Dimension-reduction in Regression

Lixing Zhu

Hong Kong Baptist University

August 25, 2008

1 Introduction

Dimension reduction was originally introduced to give a comprehensive view of the relation between predictor and response in a regression problem. The current practice of regression analysis is

- Fit a simpler model;
- Check the residual plot;
- If the residual plot does not show a systematic pattern then stop; otherwise continue to fit a more complex model.

The question is: how to give a comprehensive residual plot?

In the one dimensional case this is not a problem — we can just plot the residual e versus the predictor x or versus the predicted values $\hat{Y}$. More specifically, given $(X_1, Y_1) \ldots, (X_n, Y_n)$, we can first fit a linear regression model

$$Y = \beta_0 + \beta_1 X_i + \epsilon_i$$

and find the regression estimator:

$$\begin{cases} \hat{\beta}_1 = \sum (Y_i - \bar{Y})(X_i - \bar{X}) / \sum (X_i - \bar{X})^2, \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \end{cases}$$

Let $\widehat{Y}_i$ be the predicted values $\hat{\beta}_0 + \hat{\beta}_1 X_i$ and e_i be the residuals $Y_i - \widehat{Y}_i$. We can simply plot e_i against X_i. This is called the one-dimensional residual plot.

What should we do if X_i is a vector in $\mathbb{R}^p$? Currently two methods are in frequent use:

- We can plot e_i versus $\widehat{Y}_i$ (note that $\widehat{Y}_i$ is always one-dimensional.)
- We can use scatter plot matrix, in which we plot e_i against each predictor, and each predictor against any other predictor, forming a $(p+1) \times (p+1)$ matrix of scatter plots.

However, each of these methods are intrinsically marginal — they cannot reflect the whole picture of the regression relation. Let us see this through two examples.

Example 1.1 One hundred pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ are generated from some model, which I will show you later. Here $X_i \in \mathbb{R}^2$. We first fit a simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i.$$

We plot e_i versus $\widehat{Y}_i$ as described before. Show the scatter plot here.

From the scatter plot it appears

- The residual plot is more or less flat, which suggests that the linear model is probably adequate;
- There is heteroscedasticity — the variance appears to increase as the level of Y increases.

Thus from looking at this residual plot alone we are led to the conclusion that probably a weighted least square would be a good model for this data set. However, this is completely wrong. The data is generated by the model:

$$Y = \frac{|X_1|}{2 + (1.6 + X_2)^2} + \epsilon,$$

where $\epsilon \perp\!\!\!\perp X$, X is standard bivariate normal, and $\epsilon \sim cN(0, 1)$ for some constant $c > 0$. Show the spin view of the response surface here. The fan-shaped residual is the appearance of nonlinearity, whereas nonlinearity is completely masked in the residual plot.

Example 1.2 Again, 100 pairs, $(X_1, Y_1), \ldots, (X_{100}, Y_{100})$ are generated from some model, and the scatter plot matrix is produced. Show the scatter plot matrix here. From the scatter plot matrix the data appear to have the following features:

- Y doesn't seem to depend on X_2
- Y seem to depend on X_1 in a nonlinear way
- Y seem to depend on X_3 in a nonlinear way.

However, (X, Y) are generated from the following model:

$$Y = |X_1 + X_2| + \epsilon$$

where $\epsilon \perp\!\!\!\perp (X_1, X_2, X_3)$, (X_1, X_2, X_3) is multivariate normal with mean 0 and a non-singular covariance matrix

$$\begin{pmatrix} 1 & 0 & 0.8 \\ 0 & 0.2 & 0 \\ 0.8 & 0 & 1 \end{pmatrix}$$

Note that Y *does not* depend on X_3, and Y does depend on X_2. So what is going on here. Show the hand drawings here. First look at the equi-temperature surfaces; then look at the equi-density surfaces; explain why the effect of X_2 is masked by the collinearity in X.

So, once again, the scatter plot matrix cannot capture the true relation between X and Y. What can truly capture the relation between Y and X is the scatter plot of Y versus $X_1 + X_2$. But how can we make this plot before we know that $X_1 + X_2$ is the predictor? This is the question of dimension reduction. The goal is to find $X_1 + X_2$ before any regression analysis is performed.

A more general problem: Suppose

$$Y = f(\beta^T X) + \epsilon$$

where $\beta \in \mathbb{R}^{p\times q}$ is a matrix of dimension $p \times q$, where $q < p$; $X \perp\!\!\!\perp \epsilon$. How to estimate the direction(s) of β without estimating f? (K. C. Li, JASA (1991, 1992)).

The problem can be posed even more generally: Suppose

$$Y \perp\!\!\!\perp X|\beta^T X.$$

How to find the directions of β? Cook (1998).

2 Generalized Linear Models

2.1 Ordinary Least Square

Population derivation. $(X_1, Y_1), \ldots, (X_n, Y_n)$ independent copies of (X, Y). Choose (α, β) to minimize the loss function

$$L(\alpha, \beta) = E(Y - \alpha - \beta^T X)^2.$$

Take derivatives with respect to α and β and set the derivatives to zero and solve for α, β, it is easy to obtain:

$$\begin{aligned}\beta &= \left[E(X - EX)(X - EX)^T\right]^{-1} E(X - EX)(Y - EY)\\ &= [\mathrm{var}(X)]^{-1}\,\mathrm{cov}(X, Y)\\ \alpha &= E(Y) - \beta^T E(X).\end{aligned}$$

Here var is the variance matrix of X, and cov is the covariance vector between X and Y. In our context we frequently express this in the standardized form. That is, if $E(X) = 0$, $\mathrm{var}(X) = I$, then the ols vector can be expressed as simply:

$$\beta = E(XY).$$

Sample version. Notation: $W_1, \ldots, W_n$ independent copies of (X, Y). We write

$$E_n(W) = n^{-1}\sum_{i=1}^{n} W_i.$$

The motivation of this notation is that $E_n(W)$ is the expectation of W under the empirical distribution F_n, that put equal weight n^{-1} on each observation W_i. By the same token, we can define $\mathrm{var}_n(W, V)$, $\mathrm{cov}_n(W, V)$. For example

$$\mathrm{cov}_n(W, V) = E_n((W - E_n W)(V - E_n V)^T).$$

In this notation, we have

$$\hat{\beta} = [\mathrm{var}_n(X)]^{-1}\,\mathrm{cov}_n(X, Y).$$

We know that, if the linear model

$$Y = \alpha + \beta^T X + \epsilon,$$

where $\epsilon \perp\!\!\!\perp X$ and ϵ has a finite variance σ^2, we have

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, A),$$

where $A = \sigma^2 \left[\text{var}(X)\right]^{-1}$. For inference about β, replace $\text{var}(X)$ by its sample version $\text{var}_n(X)$.

2.2 Nonlinear Least Square

The idea is basically the same. Suppose that

$$Y = f(\beta^T X) + \epsilon,$$

where f is *known*, and ϵ has a finite variance σ^2. Estimate β by minimizing the loss function

$$E_n(Y - f(\beta^T X))^2.$$

In this case, there is no explicit solution and we have to do this numerically, usually by the Newton-Raphson algorithm. If the mentioned nonlinear model is correct, then we have

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, A^{-1}),$$

where

$$A = E\left[\{f'(\beta^T X)\}^2 X X^T\right] / \sigma^2$$

Again, for estimating the variance and so on we replace the population mean E by the sample mean E_n.

2.3 Generalized Linear Models

Reference: McCullagh & Nelder (1989). *Generalized Linear Models, 2nd edition*, Chapman and Hall.

More generally, (X, Y) may not be directly related by the linear or nonlinear function f. For example, if Y is binary, or categorical, and X is continuous, then it is impossible to describe the relation between (X, Y) by any of the above models. Furthermore, there may be heteroscedacity; that is, the variance of Y for a value of X may depend on the level of X. This is typically the case if Y is categorical.

Again, suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent copies of (X, Y). In Generalized Linear Models we assume that $Y|X$ follows a distribution in the *Linear Exponential Family*; that is, the conditional density of $Y|X$ is

$$e^{\theta y - b(\theta)}$$

with respect to some measure of y, let's say $\nu_x(y)$, that may depend on x. Here θ is a function of x.

An important fact about the linear exponential family.

Theorem 2.1 *The function $b(t+\theta) - b(\theta)$ is the cumulant generating function of the natural exponential family; that is,*

$$b(\theta + t) - b(\theta) = \log M_{Y|x}(t).$$

Here, $M_{Y|x}(t)$ is the moment generating function of the conditional density of $Y|X = x$:

$$M_{Y|x}(t) = E\left[e^{\theta Y}|X = x\right].$$

Proof. Note that

$$\begin{aligned} E(e^{tY}|X = x) &= \int e^{ty+\theta y-b(\theta)} d\nu_x(y) \\ &= \int e^{(t+\theta)y-b(t+\theta)+b(t+\theta)-b(\theta)} d\nu_x(y) \\ &= e^{b(t+\theta)-b(\theta)} \int e^{(t+\theta)y-b(t+\theta)} d\nu_x(y) \\ &= e^{b(t+\theta)-b(\theta)}. \end{aligned}$$

Now take logarithm on both sides to get the desired result. □

A consequence of this theorem is that

$$\begin{aligned} b'(\theta) &= E_\theta(Y|x) \\ b''(\theta) &= \mathrm{var}_\theta(Y|x) \\ b^{(k)}(\theta) &= \mathrm{cum}_\theta(Y|x) \end{aligned}$$

for all $k = 3, 4, \ldots$.

In generalized linear models, we assume that $E_\theta(Y|X)$ is a function of $\beta^T X$, which is called the linear predictor. We right this as $\mu(\beta^T X)$. Here μ is called the mean function and its inverse μ^{-1} is called the link function. Also, we assume that $\mathrm{var}_\theta(Y|X)$ is also a function of the linear predictor $\beta^T X$, and write this function as $V(\beta^T X)$. We call this function the variance function.

The parameter θ is called the cannonical parameter. Because μ is a function of $\beta^T X$, so is the cannonical parameter: from $b'(\theta) = \mu(\beta^T X)$ we can deduce that

$$\theta = (b')^{-1}(\mu(\beta^T X)) = \left((b')^{-1} \circ \mu\right)(\beta^T X).$$

Definition 2.1 *The link function that makes the map $\beta^T X \mapsto \theta$ the identity map is called the natural link function.*

What can make this map identity? This is

$$\mu = b' \quad \text{or} \quad \mu^{-1} = (b')^{-1}.$$

So in Generalized Linear Models $\mu^{-1} = (b')^{-1}$ is the natural link function.

Example 2.1 Suppose that $Y|x$ is distributed as $N(\theta, \sigma^2)$, where, for simplicity, assume σ is known, and take it to be 1. Then the conditional density of $Y|X$ is

$$\begin{aligned} f_\theta(y|x) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2-2\theta y+\theta^2}{2}} \\ &= e^{\theta y - \theta^2/2} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}. \end{aligned}$$

So the conditional density of $Y|x$ is

$$e^{\theta y - \theta^2/2}$$

with respect to

$$d\nu_x(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy,$$

where dy is the Lebesgue measure. (In this case $\nu_x(y)$ does not depend on x). The cumulant generating function is $b(\theta) = \theta^2/2$. Thus $b'(\theta) = \theta$ is an indentity mapping. So the natural link function is the identity mapping $\mu^{-1}(t) = t$.

Example 2.2 Suppose that $Y|x$ has a Poisson distribution $\mathrm{Po}(\lambda)$. Then

$$f_\lambda(y|x) = \frac{\lambda^y}{y!} e^{-\lambda} = e^{y \log \lambda - \lambda} \frac{1}{y!} = e^{y\theta - e^\theta} \frac{1}{y!}.$$

So the density of $Y|x$ is

$$e^{\theta y - e^\theta}$$

with respect to the measure

$$d\nu_x(y) = \frac{1}{y!} dy.$$

Here, dy is the counting measure. The cumulant generating function $b(\theta)$ is e^θ. So $b'(\theta) = e^\theta$, and the link function is $\mu^{-1}(t) = \log(t)$. So the model for the mean function is

$$E(Y|x) = e^{\beta^T x}.$$

Because $b''(\theta) = e^\theta$, the model for variance function is also

$$\mathrm{var}(Y|x) = e^{\beta^T x}.$$

This is called the log-linear or Poisson regression model.

Example 2.3 Suppose that the conditional distribution of $Y|x$ is binomial(p). Then

$$\begin{aligned} f_p(y|x) &= \left(\begin{array}{c} n \\ y \end{array} \right) p^y (1-p)^{n-y} \\ &= e^{y \log \frac{p}{1-p} + n \log(1-p)} \left(\begin{array}{c} n \\ y \end{array} \right) \\ &= e^{\theta y - n \log(1+e^\theta)} \left(\begin{array}{c} n \\ y \end{array} \right). \end{aligned}$$

So the conditional density of $Y|x$ is

$$e^{\theta - n \log(1+e^\theta)}$$

with respect to the measure

$$d\nu_x(y) = \left(\begin{array}{c} n \\ y \end{array} \right) dy,$$

where dy is the counting measure on $\{0, 1, \ldots, n\}$. The cumulant generating function is

$$b(\theta) = n \log(1 + e^\theta).$$

Hence

$$b'(\theta) = \frac{ne^\theta}{1 + e^\theta}.$$

The natural link function is

$$\mu^{-1}(t) = \log \frac{t/n}{1 - t/n}.$$

The mean and variance functions are

$$\begin{aligned} E(Y|x) &= \frac{ne^{\beta^T x}}{1 + e^{\beta^T x}} \\ \mathrm{var}(Y|x) &= \frac{ne^{\beta^T x}}{(1 + e^{\beta^T x})^2}. \end{aligned}$$

This model is called the logistic regression or logit regression.

Estimation for generalized linear models. The log-likelihood is for a single observation is

$$y\theta - b(\theta).$$

The score function (the derivative of the log likelihood function) is

$$\begin{aligned} &\theta'(\beta^T X) Y X - b'(\theta(\beta^T X)) \theta'(\beta^T X) X \\ &\qquad = X \theta'(\beta^T X)(Y - b'(\beta^T X)). \end{aligned}$$

However, recall that

$$(b' \circ \theta)(s) = \mu(s).$$

Taking derivative on both sides to obtain

$$b''(\theta(s))\theta'(s) = \mu'(s).$$

Therefore,

$$\theta'(s) = \mu'(s)/b''(\theta(s)) = \mu'(s)/V(s).$$

Therefore the score function can be rewritten, in terms of the mean and variance function, as

$$\left\{X\mu'(\beta^T X)/V(\beta X)\right\}(Y - \mu(\beta^T X)).$$

Summing up over the n observations, we have the likelihood equation

$$\sum_{i=1}^{n} \left\{X_i\mu'(\beta^T X_i)/V(\beta^T X_i)\right\}(Y_i - \mu(\beta^T X_i)) = 0$$

The estimator $\hat{\beta}$ is the solution to this equation. Again, usually there is no explicit solution to this equation, but this can be solved a numerical method such as the Newton-Raphson algorithm. Under regularity conditions, we have the following convergence:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, I^{-1}(\beta)),$$

where $I(\beta)$ is the Fisher information matrix

$$I(\beta) = E\left[\frac{\{\mu'(\beta^T X)\}^2}{V(\beta^T X)} XX^T\right].$$

In practice, we replace the Fisher information by its sample estimate:

$$E_n\left[\frac{\{\mu'(\hat{\beta}^T X)\}^2}{V(\hat{\beta}^T X)} XX^T\right],$$

which converges at the $\sqrt{n}$-rate the population version $I(\beta)$.

If we use the natural link, then the likelihood equation can be further simplified. Recall that

$$b''(\theta(s))\theta'(s) = \mu'(s).$$

Under the natural link, $\theta(s) = s$ and therefore $\theta'(s) = 1$. Recall that $b''(\theta(s)) = V(s)$. So we have

$$V(s) = \mu'(s).$$

Thus the likelihood equation is

$$\sum_{i=1}^{n} X_i\left(Y_i - \mu(\beta^T X_i)\right) = 0.$$

Correspondly, the Fisher information is simplified to

$$I(\beta) = E\left[\mu'(\beta^T X)XX^T\right].$$

3 Dimension reduction: some basic concepts

3.1 Central Space

As we explained before, the goal of dimension reduction is to seek $\beta \in \mathbb{R}^{p\times q}$ $(q < p)$, such that

$$Y \perp\!\!\!\perp X|\beta^T X.$$

Note that, if A is any $q \times q$ non-singular matrix, then $\beta^T X$ and $A^T\beta^T X$ has one-to-one correspondence. Therefore, $Y \perp\!\!\!\perp X|\beta^T X$ if and only if $Y \perp\!\!\!\perp X|(\beta A)^T X$. Note that β and βA have the same column space. So we define the column space of β as a dimension reduction subspace. For a matrix A we denote by $\mathcal{S}(A)$ the subspace spanned by the columns of A.

Also note that, if γ is another matrix such that $\mathcal{S}(\beta) \subset \mathcal{S}(\gamma)$, then $\beta^T X$ is a measurable function of $\gamma^T X$. Thus $Y \perp\!\!\!\perp X|\beta^T X$ implies $Y \perp\!\!\!\perp X|\gamma^T X$. In other words, if $\mathcal{S}_1$ is a dimension reduction space and $\mathcal{S}_1 \subset \mathcal{S}_2$, then $\mathcal{S}_2$ is also a dimension reduction space. Therefore, we are naturally interested in the smallest dimension reduction space, which achieves the maximal reduction of the dimension of X. Under mild regularity conditions, the intersection of all dimension reduction spaces is itself a dimension reduction space. This space is called the Central Space.

Definition 3.1 *The Central Space for (X, Y) is the intersection of all dimension reduction spaces for (X, Y). This space is written as $\mathcal{S}_{Y|X}$.*

Reference: Cook (1994, 1998).
Thus the goal of dimension reduction is to find the Central Space $\mathcal{S}_{Y|X}$.

3.2 Invariance of central space

An invariance property of the dimension reduction space.

Theorem 3.1 *Let $\mathcal{S}_{Y|X}$ be the Central Space for (X, Y). Let $Z = AX + b$. Then*

$$\mathcal{S}_{Y|Z} = A^{-T}\mathcal{S}_{Y|X}.$$

Proof. Let β be a p by q matrix whose columns form a basis in $\mathcal{S}_{Y|X}$. Then

$$Y \perp\!\!\!\perp X|\beta^T X.$$

Because $X = A^{-1}(Z - b)$, this means that

$$Y \perp\!\!\!\perp A^{-1}(Z - b)|\beta^T A^{-1}(Z - b).$$

This is equivalent to

$$Y \perp\!\!\!\perp Z|\beta^T A^{-1} Z.$$

Or

$$Y \perp\!\!\!\perp Z | (A^{-T}\beta)^T Z.$$

Hence $A^{-T}\mathcal{S}_{Y|X}$ is a dimension reduction space for (Z, Y). Consequently $\mathcal{S}_{Y|Z} \subset \mathcal{S}_{Y|X}$. The inverse inclusion can be proved similarly. □

In future discussions, it will prove convenient to work with standardized X. Let $\mu = E(X)$ and $\Sigma = \text{var}(X)$. Let

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

If we can find $\mathcal{S}_{Y|Z}$. Then we can use the relation $X = \Sigma^{1/2}Z + \mu$ and the above invariance property to derive $\mathcal{S}_{Y|X} = \Sigma^{-1/2}\mathcal{S}_{Y|Z}$. Hence, we will, without loss of generality, make the following assumption.

Assumption 3.1 *Throughout the following discussion we will assume that*

$$E(X) = 0 \quad var(X) = I_p.$$

3.3 Sufficient plot

Once we know the Central Space, we can construct a comprehensive scatter plot or residual plot that do not lose information as the scatter matrix plot or residual versus predictor plot. Let $\beta = (\beta_1, \ldots, \beta_q)$ be a basis for the central space. The sufficient plot is plotting Y or e versus $\beta_1^T X, \ldots, \beta_q^T X)$. In one dimensional case, this is simply a scatter plot. In two dimensional case, it is the plot of Y versus $\beta_1^T X, \beta_2^T X$. We can use a spin software to have a comprehensive view of the data. Usually this will suffice for most of the data analysis.

For example, returning to Example 1.2. The model is

$$Y = |X_1 + X_3| + \epsilon.$$

Thus the central space is spanned by $(1, 0, 1)$. The sufficient plot is the scatter plot of Y versus $X_1 + X_3$. Show this plot here.

The lower dimension central space has meanings.

- **0-D Structure**. If $q = 0$. Then that means that $Y \perp\!\!\!\perp X$; that is, there is no relation between X and Y. Thus we can use the inference procedure that we will explain later on to test whether X and Y independent without estimating the response surface.

- **1-D Structure**. This is the case where $Y \perp\!\!\!\perp X | \beta^T X$ where β is a vector. Many regression problems have 1-D structure. For example, all the generalized linear models are of 1-D structure. The so called *Single Index model* in econometrics includes precisely this instance of conditional independence. As a special case, consider

$$Y = f(\beta^T X) + \sigma(\beta^T X)\epsilon$$

That is, both the mean and the variance are functions of $\beta^T X$. This is 1-D structure.

- **2-D Structure**. For example, this occurs when the mean and the variance depends on 2 different linear combinations of X; that is

$$Y = f(\beta_1^T X) + \sigma(\beta_2^T X)\epsilon.$$

4 Estimation of CS: OLS

4.1 Linear Conditional Mean

We will make the following key assumption, which we call the *linear conditional mean* assumption.

Assumption 4.1 *Let β be a $\mathbb{R}^{p\times q}$ be a matrix whose columns form an orthonormal basis in $\mathcal{S}_{Y|X}$. We will assume that $E(X|\beta^T X)$ is a linear function of X; that is, the conditional mean of X given $\beta^T X$ is linear in X.*

We first make some notes on the intuitions and implications of this assumption.

- In practice, we do not know β at the outset. So we typically replace this assumption by $E(X|\gamma^T X)$ is linear in X for all $\gamma \in \mathbb{R}^{p\times k}$, $k = 1, \ldots, p$. This is equivalent to assuming that X has an elliptical distribution. Since X is already standardized, this is equivalent to saying that X has a circular contoured density. In other words, the density of X depends on X only through $\|X\|$, the Euclidean norm of X.

- Elliptically contoured distribution can often be achieved by appropriate transformation of the original data; for example, by taking a certain power of the data, or take logarithm of the data. This is more successful for some cases than for others. I will explain this in detail through an example in a short while.

- In regression analysis it is always preferable to transform X rather than transforming X, because transforming X does not change the interpretation of the response. Transforming X does not change the interpretation — because we are looking for a function of X any way.

- Hall & K. C. Li (1993) demonstrated that, if the original dimension p is much larger than the structural dimension q, then $E(X|\beta^T X)$ is approximately linear in X. The argument is that taking conditional example is like taking average. When you take average over a large number of variables, then the result behaves more or less like multivariate normal. But multivariate normal is elliptically-contoured distribution. Therefore the linear conditional mean assumption should hold roughly for the conditional expectation.

4.2 Transformation through examples

Example 4.1 Western economists more or less agree that the living standard of a country can be measured by how much labour hours is need for buying a big mac. There is a fast food chain called Berger King in America. A big mac is a hamburger sold at burger king. The fewer labor hours needed for buying a big mac the higher the living standard. Reversely, if a country need a lot of labor hours to buy a big mac, then the living standard of that country is low. So there are a lot of studies about big mac; the following data is collected from one such studies.

Data taken from Rudolf Enz, "Prices and Earnings Around the Globe", 1991 edition, Published by the Union Bank of Switzerland. The data give average values in 1991 on several economic indicators for 45 world cities. All prices are in US dollars, using currency conversion at the time of publication.

Variable	Meaning
BigMac	Min labor to buy a BigMac and fries
Bread	Min labor to buy 1 kg bread
BusFare	Lowest cost of 10k public transit
EngSal	Electrical eng annual salary, 1000s
EngTax	Tax rate paid by engineer
Service	Annual cost of 19 services
TeachSal	Primary teacher salary, 1000s
TeachTax	Tax rate paid by primary teacher
VacDays	Ave days vacation per year
WorkHrs	Ave hours worked per year
City	Name of city

Note that before transformation, the data look quite nonlinear and non elliptical. However, it seems that taking logarithm, on all or on some of the variables, does a good job in transforming the data into elliptically-contoured distribution.

Example 4.2 Ozone data. There are 330 measurements of the ozone levels as well as 8 covariates. This is composed of the following variables:

Variables	Meaning
Height	Vandenburg 500 millibar height (m)
Humidity	humidity, percent
InversionHt	Inversion base height, feet
Ozone	Ozone conc., ppm, at Sandbug AFB.
Pressure	Daggett pressure gradient (mm Hg)
Temp2	inversion base temperature, degrees F .
Temperature	Temperature F. (max?).
Visibility	Visibility (miles)
WindSpeed	wind speed, mph

For the purpose of demonstration, we will exclude windspeed, visibility, and InversionHt. Show here the scatter plot matrix and the effect of transformation. Again, the original data looks quite non-elliptical, but after transformation the situation is much improved. Logarithm seems to work well for all variables except humidity, for which a power transformation (about 1.75) seems to work better.

4.3 Conditional mean & projection

4.3.1 Conditionaal mean as L-2 projection

Consider two random variables U and V. Let P be the joint distribution of (U, V). We will consider the class of all functions of $f(u, v)$ that are square-integrable with respect to P. This class is usually denoted by $L^2(P)$, which reads "L-2 space with respect to the measure P."

The $L^2(P)$ space is a linear space; if f_1 is square-integrable and f_2 is also square integrable, then $f_1 + f_2$ is also squared integrable. This is because

$$\int (f_1 + f_2)^2 dP = \int f_1^2 dP + 2 \int f_1 f_2 dP + \int f_2^2 dP.$$

The first and the last term are finite. The second term, by the Cauchy-Schwarz inequality,

$$\left(\int f_1 f_2 dP \right)^2 \leq \int f_1^2 dP \int f_2^2 dP.$$

Because f_1 and f_2 are square-integrable, the right hand side is finite. Also, it is obvious that, if f is square-integrable, then so is αf. Thus $L^2(P)$ is a linear space.

Moreover, we can define in $L^2(P)$ the inner product

$$\langle f_1, f_2 \rangle = \int f_1 f_2 dP = E(f_1(X) f_2(X)).$$

It is easy to show that this is an inner product in the technical sense of the word. From the inner product we can define the length of a random variable in $L^2(P)$, as follows:

$$\|f\|^2 = \int f^2 dP = E f^2(X).$$

Finally, it can be shown that $L^2(P)$ is a closed set in terms of the metric $\|\cdot\|$ just defined. That is, any Cauchy-sequence in $L^2(P)$ converges, in terms of $\|\cdot\|$ to an element of $L^2(P)$.

In summary, $L^2(P)$ is a Hilbert space, which is an extension of the Euclidean space (in which we live), and inherits almost all the nice properties of a Euclidean space. Most importantly, it inherits the properties such as orthogonality and projection. Two functions, or two random variables, $f_1(X)$ and $f_2(X)$, in $L^2(P)$ are orthogonal if

$$\langle f_1, f_2 \rangle = 0.$$

In this case, we will write $f_1 \perp f_2$. If $\mathcal{L}$ is a subset of $L^2(P)$ and f is orthogonal to every element in $\mathcal{L}$, then we write $f \perp \mathcal{L}$. That $L^2(P)$ also inherits from a Euclidean space the property of projection is shown by the following theorem.

Theorem 4.1 *If $\mathcal{L}$ is a (closed) linear subspace of $L^2(P)$, and if f is any element in $L_2(P)$. Then there is a (almost everywhere P) unique element f_0 in $\mathcal{L}$, such that $f - f_0 \perp \mathcal{L}$. In other words,*

$$\langle f - f_0, h \rangle = 0$$

for all h in $L^2(P)$. Moreover, the element f_0 is the (almost everywhere P unique) element in $\mathcal{L}$ that is the closest to $L^2(P)$. That is

$$\|f - f_0\| \leq \|f - h\|$$

for all h in $\mathcal{L}$.

Definition 4.1 *This unique element in $\mathcal{L}$ is called the orthogonal projection (the shadow of the sun at noon) of the random variable f onto $\mathcal{L}$, and will be written as $P_{\mathcal{L}}(f)$.*

In this formulation the conditional expectation $E(U|V)$ can be viewed as the orthogonal projection of the random variable U onto the space of V.

Let P be the joint distribution of (U, V) and $L^2(P)$ be the class of all functions of (u, v) that are square-integrable with respect to P. Let Q be the marginal distribution of V and let $L^2(Q)$ be the class of all functions of V that are square-integrable with respect to Q. It turns out the conditional expectation $E(U|V)$ is nothing but the orthogonal projection of U onto $L^2(Q)$.

Theorem 4.2

$$E(U|V) = P_{L^2(Q)}(U).$$

Proof. By Theorem 4.1, it suffices to show that $U - E(U|V) \perp L^2(Q)$. That is, for any $h(V)$ in $L^2(Q)$, we have

$$E\left[(U - E(U|V))h(V)\right] = 0. \tag{1}$$

Note that

$$\begin{aligned} E(Uh(V)) &= E\left[E(Uh(V))|V\right] \\ &= E\left[E(U|V)h(V)\right]. \end{aligned}$$

Substract the right hand side from the left hand side gives (1), as desired. □

4.3.2 Projection in Euclidean space

Now let us turn to a simpler problem. Consider the Euclidean space $\mathbb{R}^p$. Let v be any vector in $\mathbb{R}^p$. It is easy to see that $\mathbb{R}^p$ is a linear space. Let's define the inner product in $\mathbb{R}^p$ as

$$\langle v_1, v_2 \rangle = a^T b.$$

It is well known that $\mathbb{R}^p$, together with this inner product, is a Hilbert space. In this subsection we will study the orthogonal projection in this Hilbert space.

Let $\beta_1, \ldots, \beta_q$ be an orthonormal subset of $\mathbb{R}^p$. (Recall that $q < p$.) and let v be any vector in $\mathbb{R}^p$. We are interested in the projection of v onto the subspace spanned by $\beta_1, \ldots, \beta_q$. Let β be the matrix $(\beta_1, \ldots, \beta_q)$. Let $\mathcal{S}(\beta)$ be the subspace spanned by the vectors $\beta_1, \ldots, \beta_q$.

Theorem 4.3

$$P_{\mathcal{S}(\beta)}(v) = \beta\beta^T v.$$

PROOF. By Theorem 4.1, it suffices to show that for any vector h in $\mathcal{S}(\beta)$, we have

$$(v - \beta\beta^T v)^T h = 0. \tag{2}$$

Because h is a vector in $\mathcal{S}(\beta)$, it is a linear combination of the vectors $\beta_1, \ldots, \beta_q$. In other words, $h = \beta w$ for some vector w in $\mathbb{R}^q$. Note that

$$\begin{aligned}(\beta\beta^T v)^T h &= (\beta\beta^T v)^T \beta w \\ &= v^T \beta\beta^T \beta w \\ &= v^T \beta w \\ &= v^T h.\end{aligned}$$

Subtract the left hand side from the right hand side gives (2), as desired. □

The matrix $\beta\beta^T$ is called the projection matrix, and will be written as P_β.

4.3.3 Conditional mean as Euclidean projection

The geometric implication of the linear conditional mean assumption is that the conditional expectation $E(X|\beta^T X)$ coincides with $P_\beta(X)$. That is, the L-2 projection and Euclidean projection are one and the same.

Theorem 4.4 *If Assumption 4.1 holds, then*

$$E(X|\beta^T X) = P_\beta(X).$$

PROOF. Let e_i be the vector in $\mathbb{R}^p$ whose ith element is 1 and the rest elements are 0. We need to show that

$$E(X_i|\beta^T X) = e_i^T P_\beta(X).$$

Because $E(X_i|\beta^T X)$ is linear in X, it can be written as $a_i + v_i^T \beta^T X$ for some $a_i \in \mathbb{R}$ and $v_i \in \mathbb{R}^q$. However, because

$$E\left[E(X_i|\beta^T X)\right] = E(X_i) = 0,$$

we have

$$E(a_i + v_i^T \beta^T X) = a_i = 0.$$

Hence

$$E(X_i|\beta^T X) = v_i^T \beta^T X.$$

Let Q be the distribution of $\beta^T X$ and $L^2(Q)$ be the class of all the square-integrable functions of $\beta^T X$. Because, by Theorem 4.2, $E(X_i|\beta^T X)$ is the orthogonal projection onto $L^2(Q)$, we have, for any $h(\beta^T X) \in L^2(Q)$,

$$E\left[(X_i - E(X_i|\beta^T X))h(\beta^T X)\right] = 0.$$

In particular, take $h(\beta^T X)$ to be $\beta_1^T X, \ldots, \beta_q^T X$, we have, for any $r = 1, \ldots, q$,

$$E\left[(X_i - v_i^T \beta^T X)\beta_r^T X\right] = 0.$$

Now let V be the matrix $(v_1, \ldots, v_p)$. The above equations (pq of them) can be re-written in the matrix form

$$E\left[(X - V^T \beta^T X)X^T \beta\right] = 0.$$

Now apply the relation $E(XX^T) = I_p$ to obtain

$$\beta - V^T \beta^T \beta = 0.$$

But recall that $\beta^T \beta = I_q$. Thus $V = \beta^T$. It follows that

$$E(X|\beta^T X) = V^T \beta^T X = \beta \beta^T X = P_\beta X,$$

as desired. □

4.3.4 Conditional mean as a self-adjoint operator

Let W be a random variable and let $A = E(\cdot|W)$; that is A is the mapping $U \mapsto E(U|W)$. Then, it is easy to see that A is a linear operator from $L^2(P)$ to $L^2(P)$. In fact, as we have already shown, A is a projection operator. Now any projection operator is self-adjoint. That is, in this case, let U and V be two random variables, then

$$\langle AU, V\rangle = \langle U, AV\rangle.$$

In other words,

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

This can also be proved directly, as follows:

$$\begin{aligned} E\left[E(U|W)V\right] &= E\left\{E\left[E(U|W)V|W\right]\right\} \\ &= E\left[E(U|W)E(V|W)\right] \\ &= E\left\{E\left[UE(V|W)|W\right]\right\} \\ &= E\left[UE(V|W)\right], \end{aligned}$$

as desire.

4.4 The OLS estimator of CS

4.4.1 Population development

We first observe a simple fact.

Lemma 4.1 *If $U \perp\!\!\!\perp V|W$, then*

$$E(U|V,W) = E(U|W).$$

PROOF. Let $f_{U|VW}(u|v,w)$ be the conditional density of $U|V,W$, and let $f_{U|W}(u|v)$ be the conditional density of $U|V$. Then,

$$f_{U|VW} = \frac{f_{UVW}}{f_{VW}} = \frac{f_{UV|W} f_W}{f_{VW}} = \frac{f_{U|W} f_{V|W}}{f_{V|W}} = f_{U|W},$$

from which the asserted result follows easily. □

As we mentioned before, when $E(X) = 0$, $\text{var}(X) = I_p$, then the population OLS vector is simply $E(XY)$.

Theorem 4.5 *Suppose that Assumption 4.1 holds. Then*

$$E(XY) \in \mathcal{S}_{Y|X}$$

PROOF. Note that

$$E(XY) = E(E(XY|X)) = E(XE(Y|X)). \tag{3}$$

However, because $X \perp\!\!\!\perp Y|\beta^T X$, we have, by Lemma 4.1,

$$E(Y|X) = E(Y|X, \beta^T X) = E(Y|\beta^T X).$$

Therefore, the right hand side of (3) is $E(XE(Y|\beta^T X))$. However, because conditional expectation is a self-adjoint operator, we have

$$E(XE(Y|\beta^T X)) = E(E(X|\beta^T X)Y).$$

Now recall that, under Assumption 4.1, the L-2 projection coincides with the Euclidean projection, as shown in Theorem 4.4, we have

$$E(E(X|\beta^T X)Y) = E(P_\beta(X)Y) = P_\beta E(XY).$$

Thus

$$E(XY) = P_\beta E(XY).$$

In other words, $E(XY)$ equals its projection onto $\mathcal{S}(\beta)$. Therefore $E(XY)$ must be in the range of projection P_β, which is $\mathcal{S}_{Y|X}$ □

Now return to Generalized Linear Models. Suppose $Y|X$ has density

$$e^{\theta y - b(\theta)}$$

with respect to some measure $\nu(y)$. Recall that, in generalized linear models,

$$\theta = \left[(b')^{-1} \circ \mu\right](\beta^T X).$$

Note that here β is a p-dimensional vector instead of a matrix. It follows that $f(Y|X)$ is a function of $\beta^T X$ and Y. So $f_{Y|X} = f_{Y|\beta^T X}$. In other words, $Y \perp\!\!\!\perp X|\beta^T X$. Consequently, $E(XY) \in \mathcal{S}(\beta)$. Therefore we have the following theorem, which we have mentioned at the beginning of this series.

Theorem 4.6 *Suppose that $Y|X$ follows a generalized linear model as we have described. Suppose that Assumption 4.1 holds. Then $E(XY)$ is proportional to β (regardless of the link function μ^{-1}).*

This is essentially (in fact, more general) the result stated in Li & Duan (1989), one of the first papers in dimension reduction.

The intuition of this theorem is clear. Show the intuition by illustrating via a piece of paper here.

4.4.2 Estimating procedure

At the population version, the estimating procedure can be described as follows. First, standardize X to be $Z = \Sigma^{-1/2}(X - \mu)$. Estimate a vector in $\mathcal{S}_{Y|Z}$. Then transform back to $\mathcal{S}_{Y|X} = \Sigma^{-1/2}\mathcal{S}_{Y|Z}$. At the sample level, we follow these steps.

- Compute

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

  Standardize $X_1, \ldots, X_n$ to be

  $$\widehat{Z}_i = \widehat{\Sigma}^{-1/2}(X_i - \hat{\mu}).$$

- Center $Y_1, \ldots, Y_n$ to $\widehat{Y}_i = Y_i - E_n(Y)$.
- Let $\hat{\gamma}$ be the the vector $E_n(\widehat{Z}\widehat{Y})$. This is an estimator of $E(ZY)$, a vector in $\mathcal{S}_{Y|Z}$.
- Let $\hat{\beta} = \widehat{\Sigma}^{-1/2}\hat{\gamma}$, and this is an estimator of $\mathcal{S}_{Y|X}$.

In this instance,

$$\begin{aligned}
\hat{\beta} &= \widehat{\Sigma}^{-1/2}\hat{\gamma} \\
&= \widehat{\Sigma}^{-1/2}\widehat{\Sigma}^{-1/2}E_n(X - E_n(X))E_n(Y - E_n(Y)) \\
&= [\mathrm{var}_n(X)]^{-1}\,\mathrm{cov}_n(X, Y).
\end{aligned}$$

This is exactly the OLS estimator of β, as described in Chapter 1.

Note that

$$\begin{aligned} E_n(X - E_n(X))(Y - E_n(Y)) &= E_n(XY) - E_n(X)E_n(Y) \\ &= E(XY) - E(X)E(Y) + O_p(n^{-1/2}) \\ &= \mathrm{cov}(X, Y) + O_p(n^{-1/2}). \end{aligned}$$

And, similarly,

$$\mathrm{var}_n(X) = \mathrm{var}(X) + O_p(n^{-1/2}).$$

So we have

$$\hat{\beta} - \beta = O_p(n^{-1/2}).$$

Thus we have proved the following theorem.

Theorem 4.7 *The OLS estimator for $\mathcal{S}_{Y|X}$ is $\sqrt{n}$-consistent. In other words, it converges at $\sqrt{n}$-rate to a vector that belongs to $\mathcal{S}_{Y|X}$.*

4.4.3 A Simulated Example

Example 4.3 Let $n = 100$, and $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent copies of (X, Y). Here X is a random vector in $\mathbb{R}^p$, where $p = 10$. We write X_i as $N(0, I_p)$ and generate Y according to the following model:

$$Y_i = e^{X_{i,1}} + \epsilon_i,$$

where $\epsilon_1, \ldots, \epsilon_n$ are iid $0.2N(0, 1)$.

Note that the central space in this example is the one spanned by the vector $(1, 0, \ldots, 0)^T$. We now compute the OLS estimator. The components of $\hat{\beta}$ is calculated to be

$$\begin{array}{ccccc} 1.821 & 0.005 & -0.170 & 0.085 & 0.014 \\ -0.120 & -0.147 & -0.086 & -0.216 & -0.033 \end{array}$$

We see that the first component is large, whereas the other component is much smaller. Thus this vector is roughly aligned with $(1, 0, \ldots, 0)$, the vector that spans the central space.

4.5 Refinement of the OLS property

Reference B. Li, D.R. Cook, F. Chiaromonte (2002, *Ann. Statist.*).

In this section we will study a special case of a theorem in this paper, which relax the Linear Conditional Mean assumption 4.1. Recall that in Assumption 4.1, we assume that $E(X|\beta^T X)$ is linear in X. Because we do not know β in advance, we have to assume this holds for any matrix γ; that is, we need to assume elliptically-contoured distribution

for X. Let $\eta = E(XY)$. In the previous section we showed that η belongs to the central space under Assumption 4.1. So it is natural to ask, would it be sufficient to require $E(X|\eta^T X)$ be linear in X? If this is true then it would bring big advantage — because, we can estimate η by OLS, and we can actually check if X is linear in this direction. In this section we will show that, for all generalized linear models with natural link, this is indeed the case; that is, if $E(X|\eta^T X)$ is linear in η, then η belongs to the central space.

First we state the following lemma.

Lemma 4.2 *The cumulant function $b(\theta)$ for a linear exponential family is a convex function.*

PROOF. Suppose $Y|\theta$ has density $e^{\theta y - b(\theta)}$ with respect to measure $\nu(y)$. Then, as we have seen before, $b''(\theta) = \mathrm{var}_\theta(Y) > 0$. Therefore b is a convex function. □

Theorem 4.8 *Suppose that $Y|X$ has a density in linear exponential family; that is, $f_\theta(y|x)$ is $e^{\theta y - b(\theta)}$ with respect to $\nu(y)$. Suppose, in the generalized linear model $E(Y|\beta^T X) = \mu(\beta^T X)$, the link function is the natural link $\mu^{-1} = (b')^{-1}$. Let η be the OLS vector $E(XY)$. Suppose that $E(X|\eta^T X)$ is linear in X. Then $\mathcal{S}_{Y|X} = span(\eta)$.*

PROOF. Let $f(x, y, \beta)$ denote the joint density of (X, Y), and let $R(\xi) = E \log f(X, Y, \xi)$. First, it is easy to see that β is the unique minimizer of the expected log likelihood

$$\begin{aligned} E\log f(X,Y,\xi) - E\log f(X,Y,\beta) &= E\log\left[\frac{f(X,Y,\xi)}{f(X,Y,\beta)}\right] \\ < \log E\left[\frac{f(X,Y,\xi)}{f(X,Y,\beta)}\right] &= \log(1) = 0. \end{aligned}$$

Now let ξ be any vector in $\mathbb{R}^p$ and let $P_\eta = \eta\eta^T$ be the projection matrix onto η. Because we have used the natural link function the conditional density $f_\beta(y|x)$ is of the form $e^{\beta^T x y - b(\beta^T x)}$. Hence,

$$\begin{aligned} R(\xi) &= E\left[\xi^T XY - b(\xi^T X)\right] \\ &= \xi^T E(XY) - Eb(\xi^T X) \\ &= \xi^T \eta - Eb(\xi^T X). \end{aligned}$$

Because P_η is the projection onto span(η), we have $\eta = P_\eta \eta$. Moreover, because b is convex, we have, by Jensen's inequality,

$$\begin{aligned} Eb(\xi^T X) &= E\{E[b(\xi^T)|\eta^T X]\} \\ &\geq Eb(E(\xi^T X|\eta^T X)). \end{aligned}$$

By the assumption that $E(X|\eta^T X)$ is linear in X, and Theorem 4.4, the L-2 projection and the Euclidean projection are on and the same; that is, $E(X|\eta^T X) = P_\eta(X)$. Therefore,

$$R(\xi) \leq \xi^T P_\eta \eta - Eb(\xi^T P_\eta X) = R(P_\eta \xi).$$

In other words, for any ξ in $\mathbb{R}^p$ we can always find a vector in the direction of η that increases the function R. Because β is the unique maximizer of $R(\xi)$ we therefore know that it must be in the direction of η. This means that η and β are in the same direction. □

5 Principle Hessian Directions

The biggest disadvantage of OLS is that it can only estimate at most one direction in the central space. Thus, if the central space is more than one dimensional, OLS cannot provide a comprehensive estimate. In this section we will introduce a method that can estimate more than one directions in the central space. For this purpose we need to make one more assumption.

5.1 Constant conditional variance

Assumption 5.1 *We assume that the conditional variance*

$$\text{var}(X|\beta^T X)$$

is a non-random matrix.

Note that this assumption is satisfied if X is multivariate normal. The next Lemma gives an important consequence of this assumption.

Lemma 5.1 *Let $P_\beta = \beta\beta^T$ be the projection matrix onto the column space of β. Let $Q_\beta = I_p - P_\beta$ be the projection on to the orthogonal complement of $\mathcal{S}(\beta)$. Then*

$$\text{var}(X|\beta^T X) = Q_\beta.$$

Proof. By the famous EV-VE formula:

$$I_p = \text{ var}(X) = E[\text{ var}(X|\beta^T X)] + \text{ var}[E(X|\beta^T X)] \tag{4}$$

Because $\text{var}(X|\beta^T X)$ is nonrandom, the first term on the right hand side is simply $\text{var}(Y|\beta^T X)$. By Theorem 4.4, the second term on the right hand side is

$$\begin{aligned} \text{var}[E(X|\beta^T X)] &= \text{ var}(P_\beta X) \\ &= P_\beta I_p P_\beta \\ &= P_\beta. \end{aligned}$$

Hence (4) becomes

$$I_p = \text{ var}(X|\beta^T X) + P_\beta.$$

Now subtract both sides by P_β to obtain the desired result. □

This condition is quite strong. There is a result saying that if $\text{var}(X|\gamma^T X)$ is constant for all γ then X is necessarily multivariate normal. Later on we will introduce a method that does not depend on this assumption.

5.2 pHd: population development

Let α be the OLS vector $E(XY)$. Let e be the residual from the simple linear regression; that is

$$e = Y - \alpha^T X.$$

Note that, in the standardized coordinate, the intersection of the OLS is zero, because it is $E(Y) - \alpha^T E(X)$, which is zero. That is why there is no constant term in e.

Definition 5.1 *The matrix $H_1 = E(YXX^T)$ is called the y-based Hessian matrix, the matrix $H_2 = E(eXX^T)$ is called the e-based Hessian matrix.*

The central result of this section is that the column space of a Hessian matrix (either one) is a subspace of the central space.

Theorem 5.1 *Suppose that Assumptions 4.1 and 5.1 hold. Then the column space of H_1 is a subspace of $\mathcal{S}_{Y|X}$.*

Proof. Note that

$$E(YXX^T) = E[E(Y|X)XX^T]$$

Because $Y \perp\!\!\!\perp X|\beta^T X$, we have, by Lemma 4.1 $E(Y|X) = E(Y|\beta^T X)$. Therefore, the right hand side becomes

$$E[E(Y|\beta^T X)XX^T].$$

Because conditional expectation is a self-adjoint operator, the above becomes:

$$E[YE(XX^T|\beta^T X)]. \tag{5}$$

Now let us analyze the inner conditional expectation.

$$E(XX^T|\beta^T X) = \mathrm{var}(X|\beta^T X) + E(X|\beta^T X)E(X^T|\beta^T X).$$

By Lemma 5.1, the first term on the right hand side is Q_β. The Theorem 4.4, the second term is $P_\beta XX^T P_\beta$. Thus the expectation in (5) becomes

$$\begin{aligned} E[YE(XX^T|\beta^T X)] &= E[Y(Q_\beta + P_\beta XX^T P_\beta)] \\ &= E(Y)Q_\beta + P_\beta E(YXX^T)P_\beta \\ &= P_\beta H_1 P_\beta. \end{aligned}$$

Thus we have proved $H_1 = P_\beta H_1 P_\beta$. Now the right hand side is of the form $P_\beta w$ where w is a vector in $\mathbb{R}^p$. Thus the columns of H_1 are linear combinations of the column vectors of P_β, which must be in $\mathcal{S}(\beta)$. This completes the proof. □

With a slight modification of the proof, we can show that the same conclusion holds for H_2 as well.

Theorem 5.2 *Suppose that Assumptions 4.1 and 5.1 hold. Then the columns space of H_2 is a subspace of $\mathcal{S}_{Y|X}$.*

PROOF. By inspecting the proof of Theorem 5.1, the only additional thing we need to prove is that

$$E(e|X) = E(e|\beta^T X).$$

Now

$$\begin{aligned} E(e|X) &= E(Y - \alpha^T X|X) \\ &= E(Y|X) - \alpha^T X. \end{aligned}$$

As argued before, the first term on the right hand side is $E(Y|\beta^T X)$. Because of Assumption 4.1, and applying Theorem 4.5, we know $\alpha \in \mathcal{S}_{Y|X}$. Therefore, $\alpha = P_\beta \alpha$. Therefore, the right hand side of the above expression becomes

$$E(e|X) = E(Y|\beta^T X) - \alpha^T P_\beta X.$$

However, by Assumption 4.1 and Theorem 4.4 we have $P_\beta X = E(X|\beta^T X)$. So the right hand side is

$$E(Y|\beta^T X) - \alpha^T E(X|\beta^T X) = E(e|\beta^T X),$$

as desired. □

5.3 Sample estimator of pHd

Again, we use the idea of first transforming to Z, estimating $\mathcal{S}_{Y|Z}$, and then transforming back to $\mathcal{S}_{Y|X}$. We summarize the computation into the following steps.

- First, standardize $X_1, \ldots, X_n$ to $\widehat{Z}_1, \ldots, \widehat{Z}_n$, and center $Y_1, \ldots, Y_n$ to $\widehat{Y}_1, \ldots, \widehat{Y}_n$, as described in the algorithm for OLS.
- Compute the OLS of $\widehat{Y}$ versus $\widehat{Z}$:

$$\begin{aligned} \hat{\alpha} &= \left[\mathrm{var}_n(\widehat{Z})\right]^{-1} \mathrm{cov}_n(\widehat{Z}, \widehat{Y}) \\ \hat{\alpha}_0 &= E_n(\widehat{Y}) - \hat{\alpha}^T E_n(\widehat{X}). \end{aligned}$$

  Much simplification can be achieved:

$$\begin{aligned} \mathrm{var}_n(\widehat{Z}) &= I_p \\ \mathrm{cov}_n(\widehat{X}, \widehat{Y}) &= E_n(\widehat{X}\widehat{Y}) - E_n(\widehat{X})E_n(\widehat{Y}) = E_n(\widehat{X}\widehat{Y}). \end{aligned}$$

  Also, because $E_n(\widehat{Y}) = 0$, $E_n(\widehat{X}) = 0$, we have $\hat{\alpha}_0 = 0$. So the OLS for $\widehat{Z}$ and $\widehat{Y}$ is simply $E_n(\widehat{Z}\widehat{Y})$. Compute the sample residual

$$\hat{e}_i = \widehat{Y}_i - \hat{\alpha}^T \widehat{X}_i.$$

- Construct the e-based and y-based Hessian matrix:

$$\widehat{H}_1 = E_n\left(\widehat{Y}\widehat{Z}\widehat{Z}^T\right) \quad \widehat{H}_2 = E_n\left(\widehat{e}\widehat{Z}\widehat{Z}^T\right)$$

- Assume, for now, we know the structural dimension q. Let $\hat{\gamma}_1, \ldots, \hat{\gamma}_q$ be the q eigenvectors corresponding to the q largest eigenvalues of $\widehat{H}_1\widehat{H}_1^T$; and let $\hat{\delta}_1, \ldots, \hat{\delta}_q$ be the q eigenvectors corresponding to the largest eigenvalues of $\widehat{H}_2\widehat{H}_2^T$. We use $\hat{\gamma}_1, \ldots, \hat{\gamma}_q$ or $\hat{\delta}_1, \ldots, \hat{\delta}_q$ as the estimator of $\mathcal{S}_{Y|Z}$.

- Let

$$\hat{\beta}_i = \widehat{\Sigma}^{-1/2}\hat{\gamma}_i$$
$$\hat{\eta}_i = \widehat{\Sigma}^{-1/2}\hat{\delta}_i$$

  We will use $\{\hat{\beta}_1, \ldots, \hat{\beta}_q\}$ or $\{\hat{\eta}_1, \ldots, \hat{\eta}_q\}$ as the estimator of $\mathcal{S}_{Y|X}$.

For now we have assumed that the structural dimension is known. In practice this must be determined by the data. Later on (if time permits) we will introduce a test that will help to determine the structural dimension.

Because this method is based on the eigenvectors of the Hessian matrix corresponding to their largest eigenvalues, or the principle directions of the Hessian matrices, we call this method the principle Hessian directions, or pHd.

5.4 Convergence rate of pHd estimators

We will only analyze the y-based pHd; the analysis for the e-based pHd is completely parallel.

First, we show that $\hat{\gamma}_1, \ldots, \hat{\gamma}_q$ converge at $\sqrt{n}$-rate to $\gamma_1, \ldots, \gamma_q$, the eigenvectors of $H_1H_1^T$ corresponding to its nonzero eigenvalues. For this it suffices to show that $\widehat{H}_1\widehat{H}_1^T$ converges at $\sqrt{n}$-rate to $H_1H_1^T$. Note that matrix $E_n(\widehat{Y}\widehat{Z}\widehat{Z}^T)$ is

$$\widehat{\Sigma}^{-1/2}E_n(Y - E_nY)(X - E_nX)(X - E_nX)^T\widehat{\Sigma}^{-1/2}.$$

The central part can be decomposed into eight terms:

$$\begin{aligned}
&E_n(YXX^T) - E_n(YX)E_n(X^T)\\
&-E_n(YX^T)E_n(X) + E_n(Y)E_n(X)E_n(X^T)\\
&+E_n(Y)E_n(XX^T) + E_n(Y)E_n(X)E_n(X^T)\\
&+E_n(Y)E_n(X)E_n(X^T) + E_n(Y)E_n(X)E_n(X^T)\\
&= E_n(XX^TY) + O_p(n^{-1/2}).
\end{aligned}$$

As we have seen before, $\widehat{\Sigma} = \Sigma + O_p(n^{-1/2}) = I + (n^{-1/2})$, and so

$$\begin{aligned}
E_n(\widehat{Y}\widehat{Z}\widehat{Z}^T) &= (I + O_p(n^{-1/2}))(E_n(XX^TY) + O_p(n^{-1/2}))\\
&\quad (I + O_p(n^{-1/2}))\\
&= E_n(ZZ^TY) + O_p(n^{-1/2}).
\end{aligned}$$

Thus we have shown that $\hat{\gamma}_1, \ldots, \hat{\gamma}_q$ converges at $\sqrt{n}$-rate to $\gamma_1, \ldots, \gamma_q$. Now

$$\begin{aligned}\hat{\beta}_i &= \widehat{\Sigma}^{-1/2}\hat{\gamma}_i \\ &= (I + O_p(n^{-1/2}))(\gamma_i + O_p(n^{-1/2})) \\ &= \gamma_i + O_p(n^{-1/2}) = \beta_i + O_p(n^{-1/2}),\end{aligned}$$

as desired.

Theorem 5.3 *If Assumptions 4.1 and 5.1 hold, then $\hat{\beta}_1, \ldots, \hat{\beta}_q$ converges at $\sqrt{n}$-rate to a set of vectors that belong to $\mathcal{S}_{Y|X}$.*

5.5 Determine q

5.5.1 Formulation of hypothesis

In this section we discuss how to use hypothesis test to determine the structural dimension q. We will derive a test statistic based on $\widehat{H}_2$ because the asymptotic structure of $\widehat{H}_2$ is much simpler than that of $\widehat{H}_1$.

We estimate the rank of H_2, which we assume to be equal to the dimension of $\mathcal{S}_{Y|X}$, by conducting a series of hypothesis tests. Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ be the eigenvalues of $H_2{H_2}^T$, and consider the sequence of tests

$$H_0 : \lambda_{j+1} = ... = \lambda_p = 0, \quad j = 0, 1, \cdots, p-1.$$

The rank q of H_2 is the smallest value of j for which this hypothesis holds. Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p$ be the eigenvalues of $n\widehat{H_2}{\widehat{H_2}}^T$.

We test H_0 using the statistic

$$T_j = C^{-1} \sum_{i=j+1}^{p} \hat{\lambda}_i$$

where C is a positive constant that depends on j and will be determined later. Relatively large values of T_j provide evidence against H_0. Tests of H_0 are used to estimate the rank q of B as follows: Beginning with $j = 0$ test H_0. If the hypothesis is rejected, increment j by one and test again, stopping with the first nonsignificant result. The corresponding value of j is the estimate $\hat{q}$ of q.

5.5.2 Analysis of $\widehat{\Sigma}^{-1/2}$

As a warm up, we first note a simple fact that we have used several times, and will be used repeatedly.

Lemma 5.2 *If $W_1, \ldots, W_n$ are iid $E(W) = 0$ and $var(W) < \infty$. Then $E_n(W) = O_p(n^{-1/2})$.*

PROOF. By Lindeberge Levy central limit theorem,

$$\sqrt{n}(E_n(W) - E(W)) = \sqrt{n}E_n(W) \xrightarrow{\mathcal{L}} N(0, E(W^2)).$$

Therefore $\sqrt{n}E_n(W)$ is bounded in probability. □

Lemma 5.3

$$\widehat{\Sigma}^{-1/2} = I - E_n(ZZ^T - I)/2 + O_p(n^{-1}). \tag{6}$$

PROOF. Note that

$$\begin{aligned}\widehat{\Sigma} &= E_n(ZZ^T) - \bar{Z}\,\bar{Z}^T \\ &= E_n(ZZ^T) + O_p(n^{-1}) \\ &= I + E_n(ZZ^T - I) + O_p(n^{-1}),\end{aligned}$$

where $E_n(ZZ^T - I)$ is of the order $O_p(n^{-1/2})$. We know that $\widehat{\Sigma}^{-1/2}$ must be of the form $I + A_n$ for some random matrix A_n of the order $O_p(n^{-1/2})$. Therefore

$$(I + A_n)^2(I + E_n(ZZ^T - I)) = I.$$

The left hand is

$$I + E_n(ZZ^T - I) + 2A_n + O_p(n^{-1}).$$

Therefore $A_n = -E_n(ZZ^T - I)/2 + O_p(n^{-1})$, as desired. □

5.5.3 Expansion of $E_n[\hat{e}(X - \bar{X})(X - \bar{X})^T]$

In this section, we will for simplicity write $\widehat{H}_2$ as $\widehat{H}$ and H_2 as H.

Lemma 5.4 *Suppose that X has a standardized multivariate normal distribution $N(0, I_p)$. Then*

$$E_n\hat{e}(X - \bar{X})(X - \bar{X})^T = H + E_n\left[e(XX^T - I) - H\right] + O_p(n^{-1}).$$

PROOF. We have

$$\begin{aligned}E_n\left[\hat{e}(X - \bar{X})(X - \bar{X})^T\right] = {} & E_n(\hat{e}XX^T) - \bar{X}E_n(\hat{e}X^T) \\ & -E_n(\hat{e}X)\bar{X}^T + O_p(n^{-1}).\end{aligned}$$

Because $\bar{X} = O_p(n^{-1/2})$, we need only expand $E_n(\hat{e}X)$ so that the error is of the order $O_p(n^{-1/2})$. Note that

$$\begin{aligned}E_n(\hat{e}X) &= E_n\left[\left((Y - \bar{Y}) - \hat{\beta}^T\widehat{\Sigma}^{-1/2}(X - \bar{X})\right)X\right] \\ &= E_n\left[(Y - \bar{Y})X\right] - E_n\left[X(X - \bar{X})^T\right]\widehat{\Sigma}^{-1/2}\hat{\beta}.\end{aligned}$$

It is easy to see that

$$\begin{aligned} E_n\left[(Y-\bar{Y})X\right] &= \beta + O_p(n^{-1/2}), \\ E_n\left[X(X-\bar{X})^T\right] &= I + O_p(n^{-1/2}), \\ \widehat{\Sigma}^{-1/2} &= I + O_p(n^{-1/2}), \\ \hat{\beta} &= \beta + O_p(n^{-1/2}). \end{aligned}$$

Therefore,

$$E_n(\hat{e}X) = \beta - \beta + O_p(n^{-1/2}) = O_p(n^{-1/2}).$$

And consequently,

$$E_n\hat{e}(X-\bar{X})(X-\bar{X})^T = E_n\hat{e}XX^T + O_p(n^{-1}). \tag{7}$$

We now expand the right hand side so that the error is of the order $O_p(n^{-1})$. We have

$$\begin{aligned} E_n\hat{e}XX^T = {} & E_n\left[(Y-\bar{Y})XX^T\right] \\ & -E_n\left[\hat{\beta}^T\widehat{\Sigma}^{-1/2}(X-\bar{X})(XX^T)\right]. \end{aligned} \tag{8}$$

The first term on the right hand side is

$$\begin{aligned} E_n\left[(Y-\bar{Y})XX^T\right] &= E_n\left[(Y-\bar{Y})(XX^T-I)\right] \\ &= E_n\left[Y(XX^T-I)\right] + O_p(n^{-1}). \end{aligned} \tag{9}$$

The second term on the right hand side of (8) is expanded as

$$\begin{aligned} & E_n\left[\hat{\beta}^T\widehat{\Sigma}^{-1/2}(X-\bar{X})(XX^T)\right] \\ &= E_n\left[\hat{\beta}^T\widehat{\Sigma}^{-1/2}(X-\bar{X})(XX^T-I)\right] \\ &= E_n\left[\hat{\beta}^T\widehat{\Sigma}^{-1/2}X(XX^T-I)\right] + O_p(n^{-1}). \end{aligned}$$

The (i,j)th element of the $p \times p$ matrix on the right hand side is

$$\sum_{k=1}^{p}\left(\widehat{\Sigma}^{-1/2}\hat{\beta}\right)_k E_n\left[X_k(X_iX_j-\delta_{ij})\right],$$

where $(\widehat{\Sigma}^{-1/2}\hat{\beta})_k$ is the k element of the vector $\widehat{\Sigma}^{-1/2}\hat{\beta}$ and δ_{ij} is the (i,j)th element of the p-dimensional identity matrix I. Because X has a standard multivariate normal distribution, the expectation of $X_k(X_iX_j-\delta_{ij})$ is zero for any i,j,k. Therefore $E_n(X_k(X_iX_j-\delta_{ij})) = O_p(n^{-1/2})$, and hence if we replace the $\widehat{\Sigma}$ and $\hat{\beta}$ by I and β then the the error incurred has the magnitude $O_p(n^{-1})$. It follows then that

$$\begin{aligned} & E_n\left[\hat{\beta}^T\widehat{\Sigma}^{-1/2}(X-\bar{X})(XX^T)\right] \\ &= E_n\left[\beta^TX(XX^T-I)\right] + O_p(n^{-1}). \end{aligned} \tag{10}$$

Now substitute (9) and (10) into (8) to obtain

$$E_n(\hat{e}XX^T) = E_n\left[e(XX^T - I)\right] + O_p(n^{-1}).$$

However, note that

$$E\left[e(XX^T - I)\right] = E(eXX^T) = H.$$

Hence,

$$E_n(\hat{e}XX^T) = H + E_n\left[e(XX^T - I) - H\right] + O_p(n^{-1}),$$

which, combined with (7), implies that

$$E_n\hat{e}(X - \bar{X})(X - \bar{X})^T = H + E_n\left[e(XX^T - I) - H\right] + O_p(n^{-1}), \tag{11}$$

as desired. □

5.5.4 Expansion of $\widehat{H}_2$

We will continue to write H_2 as H and $\widehat{H}_2$ as $\widehat{H}$. From the previous two subsections we have shown that

$$\begin{aligned}\widehat{\Sigma}^{-1/2} &= I + A_n + O_p(n^{-1})\\ E_n\hat{e}(X - \bar{X})(X - \bar{X})^T &= H + B_n + O_p(n^{-1}),\end{aligned}$$

where

$$\begin{aligned}A_n &= -E_n(ZZ^T - I)/2 + O_p(n^{-1})\\ B_n &= E_n\left[e(XX^T - I) - H\right] + O_p(n^{-1}).\end{aligned}$$

Hence

$$\begin{aligned}\widehat{H} &= (I + A_n + O_p(n^{-1}))(H + B_n + O_p(n^{-1}))\\ &\quad (I + A_n + O_p(n^{-1}))\\ &= H + A_nH + B_n + HA_n + O_p(n^{-1}).\end{aligned}$$

Substituting the definitions of A_n, B_n into the above expression to obtain the following lemma.

Lemma 5.5 *Suppose that X has a standard multivariate normal distribution. Then,*

$$\begin{aligned}\widehat{H} &= H + E_n\{e(XX^T - I_p) - H\}\\ &\quad -\frac{1}{2}E_n(XX^T - I_p)H - \frac{1}{2}HE_n(XX^T - I_p) + O_p(n^{-1}).\end{aligned}$$

5.6 Asymptotic distribution of T_j

5.6.1 Eaton and Tyler's result

Reference: Eaton, M. L. and Tyler, D. E. (1994). The Asymptotic Distribution of Singular Values with Applications to Canonical Correlations and Correspondence Analysis. *Journal of Multivariate Analysis* **34** 439–446.

Suppose $\widehat{B}$ and B are symmetric square matrix, and that $\sqrt{n}(\widehat{B} - B)$ is bounded in probability. Suppose that B has the spectrum decomposition

$$B = (\Psi_1, \Psi_0) \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \Psi_1^T \\ \Psi_0^T \end{pmatrix}, \tag{12}$$

where $\Psi_1 \in R^{p \times j}$ is the matrix whose columns are eigenvectors B corresponding to its non-zero eigenvalues, $\Psi_0 \in R^{p \times (p-j)}$ is the matrix whose columns are the eigenvectors of B corresponding to its zero eigenvalues, $D \in R^{j \times j}$ the diagonal matrix with diagonal elements equal to the nonzero eigenvalues of B. It is easy to see that Ψ_0 satisfies the relation:

$$\Psi_0 B = 0, \quad B\Psi_0 = 0. \tag{13}$$

It follows from Eaton and Tyler (1994) that the joint asymptotic distribution of the $p-j$ smallest (in absolute value) eigenvalues of the matrix $\sqrt{n}\widehat{B}$ is the same as that of the smallest eigenvalues of the matrix $\sqrt{n}\Psi_0^T(\widehat{B} - B)\Psi_0$.

Applying to our situation, we are interested in the distribution of $\sum_{i=j+1}^{p} \hat{\lambda}_i$, the sum of the smallest eigenvalues of $\widehat{H}\widehat{H}$. We will introduce the vec notation: if A is a matrix with columns $a_1, ...a_p$, then $\text{vec}(A) = (a_1^T ... a_p^T)^T$. Regarding vec, the next lemma will be useful:

Lemma 5.6 *The sum of the eigenvalues of AA^T equals to*

$$vec(A)^T vec(A)$$

.

Proof. The sum of the eigenvalues of AA^T is the same as the trace of AA^T, which is

$$\sum_{i=1}^{p} \sum_{k=1}^{p} A_{ik} A_{ik},$$

which is the same as $\text{vec}(A)^T \text{vec}(A)$. □

Now the eigenvalues of

$$\sqrt{n}\Psi_0^T(\widehat{H} - H)\Psi_0$$

is the smallest eigenvalues (in absolute values) of $\sqrt{n}\widehat{H}$. So the eigenvalues of

$$\left[\sqrt{n}\Psi_0^T(\widehat{H} - H)\Psi_0\right] \left[\sqrt{n}\Psi_0^T(\widehat{H} - H)\Psi_0\right]^T \tag{14}$$

are the smallest eigenvalues of $n\widehat{H}\widehat{H}$. So the distribution of $\sum_{i=j+1}^{p}\hat{\lambda}_i$ is the same as that of the sum of the eigenvalues of (14), which by the above lemma is the same as

$$\text{vec}^T\left[\sqrt{n}\Psi_0^T(\widehat{H}-H)\Psi_0\right]\text{vec}\left[\sqrt{n}\Psi_0^T(\widehat{H}-H)\Psi_0\right].$$

So it all boils down to the deriving the asymptotic distribution of

$$\text{vec}\left[\sqrt{n}\Psi_0^T(\widehat{H}-H)\Psi_0\right].$$

We will write the matrix $\sqrt{n}\Psi_0^T(\widehat{H}-H)\Psi_0$ as U.

5.6.2 Tensor product between matrices

Let A and B be two matrix of arbitrary order, the tensor product between them, denoted by $A\otimes B$, is defined as

$$\begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1p}B \\ a_{21}B & a_{22}B & \cdots & a_{2p}B \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mp}B \end{pmatrix}.$$

A simple fact about tensor product and vec.

Lemma 5.7 *Suppose A, B, C are matrices of orders $p_1\times p_2$, $p_2\times p_3$ and $p_3\times p_4$. Then*

$$vec(ABC) = (C^T\otimes A)vec(B).$$

The proof is a straightforward calculation, and is omitted.

Using tensor product we can express explicitly the variance matrix of the second moment of a multivariate normal vector. Let X have standard multivariate normal distribution. Then,

$$\text{var}(X\otimes X) = I_p\otimes I_p + C(e_1,\cdots,e_p),$$

where I_p is a p by p identity matrix and $e_1,\cdots,e_p$ are the standard orthonormal basis in R^p — for example, e_i is a p-dimensional vector with its ith element equal to 1 and its other elements equal to 0, and $C(e_1,\ldots,e_p)$ is the matrix

$$C(e_1,...,e_p) = \begin{pmatrix} e_1e_1^T & \cdots & e_pe_1^T \\ & \cdots & \\ e_1e_p^T & \cdots & e_pe_p^T \end{pmatrix}. \tag{15}$$

Regarding $\text{var}(X\otimes X)$, we have the following lemma.

Lemma 5.8 *The matrix*

$$D = \{I_p\otimes I_p + C(e_1,\cdots,e_p)\}/2$$

is idempotent, and has rank $p(p+1)/2$.

PROOF. It is easy to see that

$$D^2 = (1/4)\{I_p \otimes I_p + 2C(e_1, \cdots, e_p) + C^2(e_1, \cdots, c_p)\}.$$

Thus D will be idempotent if $C^2(e_1, \cdots, e_p) = I_p \otimes I_p$. This is indeed the case because the (r, s) block of $C^2(e_1, \cdots, c_p)$ is

$$\sum_{i=1}^{p} C_{ri}(e_1, \cdots, e_p) C_{is}(e_1, \cdots, e_p) = \sum_{i=1}^{p} e_i e_r^T e_s e_i^T = e_r^T e_s I_p,$$

which is the (r, s)th block of the matrix $I_p \otimes I_p$.

Now recall that the rank of a projection matrix equals to its trace. Thus the rank of D is

$$(1/2)\{\mathrm{tr}(I_p \otimes I_p) + \mathrm{tr}(C(e_1, \cdots, e_p))\} = (1/2)(p^2 + p) = p(p+1)/2,$$

which completes the proof. □

From this it is easy to derive the following corollary.

Corollary 5.1 *Suppose Q is a projection matrix of rank k. Suppose X has p-dimensional standard multivariate normal distribution. Then, the matrix*

$$var(QX \otimes QX)/2$$

is idempotent and has rank $k(k+1)/2$.

PROOF. Let γ be a $p \times k$ matrix whose columns form an orthonormal bases in Q. The $Q = \gamma\gamma^T$. Hence

$$\mathrm{var}(QX \otimes QX)/2 = (\gamma \otimes \gamma)\left[\mathrm{var}(\gamma^T X \otimes \gamma^T X)/2\right](\gamma^T \otimes \gamma^T).$$

Now $\gamma^T X$ is a k-dimensional standard multivariate normal vector, and hence, by the above lemma, $\mathrm{var}(\gamma^T X \times \gamma^T X)$ is idempotent. Therefore,

$$\begin{aligned}
&[\mathrm{var}(QX \otimes QX)/2]\,[\mathrm{var}(QX \otimes QX)/2] \\
&= (\gamma \otimes \gamma)\left[\mathrm{var}(\gamma^T X \otimes \gamma^T X)/2\right](\gamma^T \otimes \gamma^T) \\
&\quad (\gamma \otimes \gamma)\left[\mathrm{var}(\gamma^T X \otimes \gamma^T X)/2\right](\gamma^T \otimes \gamma^T) \\
&= (\gamma \otimes \gamma)\left[\mathrm{var}(\gamma^T X \otimes \gamma^T X)/2\right]^2(\gamma^T \otimes \gamma^T) \\
&= (\gamma \otimes \gamma)\left[\mathrm{var}(\gamma^T X \otimes \gamma^T X)/2\right](\gamma^T \otimes \gamma^T) \\
&= \mathrm{var}(QX \otimes QX)/2
\end{aligned}$$

Now because $(\gamma^T \otimes \gamma^T)(\gamma \otimes \gamma) = I_k \otimes I_k$, its rank is k^2. Therefore the rank of $\mathrm{var}(\gamma^T X \times \gamma^T X)$ is $\min(k^2, k(k+1)/2, k^2)$, which is $k(k+1)/2$. □

5.6.3 Asymptotic distribution of vec(U).

From Lemma 5.5, we have

$$\sqrt{n}(\widehat{H} - H) = \sqrt{n}E_n\{e(XX^T - I_p) - H\} - \frac{1}{2}\sqrt{n}E_n(XX^T - I_p)H$$
$$-\frac{1}{2}\sqrt{n}HE_n(XX^T - I_p) + O_p(n^{-\frac{1}{2}}).$$

(We see now why we needed to expand to the order $O_p(n^{-1})$). The second and the third terms, as well as the term H in the first term, vanish once we multiply from the left by Ψ_0^T and from the right by Ψ_0. And we are left with

$$\begin{aligned}\sqrt{n}\Psi_0^T(\widehat{H} - H)\Psi_0 &= \sqrt{n}\Psi_0^T E_n\{e(XX^T - I_p)\}\Psi_0 + O_p(n^{-\frac{1}{2}}) \\ &\equiv \sqrt{n}\Psi_0^T E_n(W)\Psi_0 + O_p(n^{-\frac{1}{2}}),\end{aligned}$$

where $W = e(XX^T - I_p)$.

Note that $\Psi_0^T E(W)\Psi_0 = 0$. Therefore,

$$E\text{vec}\left(\Psi_0^T E(W)\Psi_0\right) = 0.$$

Also note that

$$\begin{aligned}\Psi_0^T W\Psi_0 &= e\left(\Psi_0^T XX^T\Psi_0 - \Psi_0^T\Psi_0\right) \\ &= e\left(\Psi_0^T XX^T\Psi_0 - I_{p-j}\right)\end{aligned}$$

Therefore,

$$\text{vec}\left(\Psi_0^T W\Psi_0\right) = e\left(\Psi_0^T X \otimes \Psi_0^T X - \text{vec}(I_{p-j})\right).$$

Now

$$\begin{aligned}&\text{var}\left[\text{vec}\left(\Psi_0^T W\Psi_0\right)\right] \\ &= E\left[e^2\left(\Psi_0^T X \otimes \Psi_0^T X - \text{vec}(I_{p-j})\right)\left(\Psi_0^T X \otimes \Psi_0^T X - \text{vec}(I_{p-j})\right)^T\right] \\ &\equiv E(e^2 VV^T).\end{aligned}$$

Suppose that pHd exhausts the central space $\mathcal{S}_{Y|X}$. Then, $\Psi_0\Psi_0^T = Q$, where Q is the orthogonal projection onto the orthogonal complement of the central space. Now the right hand side is

$$E(e^2 VV^T) = E(E(e^2|X)VV^T).$$

We claim that

$$E(e^2|X) = E(e^2|\beta^T X) \tag{16}$$

This is because

$$\begin{aligned} E(e^2|X) &= E(Y^2 - 2(\alpha^T X)Y + (\alpha^T X)^2|X) \\ &= E(Y^2|X) - 2(\alpha^T X)E(Y|X) + (\alpha^T X)^2. \end{aligned}$$

As we have shown before, $Y \perp\!\!\!\perp X|\beta^T X$ implies that

$$E(Y|X) = E(Y|\beta^T X) \quad E(Y^2|X) = E(Y^2|\beta^T X).$$

Therefore,

$$(\alpha^T X)E(Y|X) = (\alpha^T X)E(Y|\beta^T X) = E((\alpha^T X)Y|\beta^T X).$$

Also,

$$E((\alpha^T X)^2|X) = (\alpha^T X)^2 = E((\alpha^T X)^2|\beta^T X).$$

Thus we have shown (16). Now

$$\begin{aligned} E(e^2 VV) &= E(E(e^2|X)VV^T) \\ &= E[E(e^2|\beta^T X)VV^T]. \end{aligned}$$

Because V is a function of $\Psi_0^T X$, which is independent of $\beta^T X$, we have $E(e^2|\beta^T X) \perp\!\!\!\perp V$, and therefore,

$$E(e^2 VV) = E(E(e^2|\beta^T X))E(VV^T) = \mathrm{var}(e)E(VV^T).$$

However, it is easy to see that

$$E(VV^T) = \mathrm{var}(\Psi_0^T X)).$$

Therefore we have

$$\mathrm{vec}(U) \xrightarrow{\mathcal{L}} N(0, \mathrm{var}(e)\mathrm{var}(\Psi_0^T X)).$$

And consequently,

$$\mathrm{vec}(U)/\sqrt{2\mathrm{var}(e)} \xrightarrow{\mathcal{L}} N(0, \mathrm{var}(\Psi_0^T X)/2). \tag{17}$$

Note that $\Psi_0^T X$ has a $p - j$ dimensional standard multivariate normal distribution. Therefore, by Lemma 5.8, $\mathrm{var}(\Psi_0^T X)/2$ is idempotent and is of rank $(p-j)(p-j+1)/2$. Now we need the following lemma:

Lemma 5.9 *If X is distributed as multivariate normal with mean 0 and variance matrix A, where A is an idempotent matrix. Then $X^T X$ is distributed as χ^2 with $rank(A)$ degrees of freedom.*

PROOF. Because A is symmetric and idempotent, it is a projection matrix. Let $a_(a_1, \ldots, a_k)$ be an orthonormal basis of the ran(A). Then, $A = aa^T$. Therefore, $a^T X$ is distributed as $N(0, a^T aa^T a) = N(0, I_k)$. Consequently $(a^T X)^T (a^T X)$ is distributed as $\chi^2_{(k)}$. Let (a, b) be an orthonormal basis in $\mathbb{R}^p$. Then

$$X^T X = X^T I_p X = X^T (aa^T + bb^T) X.$$

But $b^T X$ is distributed as $N(0, 0)$; in other words, $b^T X \equiv 0$. Therefore,

$$X^T X = X^T aa^T X = (a^T X)^T (a^T).$$

Therefore, $X^T X$ is distributed as $\chi^2_{(k)}$, as desired. □

Now combine (17) and the above lemma, we know that

$$\frac{1}{2\text{var(e)}} \text{vec}^T(U)\text{vec}(U) \xrightarrow{\mathcal{L}} \chi^2_{(p-j)(p-j+1)/2}.$$

Hence we have the following theorem.

Theorem 5.4 *Suppose that*

(a) The column space of H exhausts the Central Space; that is

$$span\{H\} = S_{Y|X}.$$

(b) The predictor X has a p-dimensional standard multivariate normal distribution.

Then, under the null hypothesis

$$H_0 : \lambda_{p-j+1} = \cdots = \lambda_p = 0,$$

the test statistic $\sum_{i=p-j+1}^{p} \hat{\lambda}_i / (2\text{var}(e))$ converges in distribution to a χ^2 distribution with $(p-j)(p-j+1)/2$ degrees of freedom.

5.6.4 Examples

Example 5.1 Let us now apply pHd to the bigmac problem. First make a transformation of the predictors — using Box-Cox transformation. The regress Big-mac onto the rest of the variables using pHd. One significant vector is obtained. Show the spin plot here.

Example 5.2 Apply pHd to the ozone data. Again, remove the variables: windspeed, visibility, and InversionHt. Use log transformation on X. Again, one significant vector is found. Show the spin plot. In the first direction the spin plot shows a v-shaped pattern.

6 Sliced Inverse Regression

Reference: K.C. Li (1991), JASA.

6.1 Population development

The central result of this section is that, under the linear conditional mean assumption (Assumption 4.1), the inverse regression vector $E(X|Y = y)$ belongs to the central subspace.

Theorem 6.1 *Suppose that Assumption 4.1 holds, and that* $E(X) = 0$ *and* $var(X) = I_p$. *Then, for any* y,

$$E(X|Y = y) \in \mathcal{S}_{Y|X}.$$

PROOF. Let β be the $p \times q$ matrix whose columns form an orthonormal basis in $\mathcal{S}_{Y|X}$. Then

$$E(X|Y) = E(E(X|Y, \beta^T X)|X).$$

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

$$E(E(X|Y, \beta^T X)|X) = E(E(X|\beta^T X)|X).$$

However, by Assumption 4.1 and Theorem 4.4, the L-2 projection $E(X|\beta^T X)$ is the same as the Euclidean projection $P_\beta X$. Therefore,

$$E(X|Y) = P_\beta E(X|Y).$$

In other words $E(X|Y)$ belongs to $\text{ran}(P)$, which is the central space. □

Corollary 6.1 *Suppose that Assumption 4.1 holds, and that* $E(X) = 0$ *and* $var(X) = I_p$. *Then the column space of the matrix*

$$\text{cov}(E(X|Y))$$

is a subspace of the central space.

PROOF. Note that

$$\begin{aligned} \text{cov}(E(X|Y)) &= \text{cov}(P_\beta E(X|Y)) \\ &= P_\beta \, \text{cov}(E(X|Y)) P_\beta, \end{aligned}$$

which completes the proof. □

In practice we will use the discretized version of the above results. Let $I_1, \ldots, I_k$ be a k intervals that partitions $\mathcal{Y}$, the space of Y. And let $\tilde{Y}$ be the discretized Y, defined by

$$\tilde{Y} = i \;\text{ if }\; Y \in I_i, \quad i = 1, \ldots, k.$$

We have the following result.

Theorem 6.2 *Suppose that Assumption 4.1 holds, and that $E(X) = 0$ and $var(X) = I_p$. Then, for any $i = 1, \ldots, k$,*

$$E(X|\tilde{Y} = i) \in \mathcal{S}_{Y|X}.$$

Consequently, the column space of the matrix

$$\mathrm{cov}(E(X|\tilde{Y}))$$

is a subspace of the central space.

PROOF. Note that $Y \perp\!\!\!\perp X|\beta^T X$ implies that $\tilde{Y} \perp\!\!\!\perp X|\beta^T X$. The rest of the proof is the same as the above. □

6.2 Sample estimator

We will use the sample version of the matrix

$$\mathrm{var}(E(X|\tilde{Y}))$$

to estimate the central space. This matrix can be written as

$$\sum_{i=1}^{k} \Pr(\tilde{Y} = i) E(X|\tilde{Y} = i) E(X^T|\tilde{Y} = i).$$

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be the sample and we have the following algorithm.

- Standardize X_i to $\widehat{Z}_i$ and centerize Y_i to $\widehat{Y}_i$ as before.
- Partition the interval $[\min\{Y_1, \ldots, Y_n\}, \max\{Y_1, \ldots, Y_n\}]$ into k intervals, say $I_1, \ldots, I_k$, and compute the mean of $\widehat{Z}$ within each slice; that is

$$\hat{\mu}_i = \frac{1}{\#(I_i)} \sum_{j \in I_i} \widehat{Z}_j.$$

- Construct the SIR matrix

$$\widehat{S} = \sum_{i=1}^{k} \frac{\#(I_i)}{n} \hat{\mu}_i \hat{\mu}_i^T.$$

- Assuming q is known. Let $v_1, \ldots v_q$ be the eigenvectors of S corresponding to the q-largest eigenvalues. This is used to estimate $\mathcal{S}_{Y|Z}$.
- Let $w_i = \widehat{\Sigma}^{-\frac{1}{2}} v_i$, $i = 1, \ldots, q$. These will be used as the estimator of $\mathcal{S}_{Y|X}$.

We call this procedure the Sliced Inverse Regression, and abbreviate it by SIR. That the SIR estimator converges at the $\sqrt{n}$-rate to a set of vectors in $\mathcal{S}_{Y|X}$ can be established similarly, and we omit it.

6.3 Determining q

The idea here is the same as the pHd case. ie we will decide q as the smallest j for which we reject the hypothesis

$$H_0 : \lambda_{p-j+1} = \cdots = \lambda_p = 0.$$

For $j = 0, \ldots, p-1$, let

$$T_j = \sum_{i=j+1}^{p} \hat{\lambda}_i,$$

where $\lambda_1 \geq \cdots \lambda_p$ are the eigenvalues of the matrix $n\widehat{S}$. We state the following theorem without proof.

Theorem 6.3 *Suppose that*

(a) $k > j+1$ *and* $p > j$.

(b) The column space of S *exhausts the Central Space; that is*

$$span\{S\} = S_{Y|X}.$$

(c) Assumptions 4.1 and 5.1 hold.

Then, under the null hypothesis

$$H_0 : \lambda_{p-j+1} = \cdots = \lambda_p = 0,$$

the test statistic $\sum_{i=p-j+1}^{p} \hat{\lambda}_i$ *converges in distribution to a* χ^2 *distribution with* $(p-j)(k-j-1)/2$ *degrees of freedom.*

The proof is in the same spirit as that for pHd, and will be omitted.

6.3.1 Examples

Example 6.1 Apply SIR to Bigmac. We find one significant vector. The sufficient plot shows a stronger pattern than does pHd. As we will see later, pHd works the best when there is a strong quadratic trend. Whereas SIR works the best when there is a monotone trend.

Example 6.2 Apply SIR to the ozone data. First make log transformation. Again, remove the three variables. One significant predictor found. The sufficient plot shows a stronger patter than does pHd.

7 Sliced Average Variance Estimator (SAVE)

Reference: Cook & Weisberg (1991, JASA).

7.1 Population development

SAVE is another method of estimating the central space based on slicing the response Y. Instead of calculating the mean within each slice, this time we compute the variance. SAVE will need both the linear conditional mean and the constant conditional variance assumption.

Theorem 7.1 *Suppose that Assumptions 4.1 and 5.1 hold and that* $E(X) = 0$ *and* $\text{var}(X) = I_p$. *Then, for any value of* y, *the column space of the matrix*

$$I_p - \text{var}(X|Y = y)$$

is a subspace of the central space. Consequently, the column space of the matrix

$$E\left[I_p - \text{var}(X|Y)\right]^2$$

is a subspace of $\mathcal{S}_{Y|X}$.

Proof. By the EV-VE formular,

$$\text{var}(X|Y) = E\left[\text{var}(X|Y, \beta^T X)|Y\right] + \text{var}\left[E(X|Y, \beta^T X)|Y\right].$$

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

$$E(X|Y, \beta^T X) = E(X|\beta^T X), \quad \text{var}(X|Y, \beta^T X) = \text{var}(X|\beta^T X)$$

However, by the linear conditional mean assumption, $E(X|\beta^T X) = P_\beta X$. And by the constant conditional variance assumption,

$$\text{var}(X|Y, \beta^T X) = Q_\beta$$

. Therefore we have

$$\begin{aligned}\text{var}(X|Y) &= Q_\beta + P_\beta \text{var}(X|Y) P_\beta \\ &= I_p - P_\beta + P_\beta \text{var}(X|Y) P_\beta.\end{aligned}$$

Subtract both sides by I_p, using the fact that P_β is idempotent, to obtain

$$\begin{aligned}I_p - \text{var}(X|Y) &= P_\beta \text{var}(X|Y) - P_\beta \\ &= P_\beta \text{var}(X|Y) - P_\beta I_p P_\beta \\ &= P_\beta \left[\text{var}(X|Y) - I_p\right] P_\beta.\end{aligned}$$

This proves that the columns of $\text{var}(X|Y) - I_p$ belong to the central space, as desire.

Now we have

$$\begin{aligned}&E\left[\text{var}(X|Y) - I_p\right]^2 \\ &= P_\beta E\left[\text{var}(X|Y) - I_p\right] P_\beta E\left[\text{var}(X|Y) - I_p\right] P_\beta.\end{aligned}$$

It is easy to see that the column space of the right hand side is contained in the central space. □

Again, in practice we use the discretized version of this. Let I_i, $i = 1, \ldots, k$ be k intervals that partitions $\mathcal{Y}$. Let $\tilde{Y}$ be as defined in the SIR section. We have

Theorem 7.2 *Suppose that Assumptions 4.1 and 5.1 hold and that* $E(X) = 0$ *and* $\mathrm{var}(X) = I_p$. *Then, for any value of* y, *the column space of the matrix*

$$I_p - \mathrm{var}(X|\tilde{Y} = k)$$

is a subspace of the central space. Consequently, the column space of the matrix

$$E\left[I_p - \mathrm{var}(X|\tilde{Y})\right]^2$$

is a subspace of $\mathcal{S}_{Y|X}$.

7.2 Sample estimator

- Standardize, and centerize, to obtain $\widehat{Z}_i$ and $\widehat{Y}_i$, $i = 1, \ldots, n$.
- Divide $\mathcal{Y}$ into k intervals, say $I_1, \ldots, I_k$. Let n_i be the number of observations in slice i; that is

  $$n_i = \sum_{j \in I_i} 1.$$

  Compute, for $i = 1, \ldots, k$,

  $$\widehat{\Sigma}_i = \frac{1}{n_i} \sum_{j \in I_i} (\widehat{Z}_j - \hat{\mu}_i)(\widehat{Z}_j - \hat{\mu}_i)^T.$$

- Construct the SAVE matrix

  $$\widehat{U} = \sum_{i=1}^{k} \frac{n_k}{n} \widehat{\Sigma}_i^2.$$

- Find $v_1, \ldots, v_q$, the eigenvalues of $\widehat{U}$ corresponding to its largest eigenvalues, and then do a back-transformation to get $w_1, \ldots, w_q$.

8 Dimension reduction for conditional mean

Reference Cook and Li (2003, *Ann. Statist.*); Cook and Li (2003, *Ann. Statist.* tentatively accepted).

In many situations regression analysis is mostly concerned with inferring about the conditional mean of the response given the predictors, and less concerned with the other aspects of the conditional distribution. In this and the next few sections we develop dimension reduction methods that incorporate this consideration. We introduce the notion of the Central Mean Subspace (CMS), a natural inferential object for dimension reduction when the mean function is of interest. We study properties of the CMS, and develop methods to estimate it. These methods include a new class of estimators which requires fewer conditions than pHd, and which displays a clear advantage when one of the conditions for pHd is violated. CMS also reveals a transparent distinction among the existing methods for dimension reduction: OLS, pHd, SIR, and SAVE. We will apply the new methods to a data set involving recumbent cows.

8.1 Central mean subspace.

When focusing on the conditional mean, dimension reduction hinges on finding a $p \times k$ matrix α, $k \leq p$, so that the $k \times 1$ random vector $\alpha^T X$ contains all the information about Y that is available from $\mathrm{E}(Y|X)$. This is less restrictive than requiring that $\alpha^T X$ contain all the information about Y that is available from X as in the current literature associated with the central subspace. The following definition formalizes this idea.

Definition 8.1 *If*

$$Y \perp\!\!\!\perp \mathrm{E}(Y|X)|\alpha^T X$$

then $\mathcal{S}(\alpha)$ is a mean dimension-reduction subspace for the regression of Y on X.

It follows from this definition that a dimension-reduction subspace is necessarily a mean dimension-reduction subspace, because $Y \perp\!\!\!\perp X|\alpha^T X$ implies $Y \perp\!\!\!\perp \mathrm{E}(Y|X)|\alpha^T X$. The next proposition gives equivalent conditions for the conditional independence used in Definition 8.1.

Proposition 8.1 *The following statements are equivalent:*

(i) $Y \perp\!\!\!\perp \mathrm{E}(Y|X)|\alpha^T X$,

(ii) $\mathrm{Cov}[(Y, \mathrm{E}(Y|X))|\alpha^T X] = 0$,

(iii) $\mathrm{E}(Y|X)$ *is a function of* $\alpha^T X$.

The first condition is the same as Definition 8.1. The second condition is that, given $\alpha^T X$, Y and $\mathrm{E}(Y|X)$ must be uncorrelated. The final condition is what might be suggested by intuition, $\mathrm{E}(Y|X) = \mathrm{E}(Y|\alpha^T X)$. Any of these three conditions could be taken as the definition of a mean dimension-reduction subspace.

Proof of Proposition 8.1 That (i) implies (ii) is immediate. That (iii) implies (i) is also immediate, because, if $\mathrm{E}(Y|X)$ is a function of $\alpha^T X$, then, given $\alpha^T X$, $\mathrm{E}(Y|X)$ is a constant, and hence independent of any other random variable. Now let's prove that (ii) implies (iii). By (ii),

$$\mathrm{E}\{Y\mathrm{E}(Y|X)|\alpha^T X\} = \mathrm{E}(Y|\alpha^T X)\mathrm{E}\{\mathrm{E}(Y|X)|\alpha^T X\}.$$

The left hand side is

$$\mathrm{E}[\mathrm{E}\{Y\mathrm{E}(Y|X)|X\}|\alpha^T X] = \mathrm{E}\{[\mathrm{E}(Y|X)]^2|\alpha^T X\},$$

and the right hand side is $\{\mathrm{E}[\mathrm{E}(Y|X)|\alpha^T X]\}^2$. Therefore

$$\mathrm{Var}[\mathrm{E}(Y|X)|\alpha^T X] = 0.$$

Thus, given $\alpha^T X$, $\mathrm{E}(Y|X)$ is a constant. □

Paralleling the development of central subspaces, we would like the smallest mean dimension-reduction subspace, as formalized in the next definition.

Definition 8.2 *Let $\mathcal{S}_{\mathrm{E}(Y|X)} = \cap\mathcal{S}_m$ where intersection is over all mean dimension-reduction subspaces $\mathcal{S}_m$. If $\mathcal{S}_{\mathrm{E}(Y|X)}$ is itself a mean dimension-reduction subspace, it is called the central mean dimension-reduction subspace, or simply the central mean subspace (*CMS*).*

The Central Mean Subspace enjoys the similar invariance property as does the central space.

Proposition 8.2 *If $Z = A^T X + b$ for some nonsingular matrix A and some vector b. Then $\mathcal{S}_{\mathrm{E}(Y|Z)} = A^{-T}\mathcal{S}_{\mathrm{E}(Y|X)}$ is the* CMS *for the regression of Y on Z.*

PROOF. We will use equivalent condition iii to prove this result. Suppose that β is a $p \times q$ matrix whose columns form a basis in $\mathcal{S}_{\mathrm{E}(Y|X)}$. Then

$$\begin{aligned} E(Y|Z) &= E(Y|X) \\ &= E(Y|\beta^T X) \\ &= E(Y|\beta^T A^{-T} A^T (X + b)) \\ &= E(Y|\beta^T A^{-T} Z) \end{aligned}$$

So $A^{-T}\mathcal{S}_{\mathrm{E}(Y|X)}$ is a dimension reduction space for Y versus Z; and hence $\mathcal{S}_{\mathrm{E}(Y|Z)} \subset A^{-T}\mathcal{S}_{\mathrm{E}(Y|X)}$. The inverse inclusion can be proved similarly. □

So, as before, we work with the standardized predictor ($E(X) = 0$, $\mathrm{var}(X) = I_p$).

9 Categorization of existing Dimension Reduction methods

Having established some basic properties of the CMS, we now turn our attention to finding population vectors in that subspace. We will survey available methods for constructing vectors in the central subspace and demonstrate that some of them in fact produces vectors in CMS. By categorizing and assessing these methods in relation to CMS, we set the stage for a new estimation method introduced in a later section.

9.1 Vectors in CMS

We will consider an objective function of the form $R(a, b) = \mathrm{E}[L(a + b^T X, Y)]$ where $a \in \mathbb{R}$ and $b \in \mathbb{R}^p$. Here, the expectation is with respect to the joint distribution of Y and X. This use of an objective function is not meant to imply that any associated model is true or even provides an adequate fit of the data. Nevertheless, there is a useful connection between $\mathcal{S}_{\mathrm{E}(Y|Z)}$ and the vectors derived from these objective functions.

Let

$$(\alpha, \beta) = \arg\min_{a,b} R(a, b) \tag{18}$$

denote the population minimizers, and let η be a basis matrix for $\mathcal{S}_{Y|Z}$.

We restrict attention to objective functions

$$L(a + b^T X, Y) = -Y(a + b^T X) + \phi(a + b^T X) \tag{19}$$

based on the linear exponential family for some strictly convex function ϕ, then $\boldsymbol{\beta}$ always belongs to $\mathcal{S}_{\mathrm{E}(Y|X)}$.

Theorem 9.1 *Let γ be a basis matrix for $\mathcal{S}_{\mathrm{E}(Y|X)}$, assume that $E(X|\gamma^T X)$ is a linear function of X and let β be as defined in (18) using an exponential family objective function (19). Then $\beta \in \mathcal{S}_{\mathrm{E}(Y|X)}$.*

The exponential family objective function (19) covers many estimation methods used in practice. In particular, OLS is obtained by setting $\phi(K) = K^2/2$.

Proof of Theorem 9.1 We first rewrite $R(a, b)$ making use of the fact that γ is a basis for the central mean subspace:

$$\begin{aligned} R(a,b) &= E[-Y(a + b^T X) + \phi(a + b^T X)] \\ &= E[-E(Y|\gamma^T X)(a + b^T X) + \phi(a + b^T X)] \\ &\geq E[-E(Y|\boldsymbol{\gamma}^T X)(a + b^T E(X|\boldsymbol{\gamma}^T X)) + \phi(a + b^T E(X|\boldsymbol{\gamma}^T X))] \\ &= E[-Y(a + b^T P_\gamma X) + \phi(a + b^T P_\gamma X)]. \end{aligned}$$

The second equality follows because $\boldsymbol{\gamma}$ is a basis for $\mathcal{S}_{\mathrm{E}(Y|X)}$. The inequality follows because of convexity. The next equality stems from the linearity of $E(X|\boldsymbol{\gamma}^T X)$ which is equivalent to requiring that $E(X|\boldsymbol{\gamma}^T X) = P_\gamma X$, where P_γ is the projection onto $\mathcal{S}_{\mathrm{E}(Y|X)}$ with respect to the usual inner product.

Thus,

$$R(a, b) \geq R(a, P_\gamma b)$$

and the conclusion now follows because β is unique. □

The next theorem shows that the pHd vectors are actually in CMS.

Theorem 9.2 *Suppose that Assumptions 4.1 and 5.1 hold and that $E(X) = 0$ and $var(X) = I_p$. Then the column space of H_1 (the y-based pHd) (or H_2 (the e-based pHd)) is a subspace of $\mathcal{S}_{\mathrm{E}(Y|X)}$.*

Proof. The proof is essentially the same as before, when we proved that the column space of H_1 (H_2) is a subspace of the central space. Recall that the key step there was the fact that

$$E(Y|X) = E(Y|\beta^T X) \text{ or } E(e|X) = E(e|\beta^T X).$$

We proved this by using the assumption $Y \perp\!\!\!\perp X \perp\!\!\!\perp \beta^T X$. However, in the present case this follows directly from the definition of the CMS: $E(Y|X) = E(Y|\beta^T X)$. That $E(e|X) = E(e|\beta^T X)$ can also be proved using this definition and an argument similar to the pHd theorem for central space. □

9.2 Vectors in CS but not in CMS

Here we point out that the SIR and SAVE vectors may not belong to CMS, though SIR belongs to CS under Assumption 4.1 and SAVE belongs to CS under Assumptions 4.1 and 5.1.

Recall that in proving the that SIR and SAVE are in the central space we used the fact that

$$E(X|Y, \beta^T X) = E(X|\beta^T X).$$

This is deduced from $X \perp\!\!\!\perp Y|\beta^T X$. However, this cannot be deduced from the definition of the CMS, which is $E(Y|X) = E(Y|\beta^T X)$.

Similarly in the derivation of CS, we used, in addition, the fact

$$\text{var}(X|\beta^T X, Y) = \text{var}(X|\beta^T X),$$

which, again, is deduced from $X \perp\!\!\!\perp Y|\beta^T X$ but cannot be deduced from the definition of CMS.

Hence SIR and SAVE vectors need not be in CMS. In summary, we have the following table:

	LCM	CCV
CMS	OLS	pHd
CS	SIR	SAVE

In a sense, if in a regression analysis we are mainly interested in the conditional mean and not the conditional distribution itself, then CMS is the parameter of interest and CS \ CMS is the nuisance parameter. Thus OLS and pHd can be viewed as the estimator of the parameter of interest.

10 Iterative Hessian Transformation

We see that OLS and pHd estimate CMS but SIR and SAVE may estimate the whole space CS. However, OLS can only estimate one direction. If the response is binary, the SIR can only estimate one direction. However, if we are only willing to use the LCM, and we think there are more vectors in the dimension reduction space, what should we do ?

Theorem 10.1 *Suppose that Assumption 4.1 holds and that* $E(X) = 0$ *and* $var(X) = I_p$. *Then the central mean space is an invariant subspace of the linear transformation* $v \mapsto Hv$ *(where* H *can be either* H_1 *or* H_2*). In symbols,*

$$H_1\mathcal{S}_{\text{E}(Y|X)} \subset \mathcal{S}_{\text{E}(Y|X)} \quad H_2\mathcal{S}_{\text{E}(Y|X)} \subset \mathcal{S}_{\text{E}(Y|X)}.$$

PROOF. We only prove this for $H = H_1$, the other case can be proved similarly. We will write H_1 as H. Let β be a matrix whose columns form a basis in $\mathcal{S}_{\mathrm{E}(Y|X)}$. Let v be a vector in $\mathcal{S}_{\mathrm{E}(Y|X)}$. We need to show that Hv also belongs to $\mathcal{S}_{\mathrm{E}(Y|X)}$. Note that

$$\begin{aligned} Hv &= E(YXX^Tv) \\ &= E(E(Y|X)X(X^Tv)) \\ &= E(E(Y|\beta^TX)X(X^Tv)) \\ &= E(YE(XX^TX|\beta^TX)) \end{aligned}$$

Because v is a vector in $\mathcal{S}_{\mathrm{E}(Y|X)}$, v^TX is a function of β^TX. Therefore, the right hand side is

$$\begin{aligned} E(YE(X|\beta^TX)v^TX) &= E(YP_\beta Xv^TX) \\ &= PE(YXX^Tv) \\ &= PHv. \end{aligned}$$

Thus Hv is equal to its projection, and therefore must be a vector in $\mathrm{ran}(P_\beta)$. □

The point of this theorem is that only one assumption — Assumption 4.1 is needed for $\mathcal{S}_{\mathrm{E}(Y|X)}$ to be an invariant subspace. Therefore, if we can find a "seek vector" in $\mathcal{S}_{\mathrm{E}(Y|X)}$ under Assumption 4.1, then the process can bring out other vectors in $\mathcal{S}_{\mathrm{E}(Y|X)}$ without evoking the Assumption 5.1. Note that the ols vector $\alpha = E(XY)$ belongs to $\mathcal{S}_{\mathrm{E}(Y|X)}$ (or for that matter any vector given in the OLS theorem for CS).

Corollary 10.1 *Under Assumption 4.1*

(i) $\mathrm{Span}\{H_1^j\,\alpha:\ j=0,1,...\} \subseteq \mathcal{S}_{\mathrm{E}(Y|X)}$, *and*

(ii) $\mathrm{Span}\{H_2^j\,\alpha:\ j=0,1,...\} \subseteq \mathcal{S}_{\mathrm{E}(Y|Z)}$.

Because this method is based on Iterative transformation of a seed vector by the Hessian matrix, we call it the Iterative Hessian Transformation estimator of the central mean space.

One question that remains is how large j must be in order for the first j vectors, $\boldsymbol{\beta}_{yz}$, ..., $\boldsymbol{\Sigma}_{yzz}^{j-1}\boldsymbol{\beta}_{yz}$, to exhaust all possible vectors in the sequence. This is important because in practice we can compute only a finite number of these vectors. This question is answered by the next proposition; its proof is straightforward and omitted.

Proposition 10.1 *Let A be a $p \times p$ matrix and α be a p-dimensional vector. If $A^j\alpha$ belongs to the subspace spanned by α, ..., $A^{j-1}\alpha$, then so does $A^s\alpha$ for any $s > j$.*

PROOF. Suppose $A^j\alpha$ belongs to the subspace spanned by α, ..., $A^{j-1}\alpha$. Then there is a vector $w \in \mathbb{R}^j$ such that

$$A^j\alpha = (\alpha, ..., A^{j-1}\alpha)w.$$

Then

$$\begin{aligned} A^{j+1}\alpha &= A(A^j\alpha) \\ &= A(\alpha, ..., A^{j-1}\alpha)w \\ &= (A\alpha, ..., A^j\alpha)w \\ &= (A\alpha, ..., A^{j-1}\alpha, (\alpha, ..., A^{j-1}\alpha)w)w. \end{aligned}$$

However, the right hand side is obviously a linear combination of $\alpha, ..., A^{j-1}\alpha$. □

Since $H^j\,\alpha$ belongs to $\mathcal{S}_{\mathrm{E}(Y|X)}$, which has dimension q, Proposition 10.1 implies that there is an integer $s \leq q$ such that the first s vectors in the sequence, α, ..., $H^{s-1}\alpha$, are linearly independent, and all the subsequent vectors are linearly dependent on them. In particular, we can focus on the vectors α, ..., $H^{p-1}\alpha$ without missing any vectors in the subsequent iteration.

This suggests the following estimation scheme :

- Standardize and centerize as before, to get $\widehat{Z}$'s and $\widehat{Y}$'s.
- Construct the OLS estimator $\hat{\alpha}$ based on $\widehat{Z}$ and $\widehat{Y}$, as described before.
- Construct Hessian matrix $\widehat{H}$ based on $\widehat{Y}$, $\widehat{Z}$, as before. Here $\widehat{H}$ can be either $\widehat{H}_1$ or $\widehat{H}_2$.
- Let

  $$\widehat{B} = (\hat{\alpha}, \ldots, \widehat{H}^{p-1}\hat{\alpha}).$$

  Assuming q is known, let $v_1, \ldots, v_q$ be the eigenvectors of $\widehat{B}\widehat{B}$ corresponding to its q largest eigenvalues. This is used to estimate $\mathcal{S}_{\mathrm{E}(Y|Z)}$.
- Transforming them back, as described before, to estimate the original space $\mathcal{S}_{\mathrm{E}(Y|X)}$.

For a numerical illustration, we generated 200 observations on 5 predictors and a response as follows:

$$\begin{aligned} X_1 &= \varepsilon_1 \\ X_2|X_1 &= X_1 + \varepsilon_2 \\ X_3 &= \varepsilon_3 \\ X_4|X_2 &= (1 + X_2/2)\varepsilon_4 \\ X_5 &= \varepsilon_5 \\ Y &= X_1 + X_2^2/2 \end{aligned}$$

All errors ε_k are independent standard normal random variables. The response Y was generated without error to emphasize the qualitative nature of the results. The CMS is spanned by $(1,0,0,0,0)^T$ and $(0,1,0,0,0)^T$. Assumption 4.1 holds, but 5.1 does not hold because $\mathrm{var}(X_4|X_2) = (1 + X_2/2)^2$. Table 1 gives the first two pHd directions, $\widehat{h}_1$ and $\widehat{h}_2$, and the first two sample IHT directions, $\widehat{u}_1$ and $\widehat{u}_2$. pHd found X_2 and X_4

Table 1: Sample pHd directions $\widehat{h}_1$ and $\widehat{h}_2$ and IHT directions $\widehat{u}_1$ and $\widehat{u}_2$ from the simulated data.

	$\widehat{h}_1$	$\widehat{h}_2$	$\widehat{u}_1$	$\widehat{u}_2$
X_1	−0.098	0.166	0.022	−0.996
X_2	−0.984	−0.011	0.999	0.024
X_3	0.050	0.151	0.001	0.062
X_4	−0.142	−0.974	0.032	−0.016
X_5	−0.017	0.035	0.016	0.017

to be the important predictors, while IHT correctly picked X_1 and X_2. In effect, pHd missed the linear component in favor of the quadratic component and X_4. Figure 1 gives a visual representation of these results. The response surface from IHT gives a very good representation of the true surface, while the surface for pHd shows only a relatively rough quadratic.

11 Local methods

11.1 Minimum Average Variance Estimator (MAVE)

Xia, Tong, Li, and Zhu (2002 JRSS-B). This is based on the minimization of the expectation of conditional variance. That is, to minimize

$$E\{\mathrm{var}(Y|\gamma^T X)\}.$$

over all $\gamma \in \mathbb{R}^{p\times q}$.

Sample estimate: Replace conditional expectation by kernel estimators (or polynomial regression estimate). Then minimize the approximated objective function.

11.2 Structural Adaptive Estimator (SAE)

Hristache, Juditsky, Polzehl, and Spokoiny (2001, AOS). This method is based on the following fact. If a function $f(x)$ depends on x only through $\beta^T x$, then the family of gradients:

$$\{\partial f(x)/\partial x :\ x \in \mathcal{X}\}$$

span the column space of β. At the sample level, the gradient are estimated by local linear regression. Then a principal component analysis is conducted on the gradient vectors.

11.3 Summary and comparison

The advantages of the global methods, such as OLS, pHd, SIR, SAVE, and IHT, are (1) they are $\sqrt{n}$-consistent, (2) they are simple to compute. The drawbacks (1) they are not guaranteed to be exhaustive. That is, they may estimate a set of vecotrs that, though *belong* to the Central Subspace, need not *span* the central subspace. (2) they require the predictor to satisfy L.C.M. and/or C.C.V., which can be restrictive for some applications.

In comparison, the global methods such as MAVE and SAE are under reasonable conditions exhaustive, and do not require L.C.M. and C.C.V. However, they are in general not $\sqrt{n}$-consistent, and their computation is more difficult.

12 Contour Regression

The global methods above are $\sqrt{n}$-consistent and computationally inexpensive, due to the fact that they exploit *global* features of the dependence of Y on X, as captured by mixed moments estimated on the data. Note that the $\sqrt{n}$-consistency is achieved regardless of the original dimension p and the structural dimension q. This important property hinges on the fact that the dimension reduction problem is intrinsically global, in the sense that the response surface is constant along the $p-q$ dimensional orthogonal complement of the central subspace. The above methods achieve the $\sqrt{n}$ rate precisely because they exploit this global property.

The methods also have common limitations. First, all of them require linearity of the mean relationships among predictors along the central subspace. When this condition fails, the methods may produce estimates that converge at the $\sqrt{n}$ rate to directions outside $\mathcal{S}_{Y|X}$. Because violations of this condition cannot be diagnosed prior to estimating β, it is often replaced by the more restrictive assumption that X be elliptically distributed. Ellipticity guarantees linearity of the mean relationships among predictors along any subspace, and can be at least partially diagnosed and remedied.

Second, none of the methods is guaranteed to be exhaustive: the estimates are $\sqrt{n}$-consistent for vectors in $\mathcal{S}_{Y|X}$, but these vectors may not span the whole central subspace. This is arguably one of the most important shortcomings of these methods. An instance is the heavy reliance of methods such as OLS and SIR on linear trends in the dependence of Y on X. For example, if $Y = (\beta^T X)^2 + \epsilon$ with $\beta \in \mathbb{R}^p$ and $X \sim N(0, I_p)$, OLS and SIR will estimate $0 \in \mathcal{S}_{Y|X} = \text{span}(\beta)$, but fail to detect β itself.

The adaptive methods relax the assumptions on X and achieve exhaustiveness but do not have $\sqrt{n}$-consistency in general.

Here, we target directly the contour directions of the response surface. Contour directions are those along which the response has small variation; they span the orthogonal complement of the central subspace. They can be extracted according to two measures of variation in the response, leading to two methods: *Simple* and *General Contour Regression* (SCR and GCR). Contour Regression guarantees exhaustive estimation of the central subspace under ellipticity of X and very mild additional assumptions. It also

proves robust to violations of ellipticity. Moreover, Contour Regression achieves $\sqrt{n}$-consistency regardless of the dimensions p and q and are computationally inexpensive.

13 Simple contour regression

Let $(X_1, Y_1), ..., (X_n, Y_n)$ be independent copies of the random pair (X, Y), where $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. Let F_n denote the empirical distribution based on the data, and F_{XY} be the joint distribution of (X, Y). We will be concerned with matrix-valued estimators of the form $T(F_n)$. If the columns of $T(F_{XY})$ belong to $\mathcal{S}_{Y|X}$, then we say $T(F_n)$ is unbiased at the population level. If the columns of $T(F_{XY})$ actually span $\mathcal{S}_{Y|X}$, then we say that $T(F_n)$ is exhaustive at the population level. If $T(F_n)$ converges at the $\sqrt{n}$ rate to $T(F_{XY})$ in the first case, then we say it is $\sqrt{n}$-consistent. If the $\sqrt{n}$ convergence holds in the second case, then we say $T(F_n)$ is $\sqrt{n}$-exhaustive.

13.1 The estimator

Let $(\tilde{X}, \tilde{Y})$ be an independent copy of (X, Y) and suppose that the central subspace $\mathcal{S}_{Y|X}$ for the regression of Y on X is spanned by the column space of a $p \times q$ matrix β with $q < p$. Consider the matrix

$$K(c) = E\left[(\tilde{X} - X)(\tilde{X} - X)^T \middle| |\tilde{Y} - Y| \leq c\right].$$

We will show that, for sufficiently small $c > 0$, the eigenvectors of $K(c)$ corresponding to its smallest q eigenvalues span the central subspace. For this purpose, we need the following assumption.

Assumption 13.1 *For any choice of vectors $v \in \mathcal{S}_{Y|X}$ and $w \in (\mathcal{S}_{Y|X})^{\perp}$ such that $\|v\| = \|w\| = 1$, and any sufficiently small $c > 0$, we have*

$$var\left[w^T(\tilde{X} - X) \middle| |\tilde{Y} - Y| \leq c\right] > var\left[v^T(\tilde{X} - X) \middle| |\tilde{Y} - Y| \leq c\right] \tag{20}$$

This assumption is a reasonable one: because the conditional distribution of $Y|X$ depends on $v^T X$ but not on $w^T X$, we expect Y to vary more with $v^T X$ than it does with $w^T X$. Hence, intuitively, within the same increment of Y, $w^T X$ should vary more than $v^T X$ does. In Section 14 we will prove that Assumption 13.1 holds under fairly general conditions. Here we give a few examples to illustrate its wide applicability.

Example 13.1 Suppose $X = (X_1, X_2)^T \sim N(0, I_2)$ and $Y = X_2^2 + \epsilon$, with $\epsilon \perp\!\!\!\perp X$ and $\epsilon \sim N(0, \sigma^2)$. For this regression $\mathcal{S}_{Y|X}$ is the one-dimensional span of $\beta = (0, 1)^T$. We verified numerically that the above sufficient condition holds.

Example 13.2 Let $X = (X_1, X_2)^T$ and ϵ be as in Example 13.1, and $Y = (X_2 - 1)^3 + \epsilon$. We used again numerical integration to verify the sufficient condition for $c = 0.1, 0.5, 1, \ldots, 3$ and $\sigma = 0.1, 0.2, 0.3, \ldots, 2$, and again obtained values below 2 in all cases.

We checked numerically numerous other $f(\cdot)$ functions such as polynomials, exponential and logorithmic functions, trigonometric functions, etc., never encountering a violation of Assumption 13.1.

In the next theorem we prove that if X is elliptically contoured and Assumption 13.1 holds, then the population vectors from SCR exhausts the central subspace $\mathcal{S}_{Y|X}$. We first consider the standardized X.

Theorem 13.1 *Suppose that X has an elliptical distribution with $E(X) = 0$ and $var(X) = I_p$. If Assumption 13.1 holds, then, for a sufficiently small c, the eigenvectors of $K(c)$ corresponding to its smallest q eigenvalues span the central subspace $\mathcal{S}_{Y|X}$.*

The estimating procedure will mimick the theoretical development in Section 13.1:

- Compute sample mean and variance matrix of the predictor X

$$\hat{\mu} = n^{-1}\sum_{i=1}^{n} X_i, \quad \widehat{\Sigma} = n^{-1}\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})^T.$$

- compute the matrix-valued U-statistic

$$\widehat{H}(c) = \frac{1}{\binom{n}{2}} \sum_{(i,j)\in N} (X_j - X_i)(X_j - X_i)^T I(|Y_j - Y_i| \leq c), \tag{21}$$

  where N is the index set $\{(i,j) : i = 2,\ldots,n; j = 1,\ldots,i-1\}$.

- Compute the spectral decomposition of $\widehat{\Sigma}^{-1/2}\widehat{H}(c)\widehat{\Sigma}^{-1/2}$ and let $\hat{\gamma}_{p+q-1},\ldots,\hat{\gamma}_p$ be the eigenvectors corresponding to the smallest q eigenvalues.

- The span of these eigenvectors estimates $\mathcal{S}_{Y|Z}$, where Z is the standardized version of X. Thus, our estimate of the central subspace is

$$\widehat{\mathcal{S}}_{Y|X} = \text{span}(\widehat{\Sigma}^{-1/2}\hat{\gamma}_{p-q+1},\ldots,\widehat{\Sigma}^{-1/2}\hat{\gamma}_p).$$

Theorem 13.2 *Suppose that Σ is nonsingular and that the components of X have finite fourth moments. Then*

$$\widehat{\Sigma}^{-1/2}\widehat{H}(c)\widehat{\Sigma}^{-1/2} = \Sigma^{-1/2}H(c)\Sigma^{-1/2} + O_p(n^{-1/2}).$$

As a consequence of this theorem, $\hat{\gamma}_{p-q+1},\ldots,\hat{\gamma}_p$ are a $\sqrt{n}$-exhaustive estimator of $\mathcal{S}_{Y|Z}$, and hence $\widehat{\Sigma}^{-1/2}\hat{\gamma}_{p-q+1},\ldots,\widehat{\Sigma}^{-1/2}\hat{\gamma}_p$ are a $\sqrt{n}$-exhaustive estimator of $\mathcal{S}_{Y|X}$.

14 Exhaustive estimation

In order to place the theory of Simple Contour Regression on a firmer foundation we devote this section to deriving a sufficient condition for Assumption 13.1. As shown in the previous sections, if this assumption holds, then SCR provides $\sqrt{n}$-*exhaustive* estimation of the central subspace $\mathcal{S}_{Y|X}$. Sufficient conditions of this type are extremely elusive – to our knowledge none has been established with reasonable generality for the other $\sqrt{n}$-consistent estimators such as OLS, PHD, SIR or SAVE.

We will need the notion of stochastic ordering. Let S and T be two random variables. We say that S is stochastically less than or equal to T if, for any real number r, $Pr(S \leq r) \geq Pr(T \leq r)$, and write this as $S \leq_d T$. If, in addition, the inequality is strict on a subset of the real line with positive Lebesgue measure, we say that S is stochastically (strictly) less than T and write $S <_d T$. The following lemma is obvious, and its proof will be omitted.

In developing a sufficient condition for Assumption 13.1, consider the location structure $Y = f(\beta^T X) + \epsilon$ with $\epsilon \perp\!\!\!\perp X$ and $E(\epsilon) = 0$. Ultimately, the sufficient condition will be imposed on the behavior of $f(\cdot)$. Let $(\tilde{X}, \tilde{\epsilon})$ be an independent copy of (X, ϵ), $\Delta = \tilde{X} - X$, $T = \tilde{\epsilon} - \epsilon$. and $F_T(\cdot)$ be the cumulative distribution function of T. Write $f(\beta^T x)$ merely as $g(x)$.

Theorem 14.1 *Suppose that X has an elliptically contoured distribution with $E(X) = 0$ and $var(X) = I_p$, and that C.C.V. holds. Moreover, suppose that the density $f_T(t)$ of $F_T(t)$ decreases as $|t|$ increases. If, for any $\alpha \in \mathcal{S}_{Y|X}$, and whenever $0 \leq \delta_1 < \delta_2$, we have*

$$\begin{aligned} |g(X+\Delta) - g(X)| \;\Big|\; \{|\alpha^T \Delta| = \delta_1\} \\ <_d \; |g(X+\Delta) - g(X)| \;\Big|\; \{|\alpha^T \Delta| = \delta_2\}. \end{aligned} \tag{22}$$

then 13.1 holds for every $c > 0$.

To understand the intuition behind condition (22), first consider the case where X is a scalar random variable. Intuitively, condition (22) should hold trivially if g is a monotone function, because it holds pointwise in $X = x$ with $<_d$ replaced by ordinary inequality $<$ (see Example 14.1 below). However, condition (22) by no means restricts $g(\cdot)$ to be monotone, because being stochastically large or small is an average behaviour of all values of X, and is not necessarily being large or small for every single value of $X = x$. It then does seem to make sense to assume that $g(X + \Delta)$ is collectively farther away from $g(X)$ if Δ is larger: this is simply requiring g to be reasonably variable. In the multivariate case, condition (22) requires this to hold along any direction α in the space $\mathcal{S}_{Y|X}$, which is the space along which $g(x)$ does vary. Also the requirement that $f_T(t)$ decreases with $|t|$ is not a severe restriction, considering that this density is symmetric about 0 by construction.

Example 14.1 Suppose that $f(x_2)$ is a continuous and monotone function, which, without loss of generality, can be assumed to be monotone increasing. Then the sufficient condition is satisfied.

Example 14.2 Let $f(x_2) = (x_2 - a)^2$. Example 13.1 is a special case of this model with $a = 0$ and $\epsilon \sim N(0, \sigma^2)$. Then the sufficient condition is satisfied.

15 General contour regression

15.1 Estimation

The idea underlying SCR is to use the inequality $|Y - \tilde{Y}| \leq c$ to identify vectors aligned with the contour directions. However, this inequality also picks up other directions when the regression function is non-monotone. Under ellipticity, such directions are averaged out, so that the method remains $\sqrt{n}$-exhaustive. Nevertheless, these "wrong" directions do tend to decrease efficiency by blurring up the "right" ones. In other words, the inequality $|Y - \tilde{Y}| \leq c$ is not a very sensitive contour identifier for non-monotone functions – even though it is sufficiently sensitive to maintain $\sqrt{n}$-exhaustiveness. We now illustrate this point using the model in Example 13.1.

To construct the left panel of Figure 1, we generated twenty observations (X_i, Y_i) $i = 1, \ldots, 20$ according to the model in Example 13.1, with $\sigma = 0.3$. We then used the threshold value $c = 0.5$, connecting by a solid line segment any two points $(X_i, X_j)^T \in \mathbb{R}^2$ satisfying $|Y_i - Y_j| \leq 0.5$. Roughly speaking , SCR picks up the contour directions by a Principal Component Analysis of the vectors represented by these line segments. We see that, though most of the segments are horizontal (i.e. aligned with the true contour direction), there are a considerable number of segments pointing to arbitrary directions. This is because Y is roughly U-shaped and the inequality $|Y_i - Y_j| \leq 0.5$ does not discriminate between the segments aligned with the contour and those across the U-shaped surface that also have small increments in Y. Though the arbitrary directions tend to average out due to the ellipticiy of the distribution of X, they make the picture less sharp, and the method less efficient.

To overcome this drawback we replace the contour identifier $|Y_i - Y_j| \leq c$ by a more sensitive one. Consider the variance of Y along the line through x_i and x_j. Formally, let $\ell(t; x_i, x_j) = (1 - t)x_i + tx_j$, $t \in \mathbb{R}$, be the straight line that goes through x_i and x_j, and define

$$V(x_i, x_j) = \text{var}\,(Y | X = \ell(t; x_i, x_j) \ \text{ for some } t\ ) .$$

We will aim at identifying the contour vectors by the smallness of this conditional variance.

The next task is to construct a sample estimate of $V(X_i, X_j)$. We will denote the line $\ell(\cdot; X_i, X_j)$ by $\ell(X_i, X_j)$. For any X_k, let $d(X_k, \ell(X_i, X_j))$ be the Euclidean distance between X_k and the line $\ell(X_i, X_j)$; that is,

$$d(X_k, \ell(X_i, X_j)) = \min_{t \in \mathbb{R}} \|X_k - \ell(t; X_i, X_j)\| ,$$

where $\|\cdot\|$ stands for the Euclidean norm. Because $\|X_k - \ell(t; X_i, X_j)\|^2$ is a quadratic function of t, this minimum distance can be expressed explicitly as

$$d(X_k, \ell(X_i, X_j)) = \frac{\|X_k - X_i\|^2 - \left\{(X_k - X_i)^T (X_j - X_i)\right\}^2}{\|X_j - X_i\|^2}.$$

For any $\rho > 0$, we define the cylinder of radius ρ connecting X_i and X_j to be the set

$$C_{ij}(\rho) = \{X_k : d(X_k, \ell(X_i, X_j)) \leq \rho,\ k = 1, \ldots, n\}.$$

According to this definition, each cylinder contains at least 2 points in the sample. Next, we estimate the variance of Y along these cylinders. Let $n_{ij}(\rho)$ be the number of points in the cylinder $C_{ij}(\rho)$, and let

$$\widehat{V}(X_i, X_j; \rho) = \frac{1}{n_{ij}(\rho)} \sum_{X_k \in C_{ij}(\rho)} \left(Y_k - \bar{Y}_{ij}(\rho)\right)^2,$$
$$\text{where } \bar{Y}_{ij}(\rho) = \frac{1}{n_{ij}(\rho)} \sum_{X_k \in C_{ij}(\rho)} Y_k.$$

We can now identify the contour directions by the smallness of $\widehat{V}(X_i, X_j; \rho)$.

Plotted in the right panel of Figure 1 are the same sample points as in the left panel, but with the line segments picked up by $\widehat{V}(X_i, X_j; \rho) \leq c$, where $c = 0.5$ and $\rho = 0.3$. We can see that many of the segments pointing to random directions in the left panel have been removed. To get a quantitative comparison, we calculated the first principal component for the line segments in each panel, which equals $(0.9169, 0.3991)^T$ for the left panel and $(0.9991, -0.0417)^T$ for the right panel. The latter is much closer to the direction $(1, 0)^T$, the population vector orthogonal to $\mathcal{S}_{Y|X}$.

We now construct the estimator of $\mathcal{S}_{Y|X}$. We standardize the predictor observations to $\widehat{Z}_i = \widehat{\Sigma}^{-1/2}(X_i - \hat{\mu})$, and form the matrix

$$\widehat{F}(c) = \frac{1}{\binom{n}{2}} \sum_{(i,j) \in N} (\widehat{Z}_j - \widehat{Z}_i)(\widehat{Z}_j - \widehat{Z}_i)^T I(\widehat{V}(\widehat{Z}_i, \widehat{Z}_j; \rho) \leq c), \tag{23}$$

where N is the same index set as used in (21). As in SCR, we take the spectral decomposition of $\widehat{F}(c)$, and use $\hat{\gamma}_{p+q-1}, \ldots, \hat{\gamma}_p$, the eigenvectors corresponding to the smallest q eigenvalues, to form

$$\hat{\mathcal{S}}_{Y|X} \quad = \quad \text{span}(\widehat{\Sigma}^{-1/2}\hat{\gamma}_{p-q+1}, \ldots, \widehat{\Sigma}^{-1/2}\hat{\gamma}_p).$$

15.2 Population-level exhaustiveness

Assume that X is already standardized to $E(X) = 0$ and $\text{var}(X) = I_p$ (so Z is X itself). The population version of the matrix $\widehat{F}(c)$ in (23) is

$$F(c) = E[(X - \tilde{X})(X - \tilde{X})^T I(V(X, \tilde{X}) \leq c)],$$

which is proportional to the matrix

$$G(c) = E\left((X - \tilde{X})(X - \tilde{X})^T \,\middle|\, V(X, \tilde{X}) \leq c\right).$$

Here we will demonstrate that, for sufficiently small c, the eigenvectors corresponding to the smallest q eigenvalues of $G(c)$ span $\mathcal{S}_{Y|X}$. For this purpose, we introduce an assumption that parallels Assumption 13.1. Again, $(\tilde{X}, \tilde{Y})$ indicates an independent copy of (X, Y).

Assumption 15.1 *For any choice of vectors $v \in \mathcal{S}_{Y|X}$ and $w \in (\mathcal{S}_{Y|X})^{\perp}$ such that $\|v\| = \|w\| = 1$, and any sufficiently small $c > 0$, we have*

$$var\left[w^T(\tilde{X} - X) \,\middle|\, V(X, \tilde{X}) \leq c\right] > var\left[v^T(\tilde{X} - X) \,\middle|\, V(X, \tilde{X}) \leq c\right]. \tag{24}$$

The interpretation of this assumption is similar to that of Assumption 13.1. We now deduce population exhaustiveness under this assumption. Once again, we do so for a spherical predictor without loss of generality.

Theorem 15.1 *Suppose that X has an elliptical distribution with $E(X) = 0$ and $var(X) = I_p$. Then, under Assumption 15.1, for sufficiently small $c > 0$, the eigenvectors of $G(c)$ corresponding to its smallest q eigenvalues span the central subspace $\mathcal{S}_{Y|X}$.*

15.3 Sufficient conditions for exhaustive estimation

Next, following a reasoning similar to that in Section 14, we derive a sufficient condition for Assumption 15.1.

Theorem 15.2 *Suppose that X has an elliptically-contoured distribution with $E(X) = 0$ and $var(X) = I_p$. Then Assumption 15.1 is satisfied for all sufficiently small $c > 0$ for which $\{(x, \tilde{x}) : V(x, \tilde{x}) \leq c\}$ is a non-empty set.*

The conditions in Theorem 15.2 are much weaker than those in Theorem 14.1. C.C.V. is not required in Theorem 15.2, and essentially no requirement is posed on the behavior of the mean function and the error term. Thus, GCR will be exhaustive under settings even more general than SCR.

16 Non-ellipticity

The population exhaustiveness of our contour-based methodology relies on ellipticity of the predictor distribution. This is because in the theoretical development we have treated the constant c in (21) and (23) as fixed with respect to the sample size n. Ellipticity of the distribution of X helps to balance out the effect of those line segments not aligned with the contour directions. However, especially when using GCR, whose contour identifier is more sensitive, we can obtain good performance even under violations of ellipticity.

We will show that the eigenvectors corresponding to the smallest $p-q$ eigenvalues of the matrix

$$A = E\left((\tilde{X}-X)(\tilde{X}-X)^T \,\middle|\, V(X,\tilde{X}) = \sigma^2\right)$$

span the orthogonal complement of the central subspace, $(\mathcal{S}_{Y|X})^{\perp}$, even when X is not elliptical. This suggests that if we let c decrease to σ^2 as n increases, then the eigenvectors corresponding to the smallest $p-q$ eigenvalues of $\widehat{F}(c)$ in (23) (after appropriate transformation by $\widehat{\Sigma}^{-1/2}$) will tend to recover the whole $\mathcal{S}_{Y|X}$, regardless of the shape of the distribution of X. In practice, if we make c small (i.e. close to the smallest value of $\widehat{V}(\widehat{Z}_i, \widehat{Z}_j)$ in (23)), then GCR is likely to estimate the central subspace exhaustively and effectively even if the shape of X does not help the process by averaging out erroneous directions, as is the case under ellipticity.

Theorem 16.1 *Suppose that X is a continuous random vector with an open support $\mathcal{X} \subset \mathbb{R}^p$. Then the matrix A has exactly $p-q$ zero eigenvalues, and their corresponding eigenvectors span $(\mathcal{S}_{Y|X})^{\perp}$. In symbols,*

$$\ker(A) = \mathcal{S}_{Y|X}$$

where $\ker(A) = \{h \in \mathbb{R}^p : Ah = 0\}$ is the kernel of A.

17 Simulation results

We now compare the performance of both versions of Contour Regression, SCR and GCR, with that of well known existing dimension reduction methods ensuring $\sqrt{n}$-consistency, such as OLS, SIR, PHD, and SAVE. For such comparisons, we need to introduce a measure of distance between two subspaces of $\mathbb{R}^p$. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two q-dimensional subspaces of $\mathbb{R}^p$ and $P_{\mathcal{S}_1}$, $P_{\mathcal{S}_2}$ be the orthogonal projections onto $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. A reasonable measure of distance between them is

$$\text{dist}(\mathcal{S}_1, \mathcal{S}_2) = \|P_{\mathcal{S}_1} - P_{\mathcal{S}_2}\|,$$

where $\|\cdot\|$ is the the Euclidean norm, i.e., the maximum singular value of a matrix.

Example 17.1 Consider the regression

$$Y = X_1^2 + X_2 + \sigma\epsilon,$$

where $X \sim N(0, I_4)$, $\epsilon \sim N(0, \sigma^2)$ and $\epsilon \perp\!\!\!\perp X$. Here, the central subspace is of dimension $q = 2$, and is spanned by the vectors $(1,0,0,0)^T$ and $(0,1,0,0)^T$. We compare SCR and GCR with SIR, SAVE, and PHD using $\sigma = 0.1$, 0.4, and 0.8. For each error value of σ, we draw 500 samples of size $n = 100$, and on each sample we apply the five estimation techniques to produce five estimates of $\mathcal{S}_{Y|X}$.

Table 1: Comparison of SCR and GCR and other estimates for Example 17.1

	SCR		GCR		SIR		SAVE		PHD	
σ	DIST	SE	DIST	SE	DIST	SE	DIST	SE	DIST	SE
0.1	0.23	0.11	0.16	0.07	0.78	0.24	0.43	0.25	0.80	0.21
0.4	0.25	0.11	0.20	0.08	0.79	0.23	0.54	0.27	0.79	0.21
0.8	0.31	0.13	0.32	0.16	0.80	0.23	0.73	0.25	0.79	0.21

18 An application

We consider data collected for Massachusetts four-year colleges in 1995, in an attempt to investigate how the percentage of freshmen that graduate (Grad) depends on variables measuring quality of incoming students and features of the colleges. There are $n = 46$ colleges and $p = 7$ predictors, which are: the percentage of freshmen that were among the top 25% percent in their graduating high school class (Top25), the median mathematics SAT score (MSAT), the median verbal SAT score (VSAT), the percentage of applicants accepted by the college (Accept), the percentage of accepted applicants who enroll (Enroll), the student-to-faculty ratio (SFRatio), and the out-of-state tuition (Tuition).

The scatter-plot matrix in Figure 2 reveals obvious curvatures in the mean dependencies among predictors. There appear to be violation of ellipticity. As discussed in Section 1, if these patterns lack a marked linear component along some of the directions they comprise, these directions may be missed by non-exhaustive methods that rely heavily on linear trends (e.g. SIR) even when ellipticity holds.

We apply GCR to the data set, taking the tube size to be $\rho = 0.03$ and including $4n = 184$ pairs of predictor differences with the smallest $\widehat{V}(\widehat{Z}_i, \widehat{Z}_j; \rho)$ values. This gives eigenvalues 2.1866, 3.6160, 7.6274, 7.7670, 8.6623, 9.6466 and 10.5777. Even though we do not have a rigorous testing theory at this stage, the clear separation between the first two eigenvalues and the following five allows us to infer the existence of two relevant directions, which correspond to the estimated linear combinations

$$\begin{aligned}
\text{GCR1} = {} & -0.6331(\text{Top25}) + 0.0168(\text{MSAT}) + 0.1519(\text{VSAT}) + 0.4068(\text{Accept}) \\
& -0.0726(\text{Enroll}) + 0.6365(\text{SFRatio}) + 0.0004(\text{Tuition}) \\
\text{GCR2} = {} & +0.1915(\text{Top25}) - 0.0605(\text{MSAT}) + 0.1336(\text{VSAT}) + 0.8642(\text{Accept}) \\
& -0.1622(\text{Enroll}) - 0.4106(\text{SFRatio}) + 0.0011(\text{Tuition})
\end{aligned}$$

Views of the 3D plot of Grad vs GCR1 and GCR2 are given in Figure 3 , revealing a peculiar "coiled" structure for the dependence of the response on the reduced predictors. While the linear component along GCR1 is strong (R-square approximately 56%), that along GCR2, which shows the bending of the coil, is much weaker (R-square approximately 8%).

Indeed, SIR applied to the same data unambiguously detects the first direction: the sample correlation between SIR1 (the first vector of SIR) and GCR1 is around 0.98, and

the p-value from the asymptotic chi-square test for SIR1 below 0.01, regardless of the number of slices employed in the SIR algorithm. However, SIR produces more ambiguous results on the existence of a second relevant direction, with p-values ranging between 0.10 and 0.30 depending on the number of slices.