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 $\mathbf{x}$:

$$\bar{H}\Sigma_{\mathbf{x}} b_j = \lambda_j b_j, \qquad j = 1, \cdots, p \qquad (1.1)$$
$$|\lambda_1| \geq \cdots \geq |\lambda_p|$$

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

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

and

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

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

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

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

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 **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

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

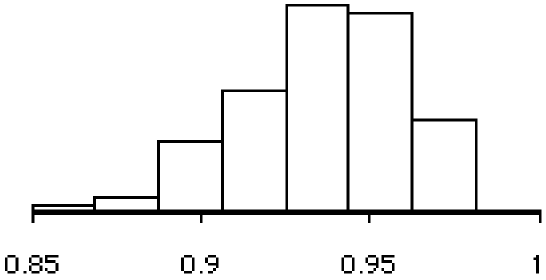

Figure 7.1: Histogram of $R^2(\hat{B})$ for 100 runs.

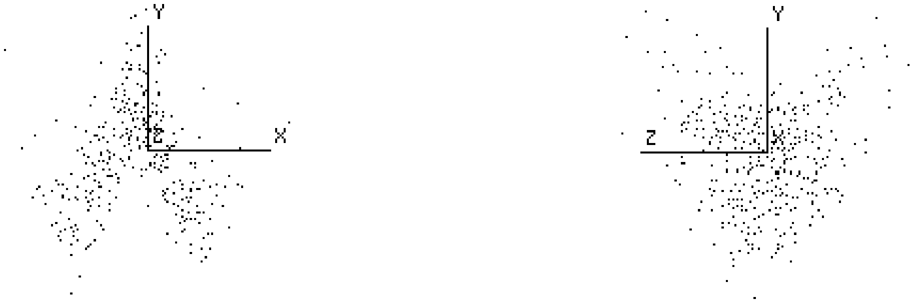

Figure 7.2: Best views of the example 7.1

where **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 **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

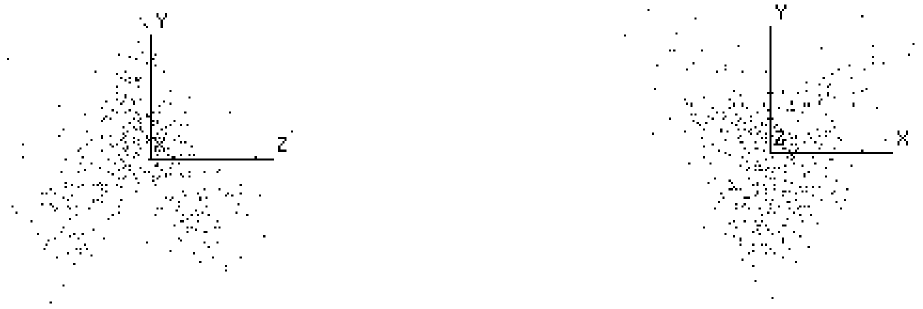

Figure 7.3: Views by the p.h.d. method for Example 7.1

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.

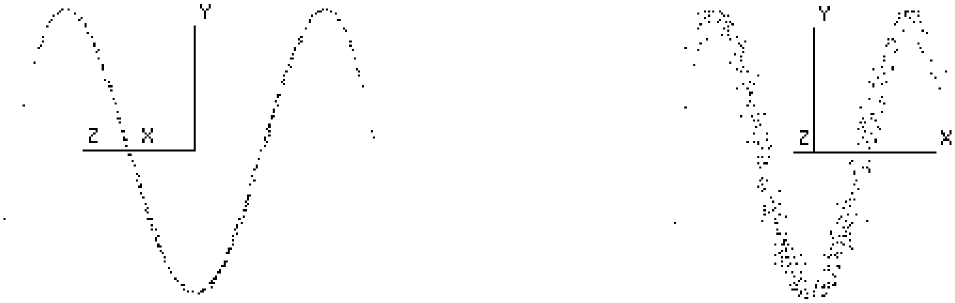

Figure 7.4: Best view(left) and the view by the p.h.d. method for Example 7.2

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

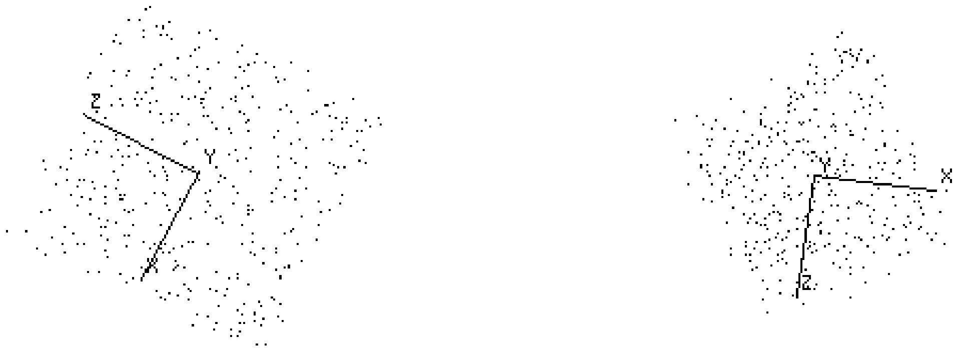

Figure 7.5: Distribution of x_1 and x_2(left), compared with distribution of first two p.h.d.'s

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.

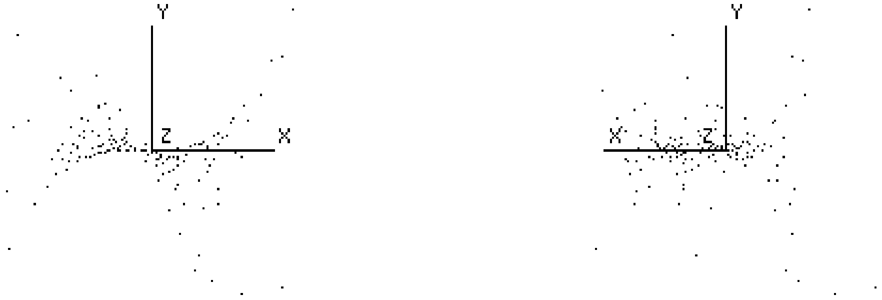

Figure 7.6: Best views of the monkey saddle.

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,

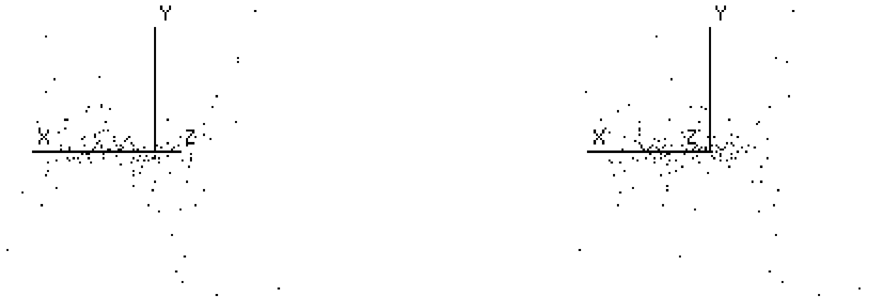

Figure 7.7: Views by the p.h.d. method for monkey saddle.

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.

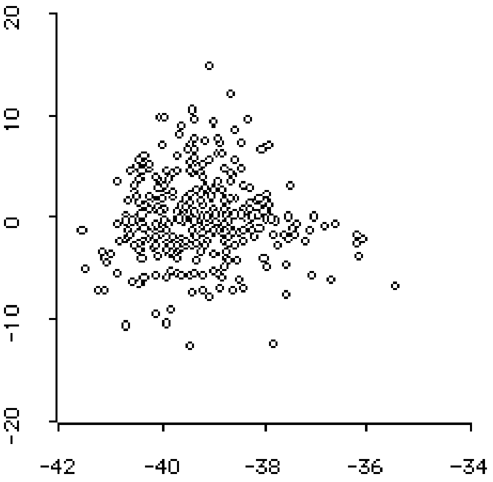

Figure 7.8: Ozone data(continued from Example 4.4.1), Residuals against the direction found by p.h.d.