-37.4 + 5.18x_1 - .0457x_2$	7.7, 1.4, .018	3.18	.70
	$-.071x_3 - .00568x_4 + .903x_6$	.022, .0011, .064		
83	$-8.11 - .00577x_4 + .613x_6$	7.6, .0009, .094	2.78	.49
103	$-4.67 + .0182x_2 - .0361x_3$	6.0, .006, .011	1.77	.62
	$-.0028x_4 + .414x_6$	.00046, .071		

accidental, we rerun this part of simulation 9 more times. The results are given in Table 12. PHDRT still has an edge. It is also significantly better than MARS. A lengthier account of our findings can be found in Lue (1994).

16.6 Concluding remarks and further study.

There are several approaches to regression surface fitting without parametric model assumptions. Compared to other smoothing techniques, tree-based regression produces surfaces which are only piece-wise continuous (constant, or linear). In fact, the lack of continuity between neighboring nodes has become a major criticism against such methods in applications where some prediction is to be made near the node boundary. Some remedy can be pursued if we are willing to sacrifice the simplicity of the functional structure. One such method recommended in Chaudhuri et al. (1994) is to suitably average out the estimates from neighboring nodes in a way similar to kernel regression.

Despite this criticism, the most distinctive feature for tree-structured analysis is that it offers a map with guided paths to visit various regions in the multi-dimensional regressor space. To enhance the merit of this feature, our approach places greater emphasis on the geometric shape of the regression surface. We choose a direction for splitting only when data projected along that direction reveals a bending nonlinear pattern. The method of PHD is used to find such directions.

Table 10: **Cross-examination for Ozone example.** Each row gives the ASE for each of the four methods applied to the data set which is simulated according to the indicated model.

	CART	SUPPORT	MARS	PHDRT
CART's model	.005	.062	.056	.084
SUPPORT's model	.078	.018	.024	.027
MARS's model	.078	.034	.014	.024
PHDRT's model	.070	.021	.019	.010

Table 11: **Cross-examination for Hitters example (256 cases).** Each row gives the ASE for each of the four methods applied to the data set which is simulated according to the indicated model.

	SUPPORT	MARS	PHDRT
SUPPORT's model	.042	.039	.036
MARS's model	.048	.024	.024
PHDRT's model	.035	.033	.017

No cross-validation rules are used yet in PHDRT. But as a trade-off, we take full advantage of modern graphical facilities and rely on the inspection of PHD plots and other graphs for diagnostic purposes. Like standard regression, the goal of tree-structured analysis should not be limited to prediction only. By combining the merits of dimension reduction, data visualization, and guided partition, PHDRT has enhanced the utility of tree-structured regression.

We have not yet addressed the issue of how to handle discrete regressors. In our view, discrete variables are geometrically rather different from continuous ones. The visualization issue of high dimensional discrete variables is often pursued via the approach of multidimensional scaling and correspondence analysis. In general, the idea is to find good scoring methods that can reasonably convert the discrete variables into continuous ones. Once this is done, then PHDRT may be applied. There are other possibilities. For example, Filliben and Li (1997) has analyzed a data set with 10 binary regressors using a tree-structured method which is also guided by PHD. We shall discuss this paper in the next chapter.

In addition to analyzing the real data, PHDRT can be used in other ways. For example,

Table 12: Cross-examination for Hitters example (256 cases) generating by SUPPORT's model.

Procedure	ASE for 10 runs	Mean of ASE	SD of ASE
SUPPORT	.040 .023 .031 .042 .030 .035 .026 .040 .040 .023	.0330	.0074
MARS	.035 .044 .048 .046 .040 .039 .048 .041 .042 .040	.0423	.0042
PHDRT	.028 .022 .032 .036 .028 .037 .026 .035 .033 .032	.0309	.0048

it can be used to help visualize the geometric structure of the complicated MARS fit. This is easily done by treating the MARS-fitted values $\hat{y}$ as the outcome variable y, and then applying PHDRT. In Section 4.2.2, we have seen major differences between the two MARS fits, (4.1) and (4.2). It is perplexing that such apparent substantial structural difference does not seem to affect the final prediction too much when applying to the testing group (see Table 8). To seek for a possible explanation, we turn to geometry. It is likely that the algebraic difference in the formula might not produce major difference in the shape of the surface.

A follow-up study has been conducted to compare the difference between the above two MARS fits given by (4.1) and (4.2). We run PHDRT twice, one run for each MARS fit. We find that the fit of PHDRT is very good in both runs. The overall R-squared value is 96.7% for the first run and 94.7% for the second run. The outputs of PHDRT are very similar. In each case, only one split is needed and the plot of the MARS fit against the splitting variable (found by PHDRT) shows a similar bending pattern. The splitting variable in each case can be explained by a single variable x_7 with a very high R-squared value, 99.5% and 97.7% respectively. This kind of study reassures the value of exploring the geometric aspect of high dimensional data. The same structure can be approximated well by drastically different algebraic formulae.

Appendices

A. Cut-point selection.

(1). The goal is to find a place in the PHD plot (residual $\hat{r}$ against $\hat{b}'_{phd}\mathbf{x}$) to break the bending pattern. Let $z = \hat{b}'_{phd}\mathbf{x}$ and denote the mean of z by μ.

(2). Tentatively set c=μ and separate the whole data points, denoted by S, on the PHD plot into two parts: $S_p = \{(z, \hat{r}) : z \geq \mu\}$ and $S_m = \{(z, \hat{r}) : z < \mu\}$. Now in each part, fit a regression line l_i of $\hat{r}$ against z separately and let the residual standard deviation be denoted by $\hat{\sigma}_i$, $i = p, m$. There are two cases to consider:

I) The case $\hat{\sigma}_p \geq \hat{\sigma}_m$ indicates that a better cut point may be found on the right hand side of the current c. Now follow these steps:

- **(a)** Denote μ_p as mean of $\{z : z \geq \mu\}$, let $S_p^+ = \{(z, \hat{r}) : z \geq \mu_p\}$. Fitted a regression line l_p^+ of $\hat{r}$ against z for $(z, \hat{r}) \in S_p^+$. Located the intersection c_p^+ of the two lines l_m and l_p^+.
- **(b)** Reset $c = c_p^+$ as the split knot and split S into two parts, S_p, S_m, according to $z \geq or < c_p^+$. Fit a simple linear regression separately to each part and evaluate $\hat{\sigma}_p$ and $\hat{\sigma}_m$. Also find the weighted standard deviation $\hat{\sigma}_w$ from $\hat{\sigma}_p$ and $\hat{\sigma}_m$, weighted by the number of cases in each part.
- **(c)** If the $\hat{\sigma}_w \geq \epsilon$ (a user-specified small value for terminating the iteration), replace μ with c_p^+, go to the beginning of (2), and iterate. Otherwise c_p^+ is declared to be the final split knot c.

II) The case $\hat{\sigma}_p < \hat{\sigma}_m$ is just the opposite of case I), indicating that a better point should can be found on the left side of the current c. The following steps are analogous:

- **(a)** Denote μ_m as mean of $\{z : z < \mu\}$, let $S_m^- = \{(z, \hat{r}) : z \leq \mu_m\}$. Fitted a regression line l_m^- of $\hat{r}$ against z for $(z, \hat{r}) \in S_m^-$. Located the intersection c_m^- of the two lines l_p and l_m^-.
- **(b)** Reset $c = c_m^-$ as the split knot and split S into two parts according to $z \geq or < c_m^-$. Compute $\hat{\sigma}_p$ and $\hat{\sigma}_m$ for the simple linear fit for each parts, then evaluate $\hat{\sigma}_w$.
- **(c)** If the $\hat{\sigma}_w \geq \epsilon$, replace μ with c_m^-, go to the beginning of Step (2) and iterate. Otherwise c_m^- is declared to be the final split knot c.

C. Program Description.

This section gives some details about how to use the software we developed for “PHDRT” (principal Hessian directions regression tree). Our object-oriented programming package is modified from a SIR-based classification tree package (Chen 1993, Chen and Li 1994). To conduct a “PHDRT” analysis, we first type `(load "phdrt-init")` in Xlispstat. On the screen, we find one menu bar which includes two menus, *pHdrt* and *Plot-tools*. For selecting a data file you want to analyze, you first highlight Open Data item from the pull-down menu pHdrt. A dialog box appears. You simply choose the item you want and click OK.

c.1 Initial Data Analysis: The main goal is to visually inspect data patterns before we make further analysis. There are five plotting options for viewing. You highlight *IDA-options* item from the pull-down menu pHdrt to find a dialog box which includes scatterplot-matrix, grand tour, principal component analysis (pca), histogram and Principal Hessian Directions (PHD) analysis. To see all plots, just click OK.

c.2 Tree-structured Analysis: Highlight *Regression-Tree* item from the pull-down menu pHdrt. This creates a dialog box which includes two items, Automatic and Interactive mode, for constructing a tree. You can choose either mode.

(a) *Automatic* mode: In this mode, the software package will analyze and construct a tree automatically. All you have to do is to type a fractional reduction α (initialized at .1), which should be between 0 and 1. The value of α will be used to see if there is a substantial reduction in $\hat{\sigma}$ after splitting.

(b) *Interactive* mode: This mode allows you to perform variable selection and simplify the PHD estimates. A dialog box is given to assist your decision. After that, highlight Continue item from the pull-down menu pHdrt, another dialog box appears which allows you to determine if a node is a terminal node or not. At this moment you can view the PHD plot, check the error reduction rate α, or the F-test for homogeneity of a node. The software will carry out the analysis, and continue until a satisfactory tree is found.

After constructing the tree, a new menu is displayed for retrieving information about the tree-structured regression analysis we just performed:

1. *Show Tree* item: Display the tree on the computer screen.

2. *Basic Information* item: Show the total numbers of nodes, intermediate nodes and terminal nodes.

3. *Node Information* item: Show a spin-plot of residuals against the first two components of PHD's and the pHdrt results. This option can also be executed by clicking any node on the tree displayed on the screen.

4. *Weighted Error* item: Provide $\hat{\sigma}$ for the whole tree, and for each terminal node.

5. *Modify Tree* item: Click it when you would like to rebuild a particular node.

References

Alexander, W. P. and Grimshaw, S. D. (1996), "Treed regression ," *J. Comp. Graphical Stat.*, **5**, 156-175.

Breiman, L. and Friedman, J. H. (1985), "Estimating optimal transformations for multiple regression and correlation," *J. Amer. Stat. Assoc.*, **80**, 580-597.

Breiman, L. and Friedman, J. H. (1988), Comment on "Tree-structured classification via generalized discrimination analysis," *J. Amer. Stat. Assoc.*, **83**, 725-727.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and regression trees*, Belmont, CA: Wadsworth.

Chaudhuri, P., Huang, M. C., Loh, W. Y. and Yao, R. (1994), "Piecewise-polynomial regression trees," *Statistica Sinica*, **4**, 143-167.

Cheng, C. S. and Li, K. C. (1995), "A study of the method of Principal Hessian Direction for analysis of data from designed experiments," *Statistica Sinica*, **5**, 617-639.

Cook, R. D. (1998), "Principal Hessian directions revisited," *J. Amer. Stat. Assoc.*, **93**, 84-94.

Duan, N and Li, K. C. (1991), "Slicing regression: A link-free regression method," *Ann. Statist.*, **19**, 505-530.

Filliben, J. J. and Li, K. C. (1997), "A systematic approach to the analysis of complex interaction patterns in two-level factorial designs," *Technometrics*, **39**, 286-297.

Friedman, J. H. (1991), "Multivariate adaptive regression splines," (with discussion), *Ann. Stat.*, **19**, 1-141.

Li, K. C. (1991), "Sliced inverse regression for dimension reduction," (with discussion), *J. Amer. Stat. Assoc.*, **86**, 316-342.

Li, K. C. (1992), "On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma," *J. Amer. Stat. Assoc.*, **87**, 1025-1039.

Loh, W. Y. and Vanichsetakul, N. (1988), "Tree-structured classification via generalized discrimination analysis," *J. Amer. Stat. Assoc.*, **83**, 715-728.

Lue, H. H. (1994), "Principal-Hessian-direction-based regression trees," unpublished Ph.D. Thesis, Department of Math., University of California, Los Angeles.

Sonquist, J. (1970), "Multivariate model building: The validation of a search strategy," Technical Report, Institute for Social Research, University of Michigan, Ann Arbor.

Stein, C. (1981), "Estimation of the mean of a multivariate normal distribution," *Ann. Stat.*, **9**, 1135-1151.

Tierney, L. (1990), *LISP-STAT: an object-oriented environment for statistical computing and dynamic graphics*, New York: John Wiley & Sons.

Chapter 17

Multivariate outcome data.

This chapter is a minor modification of Li, Aragon and Thomas-Agnan (1996).

So far the response variable Y is assumed to be one-dimensional. In this chapter, we shall show how to extend our approach to the analysis of multivariate outcome data. As a conjugate to the theory of SIR, a nonlinear theory of Hotelling's most predictable variates is developed. The interplay between both theories leads to a new approach for addressing multivariate issues on visualization, model parsimony, and the escalated complexity due to the increasing dimensionality. In the second half of this chapter, special attention is given to cases where outcomes are curves measured at fixed points.

17.1 Introduction.

Denote the outcome variable by $\mathbf{Y} = (Y_1, \cdots, Y_q)'$. We shall denote the regressor by the capital letter $\mathbf{X} = (X_1, \cdots, X_p)'$ in this Chapter. The multivariate version of the dimension reduction model (1.1) can still be used. The function g now takes values in R^q. At first sight, an extension of SIR to multivariate $\mathbf{Y}$ appears Straightforward and the theoretical development as described in Chapters 2 and 3 carries over formally. This is practically all right for $q = 2$ or 3. But as we shall soon see, several practical problems arise as we increase the dimension of $\mathbf{Y}$.

As used before, the term *variate* will be used to denote either any linear combination of the regressors or that of the outcome variables in appropriate contexts. For example, an *e.d.r. variate* means a variable $b'\mathbf{X}$ formed with an e.d.r. direction b. Similarly, we shall encounter *SIR-variates, M.P. variates* later on.

Prediction is a central goal in regression analysis. This is usually done componentwise by finding $E(Y_j|\mathbf{X})$ separately for $j = 1, \cdots, q$. This is not a multivariate issue. A genuine multivariate issue in prediction was raised by Hotelling(1935, 1936), concerning the most predictable variates: which linear combinations of $\mathbf{Y}$ are better predicted from $\mathbf{x}$ than others? Hotelling's theory uses linear predictors, leading to canonical correlation analysis. We shall revisit this issue in Section 3, but with one crucial difference: non-linear predictors are allowed. It turns out that the *most predictable variates* (*M.P. variates* hereafter) can be found from an eigenvalue decomposition which takes the same form as SIR except for the exchanged roles of **X** and **Y**. This result is based on an observation in Li(1991) which explains the descriptive nature of SIR without relying on the dimension reduction model (1.1).

In Section 4, we apply the nonlinear M.P. theory to the issues of data visualization and parsimonious modeling. A general discussion is given in Section 4.1 first. Then in Section 4.2, we focus on the case where the multivariate **Y** consists of q measurements of a curve at specified points, a situation often encountered in the longitudinal studies. Since different sample curves are likely to display different patterns, a central question here is how to explain the differences by the associated regressors. The nonlinear M.P. theory suggests a new analysis strategy for such data.

The discussion up to now assumes that the slicing step in the SIR (or M.P.) algorithm is easy to implement. This is fine if **Y** (or **X**) is low-dimensional. But if the dimensions of **X** and **Y** are both large, then we may not have enough points to make a meaningful simultaneous partition on the entire **Y** (or **X**)space. To discuss this most difficult situation, a mathematical formulation is developed in Section 5. Here we study the feasibility of slicing a low dimensional projection of **Y** (or **X**). But among different projections, some may be more informative than others. To sort out the relationship between the SIR variates (or M.P. variates) found by slicing various projections of **Y** (or **X**), anotion of partial ordering is used. Interesting duality between SIR variates and M.P. variates is observed.

In Section 6, we consider an alternating strategy. The idea is to begin with any initial projection of **Y**. Slicing is applied and a preliminary set of e.d.r. directions is found. Then they are in turn used to find the associated M.P. variates. Now replace the initial **Y** projection with these M.P. variates for slicing and iterate. For successfully carrying out this idea, a good initial projection is needed. It turns out that Hotelling's canonical variates serve this purpose well. In the population version, we show that the convergence takes place within no more than K steps, where K is the dimension of the e.d.r. space in (1.1). Examples are given to illustrate this alternating SIR procedure.

Section 7 concludes our discussion by summarizing the main findings. All technical proofs are given in Appendix.

17.2 A brief review of SIR.

Recall that the population version of SIR explores the p-dimensional inverse regression curve

$$\eta(\mathbf{Y}) = E(\mathbf{X}|\mathbf{Y})$$

It consists of a slicing step, which estimates $\eta(\mathbf{Y})$ by step functions, and an eigenvalue decomposition step:

$$\begin{aligned} \Sigma_\eta &= cov(\eta(\mathbf{Y})) \\ \Sigma_\eta b_i &= \lambda_i \Sigma_{\mathbf{X}} b_i,\ i = 1, \cdots, p \qquad (2.1) \\ & \lambda_1 \geq \cdots \geq \lambda_p \end{aligned}$$

The second step acts as the principal component analysis (with respect to the metric determined by $\Sigma_{\mathbf{X}}$) in finding the main directions of the inverse regression curve. The first few eigenvectors are used to estimate the e.d.r. directions.

Formally, it does not make much difference whether $\mathbf{Y}$ is univariate or not in defining the inverse regression $\eta(\mathbf{Y})$. It is quite easy to see that the main justification of SIR, Theorem 2.1 still holds for the multivariate case.

However, the slicing step for estimating $\eta(\mathbf{Y})$ requires simultaneous partition on all coordinates of $\mathbf{Y}$. This can cause an implementation problem if the dimension of $\mathbf{Y}$ is also large. We will come back to this issue later in Sections 5 and 6.

17.3 Most predictable variates.

Some variates of $\mathbf{Y}$ can be predicted from $\mathbf{X}$ better than others. The study of finding those with the best predictability is pioneered by Hotelling(1935). This is a genuine multivariate issue. Under the normality assumption, the best prediction rules are the linear ones. This reduces the problem to maximizing the correlation:

$$\max_{\theta,b} \rho(\theta'\mathbf{Y}, b'\mathbf{X}) \tag{3.1}$$

The solution is the first pair of canonical variates, $(\theta'_{C1}\mathbf{Y}, b'_{C1}\mathbf{X})$. Subject to being uncorrelated to the predecessors, other pairs of canonical variates $(\theta'_{C2}\mathbf{Y}, b'_{C2}\mathbf{X}), \cdots$, are similarly defined.

A natural generalization of Hotelling's theory is to allow for nonlinear predictors. For any linear combination $\theta'\mathbf{Y}$ of $\mathbf{Y}$, given $\mathbf{X} = \mathbf{x}$, the best nonlinear prediction under the squared-error loss is $E(\theta'\mathbf{Y}|\mathbf{X} = \mathbf{x})$ and the associated prediction mean squared-error is equal to $var(\theta'\mathbf{Y}|\mathbf{X} = \mathbf{x})$. The *most predictable variate* (*M.P. variate*) $\theta'_1\mathbf{Y}$ is the one that minimizes the ratio

$$\frac{E[var(\theta'\mathbf{Y}|\mathbf{X})]}{var(\theta'\mathbf{Y})}$$

Due to the ANOVA identity

$$var(\theta'\mathbf{Y}) = var[E(\theta'\mathbf{Y}|\mathbf{X})] + E[var(\theta'\mathbf{Y}|\mathbf{X})] \tag{3.2}$$

we may equivalently find the first M.P. variate $\theta'_1\mathbf{Y}$ by maximizing the proportion of variance in $\theta'\mathbf{Y}$ explained by the variation in the regression $E(\theta'\mathbf{Y}|\mathbf{X})$:

$$\max_{\theta} \frac{var[E(\theta'\mathbf{Y}|\mathbf{X})]}{var(\theta'\mathbf{Y})} \tag{3.3}$$

Similarly we can find the second M.P. variate $\theta'_2\mathbf{Y}$ from variates uncorrelated to $\theta'_1\mathbf{Y}$. Other M.P. variates are defined accordingly.

Let

$$\begin{aligned} \zeta(\mathbf{X}) &= E(\mathbf{Y}|\mathbf{X}) \\ \Sigma_\zeta &= cov[\zeta(\mathbf{X})] \\ \Sigma_{\mathbf{Y}} &= cov(\mathbf{Y}) \end{aligned}$$

and write (3.3) as

$$\max_{\theta} \frac{\theta' \Sigma_{\zeta} \theta}{\theta' \Sigma_{\mathbf{Y}} \theta}$$

Thus the M.P. variates can be found by conducting the eigenvalue decomposition:

$$\begin{aligned} \Sigma_{\zeta} \theta_j &= \rho_j \Sigma_{\mathbf{Y}} \theta_j \\ \rho_1 &\geq \cdots \geq \rho_q \end{aligned} \tag{3.4}$$

Remarkably, (3.4) is indeed the same as (2.1) except for the exchanged roles of **X** and **Y**. Thus the same SIR program can be used for estimating the M.P. directions. This twin relationship between SIR variates and M.P. variates underlies the main development in this article.

Remark 3.1. There are other ways of generalizing canonical correlation. For example, we may consider

$$\min_{\theta, b, h(\cdot)} \frac{E(\theta' \mathbf{Y} - h(b' \mathbf{X}))^2}{var(\theta' \mathbf{Y})} \tag{3.5}$$

Another possibility is

$$\min_{\theta, b, h_1(\cdot), h_2(\cdot)} \frac{E(h_2(\theta' \mathbf{Y}) - h_1(b' \mathbf{X}))^2}{var\{h_2(\theta' \mathbf{Y})\}}$$

There are no closed-form solutions for these minimization problems, however.

17.4 Plots and models.

Graphical display is another issue where special considerations are required for multivariate outcomes. In univariate regression, the vertical axis is usually reserved for the outcome variable to plot against the regressors. Conceptually this is also the usual way to imagine a regression surface. But for multivariate outcomes, it is hard to add or imagine more than one vertical axis for **Y**. An alternative is to construct two sets of plots, one for **X**−variates and another for **Y**−variates. These plots interact with each other under *linking*, a popular feature in dynamic graphics. When browsing through various sections in the **X** plot, the corresponding portions of the data points in the **Y** plot are highlighted for attention.

Section 4.1. discusses the model parsimony aspect related to graphical display. This leads to a general strategy for model building. In Section 4.2, we consider the special case where the outcome variable **Y** is a curve evaluated at q fixed points. Such data are often found in longitudinal studies for biomedical, economical or industrial applications; see, for example, Pottoff and Roy(1964), Liang and Zeger(1986), Rao(1987), Kneip and Gasser(1992), Segal(1992), etc..

17.4.1 Model parsimony.

To motivate the idea, we begin with the univariate **X** -versus-bivariate case. Suppose that conditional on the univariate **X**, the bivariate outcome $\mathbf{Y} = (Y_1, Y_2)$ has a bivariate normal distribution with covariance matrix Σ_ϵ :

$$\begin{aligned} Y_1 &= \mu_1(\mathbf{X}) + \epsilon_1 \\ Y_2 &= \mu_2(\mathbf{X}) + \epsilon_2 \end{aligned} \tag{4.1}$$

Now browse through $\mathbf{X}$ and observe the change in the scatterplot of Y_1 and Y_2. The center of the browsed data points roughly falls on the trajectory of the mean response curve, $E(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = (\mu_1(\mathbf{x}), \mu_2(\mathbf{x}))'$ as $\mathbf{x}$ varies.

We are interested in the special case when the center moves closely along some straight line with the direction, say, $(c_1^*, c_2^*)'$. If this is the case, then $\mu_1(\mathbf{X})$ and $\mu_2(\mathbf{X})$ are colinear. The projection of $\mathbf{Y}$ on the perpendicular direction, the variate $c_1^* Y_2 - c_2^* Y_1$, does not depend on $\mathbf{X}$. It is redundant in modeling. Thus instead of two regression curves, only one $\mathbf{Y}$ variate is needed in regressing against X. Which variate to use? One suggestion may be the variate $c_1^* Y_1 + c_2^* Y_2$, which represents the motion of $\mathbf{Y}$ as seen from the line of slice centers. But a disadvantage of using this variate is that the error part $c_1^*\epsilon_1 + c_2^*\epsilon_2$ would be correlated with the redundant variate $c_1^* Y_2 - c_2^* Y_1$. This means that we need an extra parameter to describe the correlation between the two.

The above discussion indicates the importance of locating the smallest affine space that contains the surface generated by $\eta(\mathbf{x}) = E(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ as $\mathbf{x}$ varies. In general, when the error in model (1.1) is additive:

$$\mathbf{Y} = g(\beta_1'\mathbf{X}, \cdots, \beta_K'\mathbf{X}) + \epsilon \tag{4.3}$$

parsimony can be achieved by first identifying the redundant variates. Then we may build a model using only the variates whose error components are uncorrelated to the redundant ones. This leads to the following expression:

$$\begin{aligned} M_1'\mathbf{Y} &= g_1(\beta_1'\mathbf{X}, \cdots, \beta_K'\mathbf{X}) + \epsilon_1 \\ M_2'\mathbf{Y} &= \mu_2 + \epsilon_2 \end{aligned} \tag{4.4}$$

where (M_1, M_2) form a q by q invertible matrix and $cov(\epsilon_1, \epsilon_2) = 0$. Based on the nonlinear M.P. theory of Section 3, such a decomposition is easy to find.

Lemma 4.1.*The eigenvectors from the eigenvalue decomposition (3.4) can be used to form $M_1'\mathbf{Y}$ and $M_2'\mathbf{Y}$ in (4.4):*

$$M_1'\mathbf{Y} = (\theta_1'\mathbf{Y}, \cdots, \theta_l'Y)', M_2'\mathbf{Y} = (\theta_{l+1}'\mathbf{Y}, \cdots, \theta_q'\mathbf{Y})'$$

where l is the number of non-zero eigenvalues. With this decomposition, the covariance matrix of the error component $(\epsilon_1', \epsilon_2')'$ in (4.4) becomes diagonal.

The verification of this lemma is straightforward; see Appendix A.1. It suggests the following three major steps for analyzing multivariate outcome data: (1). Reduce the dimension of $\mathbf{X}$ by finding the e.d.r. space. (2). Use the estimated e.d.r. variates to find the M.P. variates. (3). Apply any low-dimensional regression/smoothing technique for each M.P. variate against the e.d.r. variates.

Note that the first step is often necessary for high dimensional **X** because it would be difficult to partition the **X** space directly for finding M.P. variates. The following simulation study illustrates the recommended procedure.

Example 4.1. We generate 100 i.i.d. cases of $\mathbf{Y} = (Y_1, Y_2)'$, $\mathbf{X} = (X_1, \cdots, X_5)'$ from model (4.1) with

$$\begin{aligned} \mu_1(\mathbf{X}) &= sin(X_1 + X_2) \\ \mu_2(\mathbf{X}) &= 2\mu_1(\mathbf{X}) \\ cov(\epsilon) &= \left\{ \begin{array}{cc} 2 & -1 \\ -1 & 1 \end{array} \right\} \end{aligned}$$

The scatterplot of Y_1 and Y_2, Figure 4.1(a), indicates a negative association. Note that without proper guidance, it is hard to browse the 5 dimensional **X** space for revealing the line of centers (with a positive slope) hidden in the scatterplot of Y_1 against Y_2.

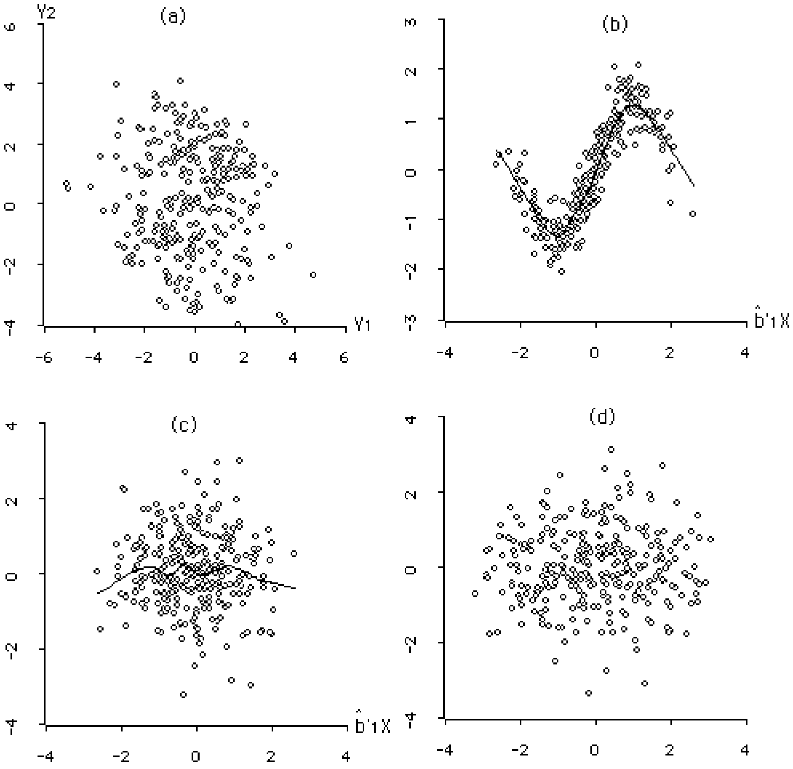


Figure 4.1: (a) is the scatterplot of Y_1 and Y_2. (b) shows a sine curve and (c) shows a flat curve for the first and the second M.P. variate respectively. (d) displays a normal uncorrelated pattern between the residuals.

We first conduct a SIR analysis to reduce the dimension of **X**. This is carried out by slicing on Y_1 and Y_2 with 6 slices in each variate, yielding a total of 36 slices. See Remark

4.1 for implementation details about the multiple slicing method and Remark 4.2 for modification of eigenvalues. From the SIR output, the eigenvalues (modified) are found to be $(0.53, 0.06, 0, 0, 0)$. The leading direction

$$\hat{b}_1 = (-0.66, -0.72, -0.03, 0.00, -0.02)' \approx (-.66, -.72, 0, 0, 0)'$$

stands out clearly.

The SIR variate $\hat{b}_1'\mathbf{X}$ is used to find the M.P. variates. This is implemented by applying the same SIR program with $\hat{b}_1'\mathbf{X}$ as the output variable and (Y_1, Y_2) as the regressor. The M.P. directions are found to be

$$\hat{\theta}_1 = (-0.32, -0.50)'; \hat{\theta}_2 = (-0.56, 0.24)'$$

with eigenvalues $(0.82, 0.01)$. The small second eigenvalue suggests that we may ignore the second M.P. direction. This leads to a tentative model:

$$\begin{aligned} -.32Y_1 - .50Y_2 &= g_1(-.66X_1 - .72X_2) + \epsilon_1 \\ -.56Y_1 + .24Y_2 &= \mu_2 + \epsilon_2 \end{aligned} \tag{4.5}$$

We still need to estimate $g_1(\cdot)$. This is a one-dimensional curve smoothing problem and there are many procedures available. For simplicity, we only apply the LOWESS provided in XLISP.STAT. A sine pattern is clear from Figure 4.1(b) which gives the scatterplot of $\hat{\theta}_1'Y$ against $\hat{b}_1'\mathbf{X}$ together with the LOWESS curve. For comparison, the same smoothing procedure is carried out for the redundant variate $\hat{\theta}_2'\mathbf{Y}$ against $\hat{b}_1'\mathbf{X}$; see Figure 4.1(c). As expected, the LOWESS curve is flat , suggesting a constant function as the fitted curve. The scatterplot of residuals, $\hat{\epsilon}_2$ versus $\hat{\epsilon}_1$, is given in Figure 4.1(d). No dependence or nonlinearity pattern is found.

After estimating $g_1(\cdot)$, we can now invert (4.5) back to the original $\mathbf{Y}$ variates:

$$\begin{aligned} Y_1 &= -.67\hat{g}_1(-.66x_1 - .72x_2) + \epsilon_1^* \\ Y_2 &= -1.55\hat{g}_1(-.66x_1 - .72x_2) + \epsilon_2^* \end{aligned}$$

The mean part now forms a line with slope $1.55/.67 \approx 2.3$, very close to the theoretical value 2.

Remark 4.1. The program for multiple slicing in all examples is based on the ordering of the variables to be sliced. The partition is first made according to the first variable, yielding h_1 slices. Then each slice is further partitioned into h_2 slices according to the second variable, and so on. This produces a total of $h_1 \times h_2 \times \cdots$ slices.

Remark 4.2. When the number of cases per slice is small, the sample slice means have sizable variances, which cause the inflation of the output eigenvalues. Li(1993) proposes a simple modification based on the second moment consideration of SIR as in the Rejoinder of Li(1991). The ith modified eigenvalue is $max\{(n\hat{\lambda}_i - H)/(n - H), 0\}$, where H is the total number of slices.

17.4.2 Curves.

Suppose the output variable $\mathbf{Y} = (Y_1, \cdots, Y_q)'$ is a time-dependent characteristic $Y(t)$ measured at $t = t_1 < t_2 < \cdots < t_q$, yielding $Y_j = Y(t_j)$, $j = 1 \cdots, q$. A natural way to display such data is the time plot which connects points (t_j, Y_j) by line segments. In the regression setting, each sample curve is further associated with one realization of **X**. Different **X** values may lead to different curve patterns.

Parametrization is one way to sort out the differences between these individual curves. Let $\mu(t) = EY(t)$ be the curve averaged over the entire population. Typically, a set of basis functions $f_1(t), \cdots, f_L(t)$ is specified first. Then we may express the difference between each individual curve and the mean curve as a linear combination of the basis functions. With additive errors, this suggests the following model to begin with:

$$Y(t) = \mu(t) + c_1(\mathbf{x}) f_1(t) + \cdots + c_L(\mathbf{x}) f_L(t) + \epsilon(t) \tag{4.6}$$

where the coefficients $c_l(\mathbf{x})$, $l = 1, \cdots, L$, depend on the value **x** of the regressor **X**.

However, unlike the traditional parametric analysis we introduce nonparametric components to this formulation by allowing both the coefficient functions $c_l(\cdot)$ and the basis functions $f_l(t)$ to be determined from the data. From our viewpoint, the role of basis functions is to help summarize the observed curve patterns parsimoniously. If this strategy is to work well, the total number L of the basis functions must be as small as possible. This requires a good choice of the basis functions. Instead of specifying the basis functions $f_l(\cdot)$ prior to the analysis, it is best to turn to the data themselves for suggestion.

The following lemma shows that a parsimonious set of basis functions can be found by applying the nonlinear M.P. theory.

Lemma 4.2.*Suppose that the basis functions in (4.6) are linearly independent and the covariance matrix for $(c_1(\mathbf{X}), \cdots, c_L(\mathbf{X}))'$ is of full rank. Then the number of nonzero eigenvalues in (3.4) is equal to L. Moreover, the vectors*

$$\phi_l = \Sigma_Y \theta_l, l = 1, \cdots, L$$

span the same space as the one generated from the L basis vectors $(f_l(t_1), \cdots, f_l(t_q))'$, $l = 1, \cdots, L$.

The proof of this lemma is given in Appendix A.2. Now suppose n individual curves $Y_1(t), \cdots, Y_n(t)$, $t = t_1, \cdots, t_q$, with covariate $\mathbf{X}_1, \cdots, \mathbf{X}_n$, are available. The key steps in our curve analysis are given in the following:

(1). Ignore the time factor for the moment and treat the data as in a standard multivariate outcome setting. Apply the nonlinear M.P. techniques in Section 3 to find the M.P. directions.

(2). From each estimated M.P. direction $\hat{\theta}_l$, compute the basis vector $\hat{\phi}_l = \hat{\Sigma}_Y \hat{\theta}_l$. Put $\hat{\phi}_l = (\hat{\phi}_1(t_1), \cdots, \hat{\phi}_l(t_q))'$ and construct the time plot for the estimated basis function $\hat{\phi}_l(t)$.

(3). Compute the mean curve $\hat{\mu}(t_j) = n^{-1} \sum_i Y_i(t_j)$.

(4). Use the basis functions obtained from (2) and fit a linear model separately to the difference between each individual curve and the mean curve: for $i = 1, \cdots, n$

$$Y_i(t_j) - \hat{\mu}(t_j) = \hat{c}_{1i}\hat{\phi}_1(t_j) + \cdots + \hat{c}_{Li}\hat{\phi}_L(t_j) + \epsilon_{ji}, \ j = 1 \cdots, q$$

(5). Find out how the estimated coefficients depend on $\mathbf{X}$. For each $l = 1, \cdots, L$, study the plot of coefficient $\hat{c}_{li}$ against $\mathbf{X}_i$ and apply a curve fitting technique to estimate the coefficient function $c_l(\cdot)$. Dimension reduction techniques may be applied to improve the estimate here if $\mathbf{X}$ is high dimensional.

(6). The final model takes the form: for $i = 1, \cdots, n$,

$$Y_i(t_j) = \hat{\mu}(t_j) + \sum_{l=1}^{L} \hat{c}_l(\mathbf{X}_i)\hat{\phi}_l(t_j) + \text{ residual }, \ j = 1 \cdots, q$$

Example 4.2. This is a simulation study for illustrating our curve analysis procedure. Consider (4.6), let $\mu(t) = 0$, and take two basis functions, $f_1(t) = sin(2\pi t)$, $f_2(t) = cos(2\pi t)$, with t ranging from 0 to 1 at the increment of .1. Suppose $\mathbf{X}$ follows a bivariate standard normal distribution and the coefficient functions are

$$c_1(\mathbf{X}) = (X_1 - X_2)/2, c_2(\mathbf{X}) = sin(X_1 + X_2)$$

Errors are independent normal random variables with standard deviations $SD(\epsilon(t_j)) = t_j$. A sample of $n = 100$ curves are generated together with their regressors. Some sample curves are shown in Figure 4.2(a). It is hard to see the common curve patterns by eyes. To find M.P. directions, we use a 5 by 5 partition on (X_1, X_2). The modified eigenvalues are .91, .86, .33, .18, 0.07, 0, $\cdots$. The first two values are much larger than others, indicating

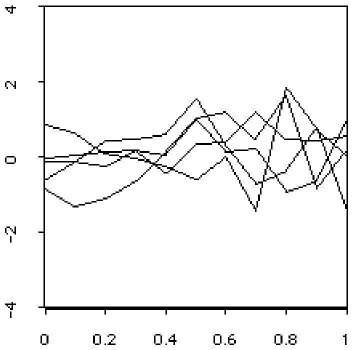


Figure 4.2: (a) some sample aurves for Example 4.2.

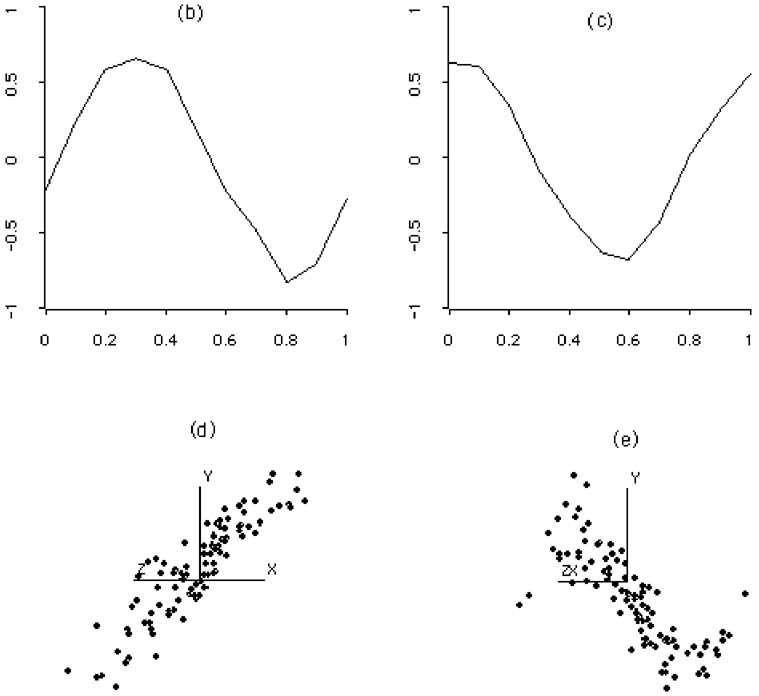


Figure 4.2: (b), (c) are the first and the second estimated basis functions, appearing like sine and cosine curves. (d) and (e) are the corresponding coefficients plotted against x1 and x2, rotated to reveal linear and sine surves.

two significant directions. From the first two M.P. directions, two basis functions, $\hat{\phi}_1, \hat{\phi}_2$ are then obtained and plotted in Figures 4.2(b),(c), revealing sine and cosine curves. Next, we compute the mean curve and find it close to zero as anticipated. Moving to Step (4), we use $\hat{\phi}_1, \hat{\phi}_2$ as the basis functions and find the coefficients $\hat{c}_{1i}, \hat{c}_{2i}, i = 1, \cdots, n$. Now for Step (5), we first plot the estimated coefficients $\hat{c}_{i1}$ against (X_{1i}, X_{2i}). By rotating the plot to the direction of the variate $X_1 - X_2$, we find a clear linear pattern, Figure 4.2 (e). This is consistent with the pattern of the true function $c_1(\mathbf{X})$. Similarly, the estimated regression coefficients for the second basis function are plotted in Figure 4.2(f), which reveals a pattern consistent with the shape of $c_2(\mathbf{X})$. If desired, we can further apply low-dimensional smoothing techniques for obtaining $\hat{c}_1(\mathbf{X})$ and $\hat{c}_2(\mathbf{X})$.

Our modeling process can be easily customized. For example, at Step (2), it is highly recommended to inspect the obtained basis curves carefully for detecting recognizable forms. We often find that the constant function is approximated well by some linear combination of these curves. In addition, at steps (1) and (5), various dimension reduction techniques can be applied. We shall come back to this later.

The next example illustrates some further practical concerns when dealing with small samples.

Example 4.3. Blood Pressure data. A small data set on blood pressures for 26 French girls with hypertension is available from an ongoing medical study by Drs. Barthe, Ph. and S. Cassadou who are interested in a medical issue related to blood pressure adjustment. A current medical practice is to adjust the raw blood pressure readings by the patient's height according to some published tables. To challenge this practice, our doctors want to know first if there is any evidence in the data suggesting that height (or other body measurement variables and age factors, not included here) is associated with the raw readings. Figure 4.3(a), gives the raw readings for each girl measured every hour from 10 a.m to 5 p.m.. Since the recommended adjustment is to add(or subtract) a fixed number (depending on the height) to (from) the reading regardless of the time of the day, we first compute the average reading for each girl and plot it against her height; Figure 4.3(b). No clear dependence relationship between the two is found. The correlation coefficient is about 0.1.

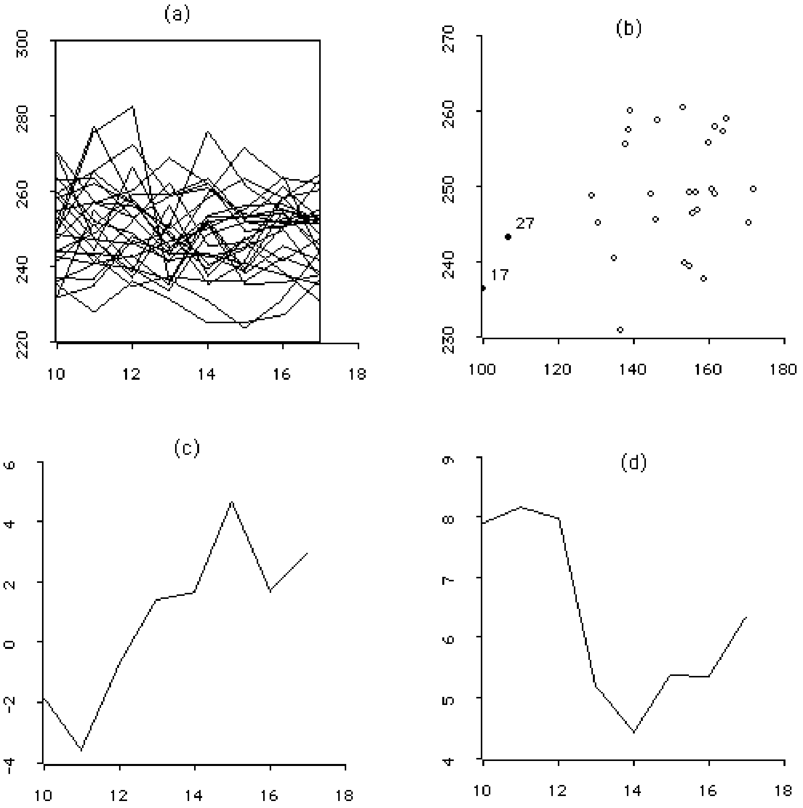


Figure 4.3: (a) raw reading curves for each girl. (b). Average against height. (c). (d) First two basis curves found.

Do these curves appear purely random? Or are there any special patterns among them that might be associated with height? To address such questions, we apply our method to find suitable basis functions first. Following Steps (1) and (2), five slices on **X** are used in

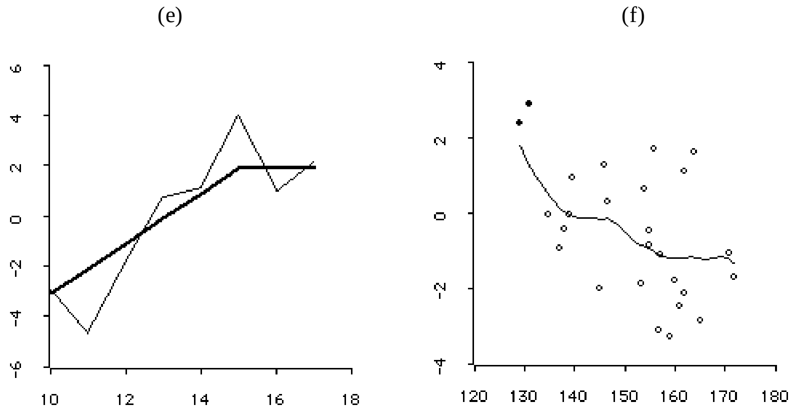


Figure 4.3: (e) simpified basis curve. (f). regression coefficient of the basis curve in (e) against height

finding the M.P. variates. The modified eigenvalues are .55, .18, 0, 0, 0, 0, 0, 0. The first two basis functions $\hat{\phi}_1(\cdot)$, $\hat{\phi}_2(\cdot)$ are plotted in Figures 4.3(c), (d).

Since our sample size is quite small, a qualitative statement about the basis functions is more relevant at this stage of analysis. We seek ways to simplify the estimated basis functions. The constant function turns out quite close to some linear combination of the two basis functions. The vector of ones has an angle of $cos^{-1}(.99)$ with the plane spanned by the vectors $\hat{\phi}_1$ and $\hat{\phi}_2$. This suggests the constant function $\tilde{\phi}_1(t) = 1$ as one of the basis functions. To find the other basis, we subtract the constant term from each individual curve to get

$$\tilde{\mathbf{Y}}_i = (\tilde{Y}_i(t_1), \cdots, \tilde{Y}_i(t_q))' = (Y_i(t_1) - \bar{Y}_i, \cdots, Y_i(t_q) - \bar{Y}_i)'$$

where $\bar{Y}_i$ is the average of $Y_i(t_j)$'s. Now apply the same basis searching method to $\tilde{Y}_i$'s as in Steps (1) and (2) again. The modified eigenvalues become .52, 0, 0, 0, 0. As expected, only the first one is significant. The obtained basis function, the shape of the thinner curve in Figure 4.3(e), appears almost the same as the curve found earlier in Figure 4.3(c). We further qualitatively approximate this function with a simple spline function,$\tilde{\phi}_2(t) = min\{t - 13, 2\}$, the truncated line shown in Figure 4.3(e). We shall use $\tilde{\phi}_1(\cdot)$ and $\tilde{\phi}_2(\cdot)$ are our basis functions to proceed the analysis.

For Step (3), we find that the mean curve is nearly a constant and decide to ignore it in the analysis. At step (4), we fit each individual curve $Y_i(t)$ with the two selected basis functions. Figure 4.3(f) shows the estimated regression coefficients for $\hat{\phi}_2(\cdot)$, plotted against heights. The descending trend, which starts from high values and then gradually flattens out, is interesting. This together with the pattern of the corresponding basis function suggests that for very short girls in this group, the blood pressures are likely to increase from 10 a.m. to 3 p.m..Taller girls on the other hand are not sensitive to this trend. We also carry out the same analysis with the unsmoothed basis function and find essentially the same pattern.

Remark 4.3. Model (4.6) can be extended in several ways. One of them is to allow for randomness in coefficients; namely, $c_i(\mathbf{x})$ is replaced by $c_i(\mathbf{x}, \epsilon_i)$. If the covariance matrix for $(E[c_1(\mathbf{X}, \epsilon_1)|\mathbf{X}], \cdots, E[c_L(\mathbf{X}, \epsilon_L)|\mathbf{X}])'$ does not degenerate, then our method still finds the basis functions. The plots constructed at Step (5) may be used to suggest forms for modeling randomness in coefficients. Another generalization of (4.6) is to allow Σ_ϵ to depend on $\mathbf{X}$. Again in principle, this does not affect our basis searching procedure although the additional complexity should not be taken lightly. Such issues deserve more attention in the future.

17.5 Duality between SIR variates and M.P. variates.

All examples we have shown up to now are the cases where either $\mathbf{X}$ or $\mathbf{Y}$ must have a small dimension so that slicing can be implemented reasonably. As argued in Li(1991), for low-dimensional $\mathbf{Y}$, SIR still works even with small slices. This is mainly because at the same time the number of slice means becomes larger so that the random fluctuations in estimating the conditional expectations can be filtered out at the principal component analysis step. Hsing and Carroll(1992) established the root n consistency for slices with only two cases each. But it does not solve the dimensionality problem. For moderate sample sizes, even the nearest neighbors in a high dimensional space are usually still too far apart to form meaningful pairs.

To rationalize the slicing step, one must reduce the dimensionality of $\mathbf{Y}$ first; c.f. Remark 5.1. Many multivariate techniques can be brought up for consideration. For example, we may conduct a principal component analysis on $\mathbf{Y}$ and then take the first few significant components for slicing. Other choices may include various low dimensional projections from different exploratory projection pursuit criteria and algorithms.

But slicing a lower dimensional projection of $\mathbf{Y}$ may not recover as many e.d.r. directions as slicing the entire $\mathbf{Y}$. For example, in the context of section 4.1, slicing the redundant $\mathbf{Y}$ variates is certainly useless. Can these directions be avoided as much as possible? This section sets up the needed mathematical formulation for discussing such issues.

To proceed, suppose slicing is based on a projection $M'\mathbf{Y}$ for some matrix M. Replace the term $\eta(\mathbf{Y})$ defined in Section 2 with $E(\mathbf{X}|M'\mathbf{Y})$ and apply the eigenvalue decomposition (2.1). Let $\kappa(M)$ be the number of nonzero eigenvalues and denote the eigenvectors by $b_i(M)$. Put the SIR directions into a p by $\kappa(M)$ matrix $S(M) = (b_1(M), \cdots, b_{\kappa(M)}(M))$.

Now reverse the roles of $\mathbf{X}$ and $\mathbf{Y}$ and consider a projection $N'\mathbf{X}$ for some matrix N. Replace the term $\zeta(\mathbf{X})$ in Section 3 with $E(\mathbf{Y}|N'\mathbf{X})$. Let $\tau(N)$ be the number of nonzero eigenvalues in (3.3) and put the first $\tau(N)$ eigenvectors $\theta_j(N)$ in a matrix $P(N) = (\theta_1(N), \cdots, \theta_{\tau(N)}(N))$.

Partial ordering notations will be used. When the column space of M_1 is contained in the column space of M_2, we may put $M_1 \preceq M_2$ (or $M_2 \succeq M_1$) and write $M_1'\mathbf{Y} \preceq M_2'\mathbf{Y}$ (or $M_2'\mathbf{Y} \succeq M_1'\mathbf{Y}$). The following lemma is obvious.

Lemma 5.1. *If $M_1'\mathbf{Y} \preceq M_2'\mathbf{Y}$, then $S(M_1) \preceq S(M_2)$. Similarly, if $N_1'\mathbf{X} \preceq N_2'\mathbf{X}$, then $P(N_1) \preceq P(N_2)$.*

Now suppose that we start with a projection $M'\mathbf{Y}$ for slicing and then use the obtained

SIR-variates in $S(M)'X$ for prediction. What can be said about the relationship between the M.P. variates in $P(S(M))'\mathbf{Y}$ and the original variates in $M'\mathbf{Y}$? We may anticipate the relationship

$$P(S(M))'\mathbf{Y} \succeq M'\mathbf{Y} \tag{5.1}$$

But this usually is not the case. There is no partial relationship between the two in general. Sometimes, we might even find the opposite relationship $\prec$ to hold. This creates a major problem for the alternating SIR strategy to be proposed in the next section.

The rest of this section digresses to the discussion on the duality property :

$$S(M) = N, P(N) = M \tag{5.2}$$

(5.2) implies $P(S(M)) = M$. Can we find pairs of (M, N) to satisfy (5.2)? The following Theorem shows the existence and uniqueness of the largest dual pair (M^*, N^*) in the sense that $M^* \succeq M$, $N^* \succeq N$ for any other dual pair (M, N).

Theorem 5.1. The maximum dual pair exists and is unique.

Remark 5.1. Dimension reduction needs not be restricted to linear projections on $\mathbf{Y}$. In the study of French city income distributions (Aragon, Li and Thomas-Agnan 1993), Theil index and other income inequality measures have been considered to transform $\mathbf{Y}$ nonlinearly.

Remark 5.2. Consider the minimization problem (3.5). Once a direction b is fixed, then the optimal solution for θ is simply $P(b)$. But conversely, given a θ direction, the optimal solution for b does not have a closed-form. In general, it can only be found by extensive searching as in projection pursuit regression. Nevertheless, $S(\theta)$ should be a good initial value for starting the search.

17.6 Alternating SIR.

The model parsimony consideration of Section 4.1 offers some guidance on what to (or not to) slice. For example, with the decomposition (4.4), slicing the redundant part $M_2'\mathbf{Y}$ is useless. On the other hand, by Lemma 4.1, slicing the M.P. variates in $M_1'\mathbf{Y}$ can recover as many e.d.r. directions as slicing the entire $\mathbf{Y}$. The problem is of course that the M.P. variates are still not found yet. Such considerations lead to an alternating strategy:

(0). Start with a $M_0'\mathbf{Y}$. Let $i = 0$.

(1). Use $M_i'\mathbf{Y}$ for slicing and apply SIR to find $S(M_i)$.

(2). Use the SIR variates $S(M_i)'\mathbf{X}$ to find the M.P. variates $P(S(M_i))'\mathbf{Y}$.

(3). Set $M_{i+1} = P(S(M_i))$ and return to Step (1).

(4). Iterate between (1) and (2).

Recall that step (2) also uses the same algorithm as SIR, We shall refer to this strategy as *alternating SIR.*

Now, will alternating SIR converge? Because of the complexity about partial ordering as we have mentioned earlier in Section 5, this cannot always be the case in general. However, as the following theorem shows, an initial projection which guarantees convergence can be found.

Theorem 6.1.*Suppose the alternating SIR starts from*

$$M_0 = (\theta_{C1}, \cdots, \theta_{C\tau(C)})$$

where θ_{Cj} 's are the canonical directions found from (3.1) and $\tau(C)$ is the number of non-zero canonical coefficients. Then the sequence $S(M_i)$, $i = 0, 1, \cdots$, converges in a finite number of steps. Furthermore, under the dimension reduction model (1.1) and condition (2.2), the convergence takes place in no more than K steps.

A proof of this theorem is given in the Appendix. The sample convergence properties are harder to study. This may raise some concerns if we were to specify the stopping rule by differences between consecutive iterations. However, encouraged by the finite step convergence result of Theorem 6.1, an interactive programming approach is to be recommended. Such a program will stop momentarily after each iteration to ask questions about how many significant variates and how many slices to use for partitioning. The users can answer the questions based on the displayed information about the current eigenvalues and eigenvectors. The users are also given two options to change the pace of iteration: (1) quit now; (2) continue for a specified number of times without pausing for answers. If the second option is taken, then the computer assumes that the same answers from the latest iteration will carry over. To implement this interactive strategy, an object-oriented package like Xlisp.stat (Tierney 1990) offers an ideal environment. The following two examples illustrate how this strategy works.

Example 6.1. Consider the curve outcome Example 4.2 again. The basis functions and error structures are kept the same as before. But we take a six dimensional regressor $\mathbf{X} = (X_1, \cdots, X_6)'$ and set the coefficient functions to be

$$c_1(\mathbf{X}) = X_1,\ c_2(\mathbf{X}) = (X_2 + 3)(cos(X_1) - 0.6)$$

Now generate $n = 200$ sample curves. The first five curves are shown in Figure 6.1(a). Again, no clear pattern is visible. Since the dimension of $\mathbf{X}$ is too big for simultaneous slicing, we want to apply the alternating SIR program to reduce the dimensionality before searching the basis functions.

By default the canonical correlation analysis is performed first and the output eigenvectors and *squared* eigenvalues are displayed. A choose-item-dialogue window soon pops out, asking how many canonical variates to be used as the initial projection of $\mathbf{Y}$. Judging from the *squared* canonical correlation coefficients, .98, .14, .07, .03, .02, .00, we decide to use only the first variate. After answering this question, another window appears, asking how many slices to be used. We input a number 10 to continue. The computer now performs the

SIR analysis and a summary is displayed on the screen. After that, the computer asks if we want to continue. After answering ”Yes”, we enter the iteration mode. Since the eigenvalues from the first SIR analysis are .95, .03, .00, ..., we use only one SIR variate for finding M.P. variates. With 10 slices, the computer then searches for the M.P. variates. From the output eigenvalues, .92, .67, 0.15, .10, .06, .03, 0, ..., the first two M.P. variates are found significant. Then we make a total of 30 slices (6 slices on the first M.P. variate and 5 slices on the second) to run SIR. This gives a list of eigenvalues : .97, .22, .035, 0, 0, 0, and the first two SIR directions are

$$(-1.00, 0.00, -0.00, 0.01, 0.01, -0.02), \text{ and, } (0.12, -0.95, -0.15, -0.09, 0.19, -0.09)$$

To continue the iteration, these two SIR variates are used to make 30 slices. We find again two significant M.P. variates. They are in turn used to make 30 slices for running SIR. The new eigenvalues become .98, .34, .07, 0, 0, 0. Compared to the last iteration, the SIR dimension does not increase and the new eigenvectors

$$(-1.00, -0.00, -0.012, 0.02, 0.02, -0.01) \text{ and } (0.10, -0.96, -0.05, 0.06, 0.05, 0.09)$$

are close to the previous ones. This is a good place to terminate the iteration. Out of curiosity, we also let the computer perform a couple more iterations. Small fluctuations in the output are noticeable. This is somewhat expected because slicing is not a continuous operation. Nevertheless, we find all answers remarkably close to the true e.d.r. space.

After reducing the dimensionality of **X** to just two, we now apply the basis searching algorithm as described in Section 4.2. The two obtained basis functions reveal clear sine and cosine patterns; Figure 6.1 (b), (c).

Example 6.2. In addition to the two basis functions as in Example 6.1, we add a third one, a constant function, $f_3(t) = 1$. The coefficient functions are

$$c_1(\mathbf{X}) = (X_2 + 3)(cos X_1 - 0.6),\ c_2(\mathbf{X}) = X_1(|X_2| - 0.8),\ c_3(\mathbf{X}) = X_1$$

Again we generate $n = 200$ curves and apply alternating SIR for reducing the dimensionality. Tables 6.1 gives the dimensions of the significant directions found during the iteration. Starting from the top row, the canonical correlation analysis finds one component significant. The first canonical **Y** variate is used as the initial projection for slicing. The second row from the table shows that the number of significant variate found by the SIR (with 10 slices) analysis is 1; see the number in the column of **X** variates. This SIR-variate is used to find M.P. variates. The next row reports that two significant M.P. variates are found; see the number in the column of **Y** variates. And so on. Overall, we see that the dimension of SIR directions increases from one to two , while the number of M.P. directions goes up to three. The number of slices used in each step is given in the parentheses following the method’s name. The eigenvalues are also recorded in the last column of Table 6.1. Table 6.2 gives the SIR directions found in each iteration.

Using the two estimated e.d.r. directions from the last iteration, we proceed with basis searching. The eigenvalues are found to be .92, .71, .53, .16, .11, .04, 0, ..., The first three

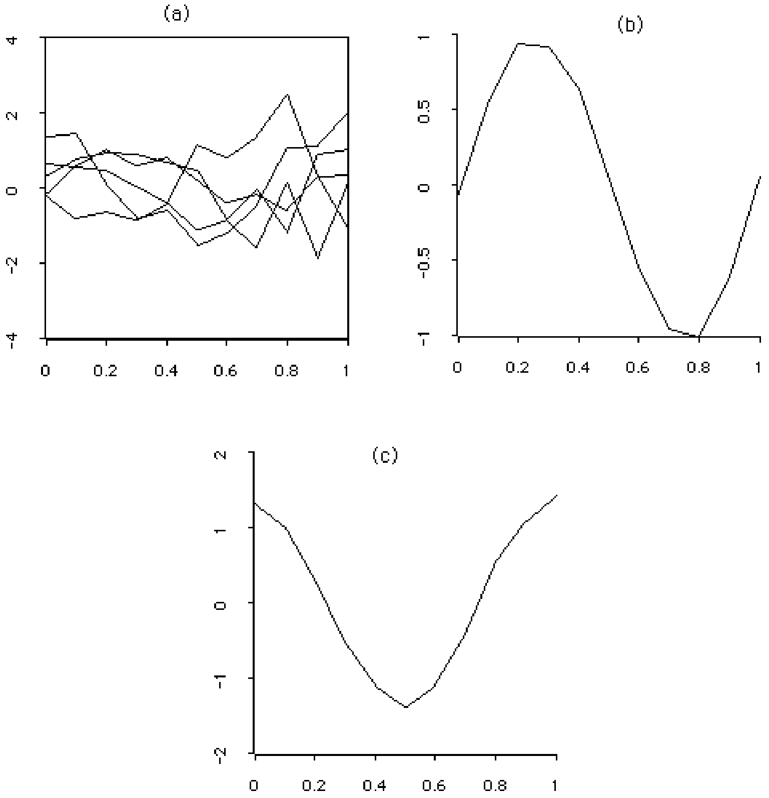


Figure 6.1: The first 5 curves from the sample in Example 6.1 are shown in (a). (b) and (c) are the results of basis searching after using alternating SIR to reduce the dimensionality of X.

significant basis functions are shown in Figures 6.2.(a)-(c). Figure 6.2 (d) shows that the constant function is in fact hidden in the linear combination:

$$1 \approx -.98\hat{\phi}_1(t) - .28\hat{\phi}_2(t) - .08\hat{\phi}_3(t).$$

Remark 6.1. Limitation of alternating SIR. Denote the limit of convergence in Theorem 6.1 by $(\tilde{M}, \tilde{N})$. Since this must be a dual pair, by Theorem 5.1, we see that $\tilde{M} \preceq M^*$ and $\tilde{N} \preceq N^*$. Thus alternating SIR may not recover as many e.d.r. directions as from slicing the entire **Y**. Systematic ways of exploring slicing directions beyond $\tilde{M}$ are not available yet. One possibility for extension is to consider the second-moment based methods like pHd, SAVE, SIR-II. Work along this direction is worth pursuing further.

17.7 Conclusion.

This paper concerns multivariate analysis on the relationship between two groups of variables without knowing the form of the regression function. We develop strategies for re-

Table 6.1: A dimension chart of alternating SIR for Example 6.2. *For canonical correlation analysis, the squared correlation coefficients are reported in the eigenvalue column.

Method	**X** variates	**Y** variates	eigenvalues
Canon. Corr.	1	1	(.99, .09, .06, .03, .02, .01)*
$SIR(10)$	1	-	(0.96, 0.07, 0.03, 0.03, 0.01, 0.01)
M.P.(10)	-	2	(0.96, 0.75, 0.09, 0.03, 0.00, $\cdots$)
SIR (6 × 6)	2	-	(0.97, 0.36, 0.07, 0, 0, 0)
M.P. (6 × 6)	-	3	(0.93, 0.76, 0.68, 0.10, 0.04, 0.01, 0, $\cdots$)
SIR(5 × 4 × 3)	2	-	(0.96, 0.21, 0.07, 0, 0, 0)

Table 6.2: SIR directions found from each iteration in Example 6.2.

Iteration	SIR directions
SIR(10)	(−1.00, 0.02, 0.02, −0.00, −0.01, 0.01)
SIR (6 × 6)	(−1.00, −0.00, −0.01, 0.01, −0.00, 0.01)
$\cdots$	(0.00, −0.96, −0.03, −0.16, 0.02, 0.01)
SIR (5 × 4 × 3)	(−1.0, −0.01, −0.03, 0.01, −0.01, 0.02)
$\cdots$	(0.04, −0.81, −0.06, −0.28, −0.26, 0.27)

vealing the underlying regression structures by systematically reducing the dimensionality in each group. Our ideas are stimulated mainly by the discovery of a twin relationship between sliced inverse regression and a proposed nonlinear generalization of Hotelling's most predictable variates. We have developed an interactive program for carrying out our alternating SIR analysis. Issues of visualization and parsimonious model building are treated simultaneously under the general framework of Li(1991).

One major application of our methodology is in analyzing data sets with curves as the outcome variable. There are at least two goals in this area: (1) fitting/predicting each individual curve; (2) explaining the curve patterns by the regressors associated with each curve. Our method offers a parsimonious and data-adaptive set of basis functions for accomplishing the dual tasks of curve fitting and pattern explanation.

There are surely many unresolved issues and some of them are mentioned in this article. Much research in this complex area of high dimensional data analysis still awaits our exploration.

APPENDIX

A.1 Proof of Lemma 4.1.

First, the eigenvalue decomposition of (3.4) implies that

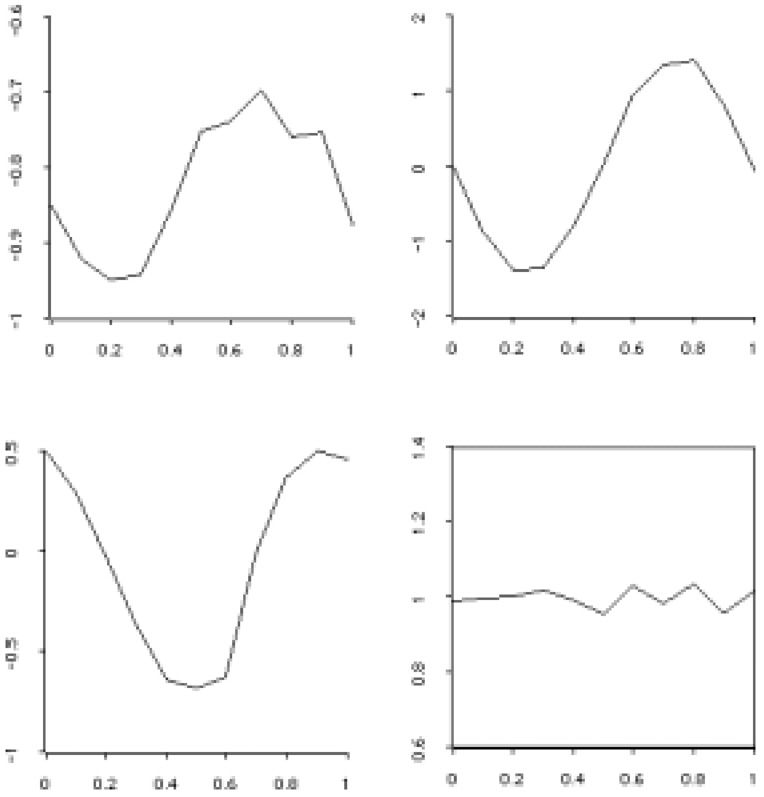


Figure 6.2: Output of Example 6.2. A nearly constant function (d) is hidden behind some linear combinations of the three obtained basis functions (a), (b), (c).

$$\Sigma_\zeta M_2 = 0 \tag{A.1}$$

$$M_1' \Sigma_Y M_2 = 0 \tag{A.2}$$

Now, (A.1) means that $cov[M_2'\zeta(X)] = 0$, or equivalently, $M_2'E(Y|X) = 0$. Thus $M_2'Y$ consists only of the error part $\epsilon_2 = M_2'\epsilon$. It remains to show that $cov(M_1'\epsilon, M_2'\epsilon) = 0$. But this follows from (3.2) (which implies $cov\,\epsilon = \Sigma_Y - \Sigma_\zeta$), and (A.1)-(A.2).

A.2. Proof of Lemma 4.2.

From (3.4), it follows that $\phi_l = \Sigma_{\mathbf{Y}}\theta_l = \rho_l^{-1}\Sigma_\zeta\theta_l$. Clearly we have

$$\Sigma_\zeta = cov(\sum_l c_l(\mathbf{X})\, f_l) = \sum_{l,l'} cov(c_l(\mathbf{X}), c_l'(\mathbf{X}))\, f_l\, f_l'$$

Thus ϕ_l can be expressed as a linear combination of f_l's. The rank of Σ_ζ must be L because of linear independence and the non-degeneracy of $cov(c_1(\mathbf{X}), \cdots, c_L(\mathbf{X})$. The proof is complete.

A.3 Proof of Theorem 5.1

Construct the sequence $N_o = I($ the p by p identity matrix), $N_1 = S(P(N_o))$, $N_2 = S(P(N_1))$, $\cdots$. This sequence is non-increasing. To establish this property, we start from $N_1 \preceq N_o$, which is of course obvious. We apply Lemma 5.1 twice to get $P(N_1) \preceq P(N_o)$ first and then $S(P(N_1)) \preceq S(P(N_o))$. This shows $N_2 \preceq N_1$. We can apply the same argument recursively to obtain the monotonicity for the entire sequence. Since our spaces have finite dimensions, this sequence must converge. Denote the limit of this sequence by N^*. The pair $(M = P(N^*), N = N^*)$ must satisfy the duality condition (5.2). To show that this pair is maximum, suppose (M, N) is another dual pair. Starting from $N \preceq N_o$, we can again apply Lemma 5.1 twice to get $N = S(P(N)) \preceq S(P(N_o)) = N_1$. Then the same argument leads to $N \preceq S(P(N_1) = N_2$, etc.. Take the limit to obtain $N \preceq N^*$. Apply Lemma 5.1 once more and we get $M = S(N) \preceq S(N^*) = M^*$, as desired.

Remark. We may also find the maximum pair by starting from $M_o = I($ the q by q identity matrix) and construct the non-increasing sequence $M_1 = P(S(M_o)), M_2 = P(S(M_1)), \cdots$. Denote the limit by M^{**}. Since $(M^{**}, S(M^{**}))$ must also be a maximum dual pair, the uniqueness result implies that $(M^{**}, S(M^{**})) = (P(N^*), N^*)$. In fact, this sequence and the previously constructed one can be mixed together to get two monotone sequences: $M_o \succeq P(N_o) \succeq M_1 \succeq P(N_1) \succeq M_2 \succeq \cdots$ and $N_o \succeq S(M_o) \succeq N_1 \succeq S(M_1) \succeq N_2 \succeq \cdots$.

A.4. Proof of Theorem 6.1

We need only to show that $M_1 \succeq M_o$. By applying Lemma 5.1 repetitively as in the proof of Theorem 5.1, we can establish the partial ordering for the entire sequence.

Now denote $N_o = (b_{C1}, \cdots, b_{C\tau(C)})$ and recall that $M_1 = P(S(M_o))$. If we can show that

$$S(M_o) \succeq N_o \text{ and } P(N_o) \succeq M_o \tag{A.1}$$

then by Lemma 5.1, we have $P(S(M_o)) \succeq P(N_o) \succeq M_o$, establishing the partial ordering $M_1 \succeq M_o$.

To show (A.1), without loss of generality we may assume that $cov\,\mathbf{X}$ and $cov\,\mathbf{Y}$ are both identity matrices because of affine invariance. The canonical directions have the relationship

$$b_{Cl} = E(\mathbf{X}\mathbf{Y}'\theta_{Cl}) = E(\mathbf{Y}'\theta_{Cl} \cdot E(\mathbf{X}|M_o'\mathbf{Y}))$$

Since each column of $E(\mathbf{X}|M_o'\mathbf{Y})$ must be in the column space of $S(M_o)$, b_{Cl} also belongs to the column space of $S(M_o)$. Hence we have shown $S(M_o) \succeq N_o$. The second part of (A.1) can be shown in the same way after reserving the roles of $\mathbf{X}$ and $\mathbf{Y}$. This completes the proof of Theorem 6.1.

Remark. The intuitive reason behind (A.1) is because nonlinear predictors are more flexible than linear ones. Since the canonical $\mathbf{Y}$ variates $M_o'\mathbf{Y}$ are linearly predictable from $N_o'\mathbf{X}$, they must also be predictable from $N_o'\mathbf{X}$ nonlinearly. This shows $M_o'\mathbf{Y} \preceq P(N_o)'\mathbf{Y}$, establising the second part of (A.1). The first part of (A.1) follows the same argument with **X** and **Y** exchanged.

References

Aragon, Y., Li, K.C., and Thomas-Agnan, C.(1995), “Modeling income distributions using multivariate sliced inverse regression”, *Technical Report.*

Brieman, L., and Friedman, J.(1985), “Estimating optimal transformations for multiple regression and correlation,” (with discussion), *J. Amer. Stat. Assoc.* **80**, 580-619.

Brillinger, D. R. (1991). Discussion of “Sliced Inverse Regression”, *J. Amer. Statist. Assoc.* **86**, 333-333.

Carroll, R. J. and Li, K. C. (1992). “Measurement error regression with unknown link : dimension reduction and data visualization”, *J. Amer. Stat. Assoc.*, **87**, 1040-1050.

Carroll, R.J. and Li, K.C.(1993).“Binary regressors in dimension reduction models: a new look at treatment comparisons”, *Technical Report.*.

Chaudhuri, P., Huang, M.C., Loh, W.Y., and Yao, R. (1994), “Piecewise-polynomial regression trees”, *Statistica Sinica*, **4**, 143-167.

Chen, H. (1991). “Estimation of a projection-pursuit type regression model”, *Ann. Stat.*, **19** 142-157.

Cook, R. D. (1994). “On the interpretation of regression plots”, *J. Am. Statist. Assoc.*, **89** , 177-189.

Cook, R. D., and Nachtsheim, C. J. (1994). “Re-weighting to achieve elliptically contoured covariates in regression”, *J. American Stat. Assoc.* **89** 592-599.

Cook, R. D. and Weisberg, S. (1991). Discussion of “Sliced inverse regression”, *J. Amer. Statist. Assoc.*, **86**, 328-332.

Duan, N. and Li, K. C. (1991). “Slicing regression: a link-free regression method”,*Ann. Stat.*, **19**, 505-530.

Friedman, J., and Stuetzel, W. (1981), “Projection pursuit regression,” *J. Amer. Stat. Asscoc.*, **76**, 817-823.

Friedman, J. (1991), “Multivariate adaptive regression splines,” (with discussion), *Ann. Stat.*, **19**, 1-141.

Härdle, W., Hall, P. and Ichimura, H. (1993). “Optimal smoothing in single index models”,*Ann. Stat.*, **21** 157-178. Härdle, W. and Marron, J. S. (1990). “Semiparametric comparison of regression curves”, *Ann. Stat.*, **18**, 63-89.

Härdle, W. and Stoker, T. M. (1989). “Investigating smooth multiple regression by the method of average derivatives”, *J. Amer. Statist. Assoc.*, **84** 986-995.

Hall, P. (1989). “On projection pursuit regression”, *Ann. Stat.*, **17** 573-588.

Hall, P. and Li, K. C. (1993). “On almost linearity of low dimensional projections from high dimensional data”, *Ann. Stat.*, **21**, 867-889.

Hotelling, H. (1935). “The most predictable criterion”, *I. Educ. Psych.,* **26**, 139-142.

Hotelling, H. (1936). “Relationship between two sets of variates”, *Biometrika*, **28**, 321-377.

Hsing, T. and Carroll, R. J. (1992). “Asymptotic properties of sliced inverse regression”, *Ann. Stat.*, **20**, 1040-1061.

Kneip, A. and Gasser, T. (1992). “Statistical tools to analyze data representing a sample of curves”, *Ann. Stat.*, **20**, 1266-1305.

Li, K. C. (1990). Data-visualization with SIR : a transformation-based projection pursuit method. *Technical Report.*

Li, K. C. (1991). “Sliced inverse regression for dimension reduction”, with discussions. *J. Amer. Statist. Assoc.*, **86**, 316-342.

Li, K. C. (1992a). “Uncertainty analysis for mathematical models with SIR”, *In Probability and Statistics,* 138-162, edited by Jiang Ze-Pei, Yan Shi-Jian, Cheng Ping, and Wu Rong, World Scientific, Singapore.

Li, K. C. (1992b). “On principal Hessian directions for data visualization and dimension reduction: another application of Stein’s lemma”, *J. Amer. Stat. Assoc.*, **87**, 1025-1039.

Li, K.C.(1993). Quasi-helices in high dimensional regression. *Technical report, UCLA.*.

Liang, and Zeger(1986).“Longitudinal data analysis using generalized linear models”, *Biometrika,* 73, 13-22.

Pottoff, R.F. and S.F. Roy (1964). “A generalized multivariate analysis of variance model useful especially for growth curve problems”, *Biometrika,* **51**, 313-326.

Samarov, A. M. (1993). “Exploring regression structure using nonparametric functional estimation,” *J. Amer. Stat. Assoc.*, **88** 836-847.

Rao, C.R. (1987). “Prediction of future observations in growth curve models,” *Statistical Science,* 2, 434-471.

Schott, J.R. (1994), “Determining the dimensionality in sliced inverse regression”, *J. Ameri. Stat. Assoc.*, **89**, 141-148.

Segal, M. R. (1992) “Tree-structured methods for longitudinal data”, *J. Amer. Stat. Assoc.* , 87, 407-418.

Tierney, L. (1990). *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics,* New York: John Wiley & Sons.

References for PART I

Aragon Y, and Saracco J (1997). "Sliced inversed regression (SIR): An appraisal of small sample alternatives to slicing", *COMPUTATION STAT* ,**12**, 109-130.

Bickel, P. J., and Doksum, K. A. (1981), "An Analysis of Transformations Revisited," *J. Amer. Stat. Assoc.*, **76**, 296-311.

Bickel, P. J. Klassen, C.A.J., Ritov, Y. and Wellner, J.A.(1992), *Efficient and adaptive estimation for semiparametric models.* Baltimore : Johns Hopkins University Press.

Box, G. E., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, Ser. B,* **26**, 211-252.

Box, G.E.P. and Draper, N.R. (1986). *Empirical model building and response surfaces.* Wiley, New York.

Box, G.E.P., Hunter, W.G. and Hunter, J.S.(1978). *Statistics for experimenters.* Wiley, New York.

Breiman, L., and Friedman, J.(1985). "Estimating optimal transformations for multiple regression and correlation," *J. Amer. Statist. Assoc.* **80**, 580-597.

Breiman,L., Friedman, J., Olshen, R., and Stone, C.(1984).*Classification and regression trees.* Wadsworth.

Brillinger, D. R. (1977). The identification of a particular nonlinear time series system. *Biometrika*, 64, 509-515.

Brillinger, D. R. (1983). A generalized linear model with Gaussian regressor variables. In *A Festschrift for Erick L. Lehmann*, pp. 97-114, Wadsworth.

Brillinger, D. R. (1991), Comments on "Sliced Inverse Regression for Dimension Reduction", by K. C. Li, *J. Amer. Stat. Assoc.*, **86**, 333-333.

Carroll, R. J. (1989). Covariance analysis in generalized linear measurement error models. *Statistics in Medicine*, 8, 1075-1093.

Carroll, R. J., and Li, K. C. (1992), "Measurement Error Regression with Unknown Link: Dimension Reduction and Data Visualization," *J. Amer. Stat. Assoc.*, **87**, 1040-1050.

Carroll R.J, Li. K.C. (1995) "Binary regressors in dimension reduction models - a new look at treatment comparisons," *Statistica Sinica* 5: 667-688.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman & Hall, London.

Chen, C.H. and Li, K. C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica.* **8**, 289-316.

Cheng, C. S. and Li, K. C. (1995) A study of the method of principal Hessian direction for analysis of data from designed experiments, *Statistica Sinica*, **5**, 617-640.

Cleveland, W.S. and MacGill, M.E.(1988). *Dynamic Graphics for Statistics.* Wadsworth

& Brooks/Cole.

Cook, R. D. (1994), "On the Interpretation of Regression Plots," *J. Amer. Stat. Assoc.*, **89**, 177-189.

Cook, R. D. (1998a), "Principal Hessian directions revisited," *J. Amer. Stat. Assoc.*, **93**, 84-94.

Cook, R.D. (1998b). *Regression Graphics.* Wiley, New York.

Cook, R.D., Bura, E. (1997) "Testing the adequacy of regression functions". *BIOMETRIKA* , **84**, 949-956.

Cook, R. D., Hawkins, D. M. and Weisberg, S. (1992), "Comparison of Model Misspecification Diagnostics Using Residuals from Least Mean of Squares and Least Median of Squares Fits," *J. Amer. Stat. Assoc.*, **87**, 419-424.

Cook, R. D., and Weisberg, S. (1991), Comments on "Sliced Inverse Regression for Dimension Reduction", by K. C. Li. *J. Amer. Stat. Assoc.*, **86**, 328-333.

Cook, R. D., and Weisberg, S. (1994). *An introduction to regression graphics.* John Wiley , New York.

Cook, R. D., and Wetzel, N. (1994), "Exploring Regression Structure with Graphics," (with discussion), *Test*, **2**, 33-100.

Cook, R. D., and Nachtsheim, J. C. (1994), "Reweighting to Achieve Elliptically Contoured Covariates in Regression," *J. Amer. Stat. Assoc.*, **89**, 592-599.

Cox, D.R., and Snell (1981). *Applied Statistics: Principles and Examples.* New York: Chapman & Hall.

Dabrowska, D. M. (1987), "Non-parametric Regression with Censored Survival Time Data," *Scandinavian J. Statist.*, **14**, 181- 197.

Dabrowska, D. M. (1992), "Variable Bandwidth Conditional Kaplan-Meier Estimate," *Scandinavian J. Stat.*, **19**, 351- 361.

Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Annals of Statistics*, 12, 793-815.

Doksum, K. A. (1987), "An Extension of Partial Likelihood Methods for Proportional Hazard Models to General Transformation Models," *Ann. Stat.*, **15**, 325- 345.

Doksum, K. A., and Gasko, M. (1990), "On a Correspondence between Models in Binary Regression Analysis and in Survival Analysis," *Internat'l Stat. Rev.*, **58**, 243- 252.

Duan, N. and Li, K. C. (1991), "Slicing Regression : a Link-Free Regression Method," *Ann. Stat.*, **19**, 505-530.

Fan, J., and Gijbels, I. (1995), "Censored Regression: Local Linear Approximations and Their Applications," To appear in *J. Amer. Stat. Assoc.*.

Ferre, L(1998). "Determining the dimension in sliced inverse regression and related methods." *J. Amer. Stat. Assoc.* 93, 132-140.

Filliben, J. and Li, K.C. (1997), “A systematic approach to the analysis of complex interaction patterns in 2-level factorial Designs”, *Technometrics*, . **39**, 286-297.

Fiskerkeller, M.A.,Friedman, J.H., and Tukey, J.W. (1974) ”PRIM-9:An interactive multidimensional data display and analysis system. SLAC-PUB-1408. Stanford. CA: Stanford Linear Accelerator Center.

Fleming, T. R., and Harrington, D. P. (1991), *Counting Processes and Survival Analysis*, New York: John Wiley & Sons.

Friedman, J., and Stuetzle, W. (1981), “Projection Pursuit Regression,” *J. Amer. Stat. Assoc.*, **76**, 817-823.

Friedman, J. (1991), “Multivariate Adaptive Regression Splines,” (with discussion), *Ann. Stat.*, **19**, 1-141.

Fuller, W. A. (1987). *Measurement Error Models.* Wiley, New York.

Geisser, S.(1975). łThe predictive sample reuse method with applications.” *J. Amer. Statist. Assoc.* **70** 320-328.

Gifi, A. (1991), *Nonlinear Multivariate Analysis*, Chichester: John Wiley & Sons.

Golub, G., Heath, M., and Wahba, G.(1979). łGeneralized cross-validation as a method of choosing a good ridge parameter. *Technometrics* **21** 215-223.

Hall, P. (1989). On projection pursuit regression. *Annals of Statistics*, 17, 573-588.

Hall, P., and Li, K. C. (1993), “On Almost Linearity of Low Dimensional Projection from High Dimensional Data,” *Ann. Stat.*, **21**, 867-889.

Härdle, W., and Stoker, T. (1989), “Investigating Smooth Multiple Regression by the Method of Average Derivatives,” *J. Amer. Stat. Assoc.*, **84**, 986-995.

Harrison, D., and Rubinfeld, D. L. (1978), “Hedonic Housing Prices and the Demand for Clean Air,” *Journal of Environmental Economics Management*, **5**, 81-102.

Hinkley, D. V., and Runger, G. (1984), “The Analysis of Transformed Data,” (with discussion), *J. Amer. Stat. Assoc.*, **79**, 302-320.

Hooper, J.(1959). Simultaneous equations and canonical correlation theory, *Econometrica.* **27**, 245-256.

Hsing, T and Carroll, R. J. (1992), “An Asymptotic Theory for Sliced Inverse Regression,” *Ann. Stat.*, **20**, 1040-1061.

Hsing, T. (1999) “Nearest neighbor inverse regression.”, *Ann. Stat.* **27**, 697-731

Huber, P. (1985), “Projection Pursuit,” (with discussion), *Ann. Stat.*, **13**, 435-526.

Huber, P.(1987). Experiences with three-dimensional scatterplots. *J. Amer. Statist. Assoc.* **82**, 448-454.

Kato, T. (1976). *Perturbation Theory for Linear Operators* (2nd ed.), Berlin: Springer-Verlag.

Koyak, R. (1987), “On Measuring Internal Dependence in a Set of Random Variables,” *Ann. Stat.*, **15**, 1215-1228.

Lenth, R.S. (1989). łQuick and easy analysis of unreplicated factorials” *Technometrics, 31 p469-473.*

Li, K. C. (1991), “Sliced Inverse Regression for Dimension Reduction,” (with discussion), *J. Amer. Stat. Assoc.*, **86**, 316-342.

Li, K. C. (1992a), “Uncertainty Analysis for Mathematical Models with SIR”, in *Probability and Statistics*, eds. Z. P. Jiang, S. H. Yan, P. Cheng, and R. Wu, Singapore: World Scientific, pp. 138-162.

Li, K. C. (1992b), “On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein’s Lemma,” *J. Amer. Stat. Assoc.*, **87**, 1025-1039.

Li, K.C. (1993), Discussion of ”Exploring regression structure with graphics”, *TEST*, **2**, 33-100.

Li, K.C. “Sliced inverse regression” (1995). To appear in *Encyclopedia in Statistical Sciences.*

Li, K. C. (1997), “Nonlinear confounding in High Dimensional Regression,” *Ann. Stat.* **25**, 577-612.

Li, K.C.(1998), “Statistical Inference in high dimensional data analysis - in discussion of *Principal Hessian directions revisted* by Cook.” *Journal American Statistical Association.* 93, 94-97.

Li, K.C., Aragon, Y, and Thomos-Agan, C.(1996). “Analysis of multivariate outcome data: SIR and a nonlinear theory of Hotelling’s most predictable variates”.

Li, K. C., and Duan, N. (1989), “Regression Analysis under Link Violation,” *Ann. Stat.*, **17**, 1009-1052.

Li, K.C. and Chen, C.H. (1999) ”A three-way subclassification approach to multiple-class discriminant analysis.”, Technical Report.

Li, K.C. and Lue, H.H. (1998). ”Tree-structure regression via principal Hessian directions.”, Journal of American Statistical Association, vol. , (1998). Submitted

Li, K.C. , Wang, J.L., and Chen, C.H.(1999) ”Dimension reduction for censored regression data”, Annals of Statistics, **27**, 1-21.

Loh, W.Y., and N. Vanichsetakul (1988). łTree-structured classification via generalized discriminant analysis”, *J. Amer. Stat. Assoc.*, 83, 715- 728.

Mallows, C. L. (1973). łSome comments on C_p”, *Technometrics* **15** 661-675. Mallows, C.L. (1986). “Augmented partial residuals”. *Technometrics,* **28,** 313-319.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, (2nd. ed.), London: Chapmn and Hall.

Mitchell, T.J. and J.J. Beauchamp (1988). łBayesian variable selection in linear regression, with discussions" *J. Amer. Stat. Assoc.* 1023-1036. Mosteller, F., and Tukey, J. W. (1977). *Data Analysis and Regression.* Reading, MA: Addison-Wesley.

Nelder, J.A., and Wedderburn, R.W.M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A.* **135**, 370-384.

Rice, John A.(1988) *Mathematical Statistics and Data Analysis.* Wadsworth & Brooks/Cole, Belmont, California.

Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis.* Springer, New York

Rosner, B., Willett, W. C. and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8, 1051-1070.

Schott, J. R. (1994), "Determining the Dimensionality in Sliced Inverse Regression," *J. Amer. Stat. Assoc.*, **89**, 141-148.

Souders T.M., and G. N. Stenbakken (1990) łA comprehensive approach for modeling and testing analog and mix-signal devices." *in Proc. . Int. Test Conf. 1990*, pp 169-176.

Spiegelman, C. H. (1986). Two pitfalls of using standard regression diagnostics when both X and Y have measurement error. *The American Statistician*, 40, 245-248.

Stefanski, L. A. and Carroll, R. J. (1990). Score tests in generalized linear measurement error models. *Journal of the Royal Statistical Society, Series B*, 52, 345–360.

Stenbakken, G.N. and T. M. Souders (1987) łTest point selection and testability measures via QR factorization of linear models", *IEEE Trans. Instrum. Meas.,* Vol IM-36, No 2, Jun. 1987, pp 406-410.

Stenbakken, G.N., T. M. Souders, and G.W. Stewart (1989) łAmbiguity groups and testability", *IEEE Trans. Instrum. Meas.,* Vol IM-38, No 5, Oct. 1989, pp 941-947..

Stein, C. (1981), "Estimation of the mean of a multivariate normal distribution," *Ann. Stat.*, **9**, 1135-1151.

Stone, M. (1974). łCross-validatory choice and assessment of statistical predictions." *J. Royal Statist. Soc. Ser. B.* **36** 111-147.

Tierney, L. (1990), *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*, New York: John Wiley & Sons.

Tosteson, T. and Tsiatis, A. (1988). The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates. *Biometrika*, 75, 507-514.

Tosteson, T., Stefanski, L. A. and Schafer D.W. (1989). A measurement error model for binary and ordinal regression. *Statistics in Medicine*, 8, 1139-1147.

Velilla, S (1998). "Assessing the number of linear components in a general regression problem", *J. Amer. Stat. Assoc.* **93**, 1088-1098.

White, H (1989), “Some asymptotic results for learning in single hidden-layer feed-forward network models. JASA vol 84, 1003-1013

Zhu, L. X., and Fang, K. T. (1996), “Asymptotics for Kernel Estimate of Sliced Inverse Regression,” *Ann. Stat.* **24**, 1053-1068.

Zhu, L. X., and Ng. (1995), “Asymptotics of Sliced Inverse Regression,” *Statistica Sinica*, **5**, 727-736.