Lecture 1: Dimension reduction estimation: can linear methods solve nonlinear problems?

Fall 2014

Hong Kong Baptist University
University of Connecticut

Part I: Why dimension reduction?

Question: Why dimension reduction?

- Regression modelling: Dimension reduction makes data visualization. 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.

Why dimension reduction?

Question: How to give a comprehensive residual plot?

- In the one-dimensional cases – the residual plot is informative. Given $(X_1, Y_1), \cdots, (X_n, Y_n)$, first fit a linear regression model:
- For $1 \leq i \leq n$,

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

- Find a regression estimator:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{(\sum_{i=1}^n)((X_i - \bar{X}))^2};$$
$$\beta_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \quad (1)$$

Why dimension reduction?

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

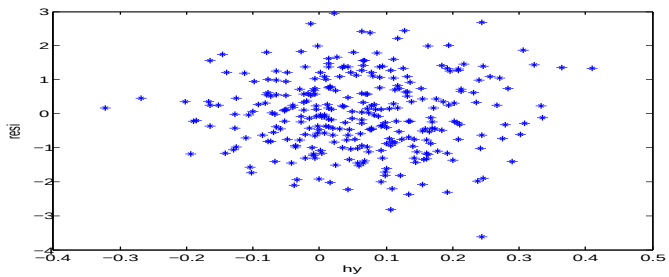


Figure: Residual Plot

Why dimension reduction?

- What should we do if X_i is a vector in R^p? Currently two methods are in frequent use:

  (1) Residual plot e_i versus $\hat{Y}_i$ (note that $\hat{Y}_i$ is always one-dimensional.)

  (2) 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 an example.

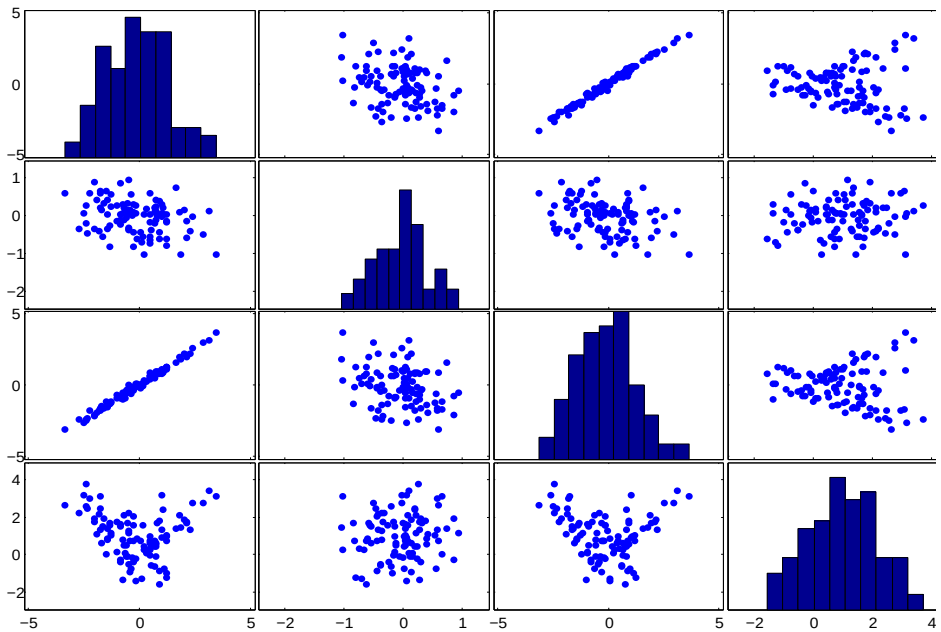
Why dimension reduction?

- Example 1. 100 pairs, $(X_1; Y_1); \cdots ; (X_{100}; Y_{100})$, are generated from some model, where X_i are in R^3 (three-dimensional vector $X_i = (X_{i1}, X_{i2}, X_{i3})$).
- The scatter plot matrix is produced. Show the scatter plot matrix below. From the scatter plot matrix the data appear to have the following features:

(1) Y doesn't seem to depend on X_2
(2) Y seems to depend on X_1 in a nonlinear way
(3) Y seems to depend on X_3 in a nonlinear way.

Why dimension reduction?

Figure: Scatter Plot Matrix



Why need dimension reduction?

- However, $(X; Y)$ are actually generated from the following model:

$$Y = |X_1 + X2| + \varepsilon,$$

where ε is independent of $X = (X_1; X_2; X_3)$ that is multivariate normal with mean 0 and covariance matrix Σ

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

- Note that Y does not depend on X_3, and Y does depend on X_2.
- X_2 has much smaller variance than that of X_1 or X_3 both have high correlation.

Why dimension reduction?

- Once again, the scatter plot matrix cannot capture the true relation between X and Y.
- What can truly capture the relation between X and Y 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.**
- Find the linear combination $X_1 + X_2$ before any regression modelling is performed!

Models under dimension reduction structure

- **Some examples:**
- Linear model: $Y = \beta^T X + \epsilon$,
- Generalized linear model: $Y = g(\beta^T X) + \epsilon$ for a given monotonic function g,
- Single-index model: $Y = g(\beta^T X) + \epsilon$ for an unknown function g,
- Multi-index model (1): $Y = g(\beta^T X) + \epsilon$, β is an $p \times q$ orthogonal matrix
- Multi-index model (2): $Y = g_1(\beta_1^T X) + g_2(\beta_2^T X)\epsilon$, where $\beta = (\beta_1, \beta_2)$,
- Multi-index model (3): $Y = g(\beta^T X, \epsilon)$.

Further observation on dimension reduction structure

- In these models, all the information on the response Y can be captured through $\beta^T X$, rather than through the original X!
- In other words, when $\beta^T X$ is given, no more information on Y can be acquired from the rest part of X. Y is then conditionally independent of X if ϵ is independent of X.
- We call this independence *the conditional independence,* and write this independence as $Y \perp\!\!\!\perp X | \beta^T X$.
- **Based on this, consider a generic framework such that "model-free" methods can be developed to estimate the parameter of interest and then to establish a model.**

Central Subspace

- The goal of dimension reduction is to seek $\beta \in R^{p \times q}$, $q < p$ such that
$$Y \perp\!\!\!\perp X | \beta^T X.$$
- However, β is not unique such that $Y \perp\!\!\!\perp X | \beta^T X$ unless $q = 1$ such that β is a vector. To make the notion clearly, define column space first.
- **Proposition 1.**
  1. If A is any $q \times q$ non-singular matrix, $Y \perp\!\!\!\perp X | \beta^T X$ if and only if $Y \perp\!\!\!\perp X | (\beta A)^T X$.
- **Column space**. For a matrix B we denote by $S(B)$ the subspace spanned by the columns of B: Let $b_1, \cdots, b_q$ be the columns of the matrix B. The space consists of all the linear combinations $c_1 b_1 + \cdots + c_q b_q$ for constants $c_1, \cdots, c_q$.

Central Subspace

- If γ is another matrix such that $S(\beta) \subseteq S(\gamma)$, then $Y \perp\!\!\!\perp X|\beta^T X$ implies $Y \perp\!\!\!\perp X|\gamma^T X$. Therefore, we are naturally interested in the smallest dimension reduction space, which achieves the maximal reduction of the dimension of X.

Definition

Definition *If the intersection of all dimension reduction spaces for* (X, Y) *is itself a dimension reduction space, this space is called the Central Space. Write* $S_{Y|X}$.

- Reference: Cook (1994, 1998). Thus the goal of dimension reduction is to find the central space $S_{Y|X}$.

Sufficient plot

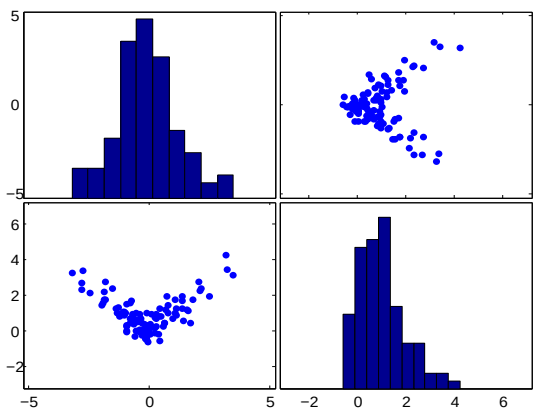
- Once we know the central space or equivalently its $p \times q$ base matrix $\beta = (\beta_1, \cdots, \beta_q)$, a comprehensive scatter plot or residual plot relating to $\beta^T X$ can be informative.
- When $q = 1$, scatter plot is informative and $q = 2$, use spin software to have a comprehensive view of the data. Usually this will suffice for most of the data analysis.
- Example 1 (continued) The model is

$$Y = |X_1 + X_2| + \varepsilon.$$

Thus the central space is spanned by $(1, 1, 0)$. The sufficient plot is the scatter plot of Y versus $X_1 + X_2$. Show this plot here.

Sufficient plot

Figure: Scatter Plot Matrix



Part II: Ordinary least squares

Assumption

Assumption 2.1 *Let β be a $R^{p\times q}$ matrix whose columns form an orthonormal basis in $S_{Y|X}$. Assume that $E(X|\beta^T X)$ is a linear function of X, that is, For a constant c and an $p \times q$ matrix C $E(X|\beta^T X) = c + C\beta^T X$.*

- 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 $\gamma \in R^{p\times q}$. This is equivalent to elliptical symmetry of X.

Ordinary least squares

- Elliptically symmetric distribution can often approximately be achieved by appropriate transformation of the original data; a certain power of the data, or logarithm of the data. See Cook and Weisberg (1994).

- Hall and 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.

Ordinary least squares

- We can get $C = \beta$.

Theorem

Theorem 2.2. *If Assumption 2.1 holds, then*

$$E(X|\beta^T X) = c + C\beta^T X =: P_\beta(X),$$

where $P_\beta = \beta\beta^T$ is the projection operator.

Ordinary least squares

- Recall the assumption that $E(X) = 0$, $var(X) = I_p$. Then we have

Theorem

Theorem 2.3. *Suppose that Assumption 2.1 holds. The vector $E(XY) = \beta c$ for an $1 \times q$ vector c is a vector in $S_{Y|X}$. In other words, $E(XY)$ can identify a vector in the central subspace.*

Ordinary least squares

Theorem

Theorem 2.4 *Suppose that $Y|X$ follows the model $Y = g(\beta^T X) + varepsilon$ where $\beta = (\beta_1, \cdots, \beta_p)$ is a vector. Suppose that Assumption 2.1 holds. Then $E(XY)$ is proportional to β in the model.*

Ordinary least squares

- At the population level, the identifying procedure can be described as follows. First, standardize X to be $Z = \Sigma_x^{-1/2}(X - \mu)$. Identify a vector in $S_{Y|Z}$, then transfer back to $S_{Y|X} = \Sigma_x^{-1/2} S_{Y|Z}$. At the sample level, the estimating procedure is as follows.
- Step 1. Compute the sample mean and variance:

$$\hat{\mu} = E_n(X) \quad \hat{\Sigma}_x = var_n(X)$$

  and standardize X_i to be $\hat{Z}_i = \hat{\Sigma}_x^{-1/2}(X - \hat{\mu})$.
- Step 2. Center Y_i to be $\tilde{Y}_i = Y_i - E(Y)$ and it is estimated by $\hat{Y}_i = Y_i - E_n(Y)$.
- Step 3. Let $\hat{\gamma} = E_n(\hat{Z}\hat{Y})$ estimate $E(Z\tilde{Y}) \in S_{Y|Z}$.
- Step 4. Let $\hat{\beta} = \hat{\Sigma}_x^{-1/2}\hat{\gamma}$ estimate $E(XY) \in S_{Y|X}$.

Applications

- For generalized linear model: $Y = g(\beta^T X) + \epsilon$, single-index model $Y = g(\beta^T X) + \epsilon$, and transformation model $H(Y) = \beta^T X + \epsilon$, the above result shows that OLS can be applied to estimate $\gamma = \beta/\|\beta\|$ if the L_2-norm $\|\beta\|$ of β is not 1.
- As OLS has a close form and thus, estimating γ is very computational efficient under the above nonlinear models.
- After that, we can define $z = \gamma^T X$ that is one-dimensional, rewrite the model as $Y = g(\alpha z) + \varepsilon$ to estimate α in a one-dimensional setting.
- Thus, estimating $\beta = \alpha\gamma$ can be performed in this two-step procedure. (see Feng and Zhu 2013 CSDA)

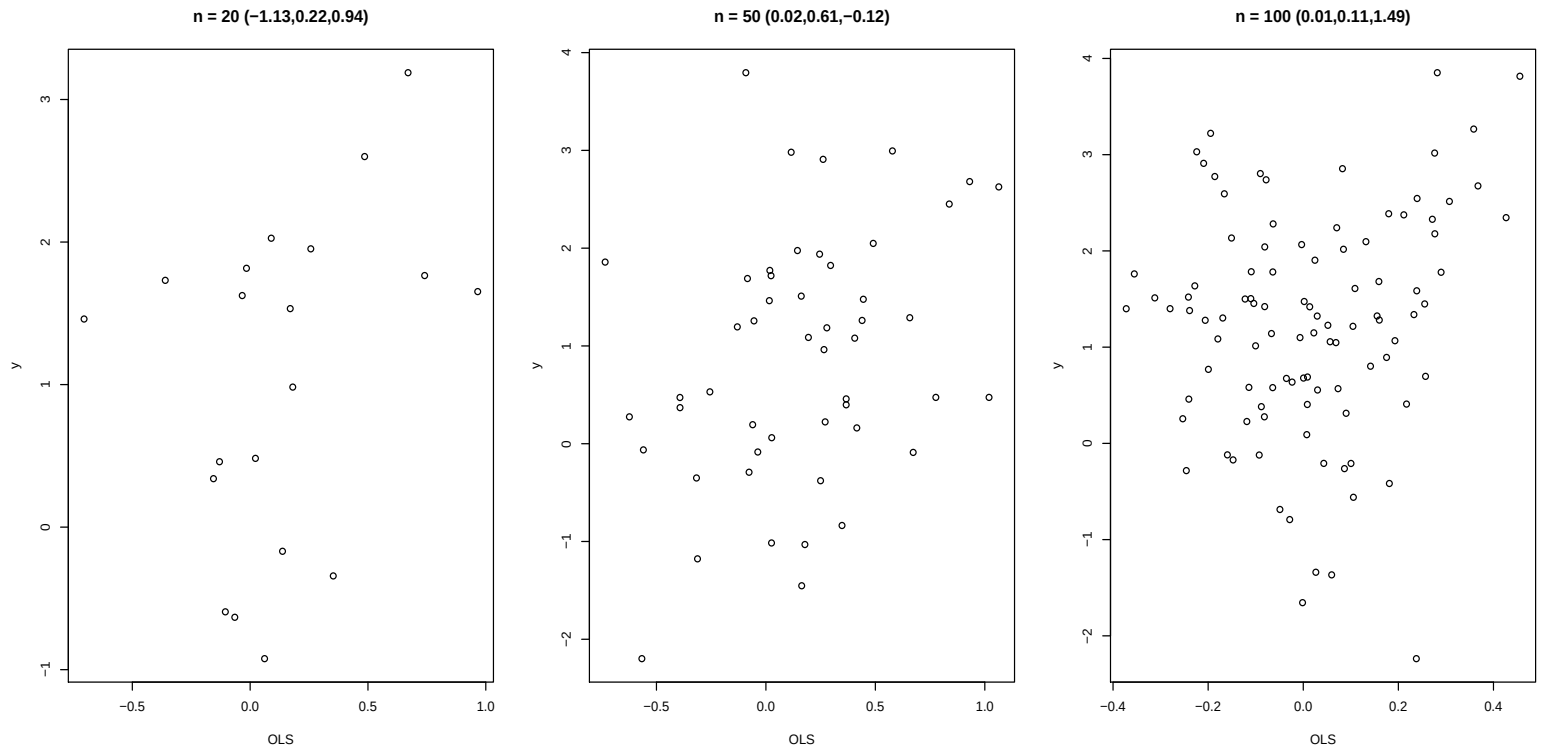
Part III: Principal Hessian Directions

- The biggest disadvantages of OLS: 1) it can only estimate at most one direction in the central space; 2) it cannot identify the direction in symmetric function such as the one in $Y = |\beta^T X| + \varepsilon$.

- For example, $Y = |X_1 + X_2| + \varepsilon$: $\beta = (1, 1, 0)$. But it cannot well identified by OLS.

OLS plot

Figure: OLS Scatter Plot



Principal Hessian Directions

- **Another method:** Consider the conditional mean $E(Y|X)$ of Y given X. When $E(Y|X) = E(Y|\beta^T X)$, its second derivative is $\partial^2 E(Y|X)/\partial(XX^T) = \beta\partial^2 E(Y|\beta^T X)/\partial(\beta^T XX^T\beta)\beta^T$ which is a $p \times p$ matrix.
- An application of Stein Lemma (1956). When the distribution of X is normal,

$$\begin{aligned} E\Big(\partial^2 E(Y|X)/\partial(XX^T)\Big) &= E(YXX^T) \\ = \quad \beta E\Big(\partial^2 E(Y|\beta^T X)/\partial(\beta^T XX^T\beta)\Big)\beta^T, \end{aligned}$$

- where $E\Big(\partial^2 E(Y|\beta^T X)/\partial(\beta^T XX^T\beta)\Big)$ is a $q \times q$ matrix.
- The $p \times p$ matrix $\beta E\Big(\partial^2 E(Y|\beta^T X)/\partial(\beta^T XX^T\beta)\Big)\beta^T$ has at most q non-zero eigenvalues and the corresponding eigenvectors will be proved to be in the central subspace.

Principal Hessian Directions

Assumption

Assumption 3.1 *Assume that the conditional variance*

$$var(X|\beta^T X) = C$$

is a $p \times p$ non-random matrix.

- This assumption is satisfied if X is multivariate normal.

Principal Hessian directions: 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 $E(Y) = 0$ and $E(X) = 0$ and thus $E(Y) - \alpha^T E(X) = 0$. That is why there is no constant term in e.

Definition

Definition 3.1. *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 Part III is that the column space of a Hessian matrix (either one) is a subspace of the central space.

Principal Hessian directions: population development

Theorem

Theorem 3.1. *Suppose that Assumptions 2.1 and 3.1 hold. Then the column space of H_1 is a subspace of $S_{Y|X}$.*

Theorem

Theorem 3.2 *Suppose that Assumptions 2.1 and 3.1 hold. Then the column space of H_2 is a subspace of $S_{Y|X}$.*

Sample estimator of pHd

- Again, we use the idea of first transforming to Z, estimating $S_{Y|Z}$, and then transforming back to $S_{Y|X}$. We summarize the computation into the following steps.
- Step 1. standardize $X_1, \cdots, X_n$ to be $\hat{Z}_1, \cdots \hat{Z}_n$, and center $Y_1, \cdots, Y_n$ to be $\hat{Y}_1, \cdots \hat{Y}_n$, as described in the algorithm for OLS.
- Step 2. Compute the OLS of $\hat{Y}_i$ versus $\hat{Z}_i$ to get $\hat{\alpha}$:

$$\hat{\alpha} = (var_n(\hat{Z}))^{-1} cov_n(\hat{Z}, \hat{Y}),$$

and

$$\hat{\alpha}_0 = E_n(\hat{Y}) - \hat{\alpha}^T E_n(\hat{Z}) = 0.$$

- Because of standardization, we have:

Sample estimator of pHd

- $var_n(\hat{Z}) = I_p$, and
  $cov_n(\hat{Z}, \hat{Y}) = E_n(\hat{Z}\hat{Y}) - E_n(\hat{Z})E_n(\hat{Y}) = E_n(\hat{Z}\hat{Y})$.
- This means the OLS for $\hat{Y}$ and $\hat{Z}$ is $\hat{\alpha} = E_n(\hat{Z}\hat{Y})$. The residual is $\hat{e}_i = \hat{Y}_i - \hat{\alpha}^T\hat{Z}_i$.
- Step 3. Construct the e-based and y-based Hessian matrix:

$$\hat{H}_1 = E_n(\hat{Y}\hat{Z}\hat{Z}^T), \quad \hat{H}_2 = E_n(\hat{e}\hat{Z}\hat{Z}^T).$$

- Step 4. Assume, for now, we know the structural dimension q. Let $\hat{\gamma}_1 \cdots \hat{\gamma}_q$ be the q eigenvectors corresponding to the q largest eigenvalues of $\hat{H}_1\hat{H}_1^T$ and let $\hat{\delta}_1 \cdots \hat{\delta}_q$ be the q eigenvectors corresponding to the q largest eigenvalues of $\hat{H}_2\hat{H}_2^T$. We use $\hat{\gamma}_1 \cdots \hat{\gamma}_q$ and $\hat{\delta}_1 \cdots \hat{\delta}_q$ as the estimators of $S_{Y|Z}$.

Sample estimator of pHd

- Let
$$\hat{\beta}_i = \hat{\Sigma}^{-1/2}\hat{\gamma}_i, \quad \hat{\eta}_i = \hat{\Sigma}^{-1/2}\hat{\delta}_i$$
- We then use $\hat{\beta}_1 \cdots \hat{\beta}_q$ and $\hat{\eta}_1 \cdots \hat{\eta}_q$ as the estimators of $S_{Y|X}$.
- We have assumed that the structural dimension q is known. In practice this must be determined by the data. There are several proposals in the literature.
- The following is the pHd scatter plot for the model $Y = |X_1 + X_2| + \varepsilon$, described before.

pHd plot

Figure: pHd Scatter Plot