Partial Least Squares

regression

Let $X = (X_1, X_2, \dots X_p)$ be

a random vector with values in $\mathbb{R}^p$.

and

$Y = (Y_1, \dots, Y_q)$ a r.v. $Y \in \mathbb{R}^q$

X : predictor

Y : response.

Regression problem :

Approximate : $\mathbb{E}(Y|X)$

(conditional expectation).

(2)

$$\mathbb{E}(Y|X) = f(X), \quad f : \mathbb{R}^p \to \mathbb{R}^q$$

$$f = (f_1, \dots, f_q), \quad f_i : \mathbb{R}^p \to \mathbb{R}.$$

<u>Linear model</u>:

$$f(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p,$$

with $\beta_i \in \mathbb{R}^q$, $i = 0, \dots, p$.

$$\Updownarrow$$

$$\begin{cases} f_1(X) = \beta_0^1 + \beta_1^1 X_1 + \dots + \beta_p^1 X_p \\ \vdots \\ f_q(X) = \beta_0^q + \beta_1^q X_1 + \dots + \beta_p^q X_p \end{cases}$$

Put $\beta = (\beta_0, \beta_1, \dots \beta_p)$

③ Least Squares estimator:

$$\boxed{\hat{\beta} = \arg\min_{\beta} \mathbb{E}\left(\left\| Y - \langle \tilde{X}, \beta \rangle_{\mathbb{R}^{q+1}} \right\|^2_{\mathbb{R}^q}\right)}$$

where $\tilde{X} = (1, X_1, \dots X_p)$.

and

$$\langle \tilde{X}, \beta \rangle_{\mathbb{R}^{q+1}} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p .$$

$$\boxed{\hat{\beta} = \arg\min_{\beta} \mathbb{E}\left(\sum_{i=1}^{q} \left(Y_i - f_i(x)\right)^2 \right)}$$

But

$$\min \mathbb{E}\left(\sum_{i=1}^{q} \left(Y_i - f_i(x)\right)^2 \right) = \sum_{i=1}^{q} \min \mathbb{E}\left(Y_i - f_i(x)\right)^2$$

$\Leftrightarrow$ q independent minimization problems

$(q = 1)$.

Remark : The estimation of the linear model by the least squares method does not take into account the (link) dependence between Y_i's!

5 Let us consider $q = 1$ $(Y \in \mathbb{R})$

and $\left.\begin{array}{l} \mathbb{E}(Y) = 0 \\ \mathbb{E}(X_i) = 0, \ i = 1, \dots, p \end{array}\right\} \Leftrightarrow \beta_0 = 0.$

Then

$$\boxed{\hat{\beta} = V^{-1} \cdot \mathbb{E}(X \cdot Y)}$$ where

$$V = \{v_{ij}\}_{1 \le i,j \le p}, \quad v_{ij} = \mathbb{E}(x_i \cdot x_j)$$

and

$$\mathbb{E}(X \cdot Y) = \begin{pmatrix} \mathbb{E}(X_1 \cdot Y) \\ \vdots \\ \mathbb{E}(X_p \cdot Y) \end{pmatrix} \in \mathbb{R}^p$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} \in \mathbb{R}^p$$

6

Denote by

$$\hat{Y} = X_1 \hat{\beta}_1 + \dots + X_p \hat{\beta}_p$$

$$= X^T V^{-1} \mathbb{E}(X \cdot Y)$$

the least square estimation of the conditional expectation in the linear model.

<u>Remark</u>

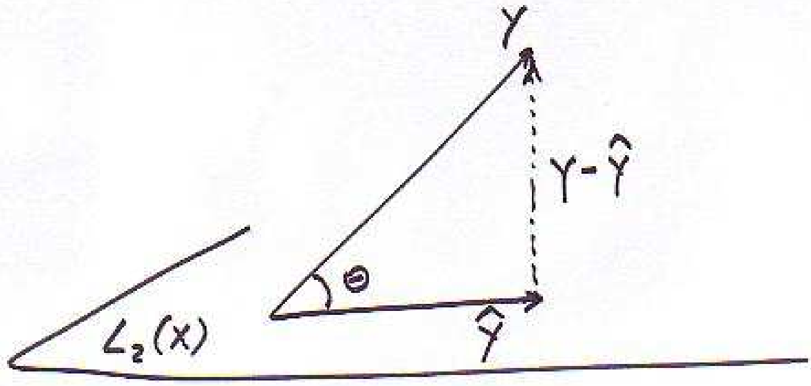
$P_X = X^T V^{-1} \mathbb{E}(X \cdot\ )$ is the orthogonal projector on the linear space spanned by $\{X_1, \dots, X_p\}$ with the dot product $\langle X_1, X_2 \rangle = \mathbb{E}(X_1 X_2)$

$$P_X(Z) = X^T V^{-1} \mathbb{E}(X \cdot Z) \quad \forall Z \in L_2(\Omega)$$

and $P_X^2(Z) = \mathbb{P}(Z)$.

Thus $\hat{Y} = \mathbb{P}_X(Y)$.

(7)



$$\varepsilon = y - \hat{y} \quad \text{(error)}$$

Analyse of variance :

$$\boxed{\mathbb{V}(y) = \mathbb{V}(\hat{y}) + \mathbb{V}(y-\hat{y})}$$

(1) and since $\mathbb{E}(y)=0$ and $\mathbb{E}(x_i)=0$:

$$\boxed{\mathbb{E}(y^2) = \mathbb{E}(\hat{y}^2) + \underline{\mathbb{E}(y-\hat{y})^2}}$$

objective function of the least squares criterion!

(2) $\boxed{\mathbb{E}(y-\hat{y})^2 = (1-R^2)\,\mathbb{E}(y^2)}$ $(R^2 = \cos^2(\Theta))$.

$$\mathbb{E}(Y-\hat{Y})^2 = (1-\underline{R^2})\,\mathbb{E}(Y^2)$$

where R^2 is the correlation coefficient between $\underline{Y}$ and $\underline{\hat{Y} = \beta_0 X_1 + \cdots + \beta_p X_p}$.

Thus <u>minimize</u> the least squares criterion is equivalent to <u>maximize</u> the value of $\underline{R^2}$!

9

Estimation from a sample of size n

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad \text{and}$$

$$\hat{Y} = X(X^T X)^{-1} X^T Y = X\hat{\beta}.$$

<u>Remark</u> if data is not centered then:

$$\hat{\beta}_0 = \bar{Y} - (\hat{\beta}_1 \bar{X}_1 + \cdots + \hat{\beta}_p \bar{X}_p) \quad \text{with}$$

$$\bar{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$

$$\bar{x}_i = \frac{x_{1i} + \cdots x_{ni}}{n}, \quad i = 1, \dots p.$$

Problems of the linear model estimation

- Multicollinearity of the predictor

$\exists\ \alpha_1, \alpha_2, \ldots, \alpha_p$ constants

$$\alpha_1 X_1 + \cdots + \alpha_p X_p = 0 \quad \text{p.s.}$$

It occurs

- by the nature of the X_i's
- when $n \leq p$

Consequence : V^{-1} does not exist!

$\Rightarrow$ Bad estimation of β_i's.

(11)

$$\boxed{V(\hat{\beta}) = \frac{\sigma^2}{n} \cdot V^{-1}}$$

where $\sigma^2 = \mathbb{E}(y-\hat{y})^2 = Var(y-\hat{y})$

residual variance.

$V(\hat{\beta})$ is then estimated by

$$\boxed{V(\hat{\beta}) \simeq \hat{\sigma}^2 . (X^T X)^{-1}}$$ with

$$\hat{\sigma}_2 = \frac{1}{n-p-1} \cdot \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

If multicollinearity :

$V(\hat{\beta}_i) \longrightarrow \infty$ when $\det(X^T X) \to 0$

$\Rightarrow$ student test for $\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$

<u>is not significant!</u>

(12)

Remark The student test for

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$$

is a test for the contribution of the variable $\underline{X_i}$, conditionaly to $\underline{X_1, \ldots X_{i-1}, X_{i+1} \ldots X_p}$ to the prediction of Y

Obviously, if multicollinearity

contribution of $X_i \simeq 0$.

Let see an example:

"The price of the cars."

13

How to detect the multicollinearity?

- **Principal Component Analysis**

Find linear combinations of X_i's of maximal variance.

$$\boxed{C = u_1 X_1 + \cdots + u_p X_p} \quad , \quad u = \begin{pmatrix} u_1 \\ \vdots \\ u_p \end{pmatrix} \in \mathbb{R}^p$$

such that

$$\boxed{u = \underset{\substack{u \in \mathbb{R}^p \\ \|u\| = 1}}{\arg\max} \; V(X^T u) = V(C).}$$

u is the eigen vector associated to the largest eigen value of matrix of variance-covariance of X's

$$\underline{Vu = \lambda u}$$

Let $\{(\lambda_1, u_1), \dots, (\lambda_p, u_p)\}$ be the set of eigen values/vectors of the covariance matrix $\mathbb{V}$, such that

$$\lambda_1 \geq \lambda_2, \geq \dots \geq \lambda_p \geq 0.$$

Let also define the i^{th} principal component by

$$\boxed{c_i = X^T \cdot u_i = X_1 \cdot u_{1,1} + \dots + X_p \cdot u_{1,p}}$$

Then, it is well known that :

1) $\lambda_1 + \lambda_2 + \dots + \lambda_p = \sum_{i=1}^{p} \mathbb{V}(x_i) \; (= p)$

scaled data

2) $\begin{cases} \mathbb{V}(c_i) = \lambda_i \,, & \forall i = 1, \dots, p \\ E(c_i) = 0 \end{cases}$

3) $X = c_1 u_1 + c_2 \cdot u_2 + \dots + c_p \cdot u_p$

(expansion formula).

By 2), if there exists $k \in \{1, \dots, p\}$ such that

$$\lambda_k \simeq 0$$

then

$$V(c_i) \simeq 0 \iff c_i \simeq 0$$

$$\Updownarrow$$

$$\underbrace{X_1 u_{k,1} + X_2 u_{k,2} + \dots + X_p u_{k,p} \simeq 0}_{\text{multicollinearity}}.$$

See on the example. (car example).

Solutions to multicolinearity for linear model stimation

Remark that, because of the expansion formula 2), the linear space spanned by X_i's is the same as that spanned by the principal components c_i's.

Principal component Regression (PCR)

Regression of Y using the principal components with variance > 0.

$$\boxed{\hat{Y} = \gamma_1 c_1 + \gamma_2 c_2 + \dots + \gamma_q c_q}$$

(17) Notice that

$$\boxed{V(\hat{\gamma}_i) = \frac{\sigma^2}{n} \cdot \frac{1}{\lambda_i}}$$ and

$$\boxed{\hat{\beta}_j = \hat{\gamma}_1 \cdot u_{1,j} + \hat{\gamma}_2 \cdot u_{2,j} + \dots + \hat{\gamma}_q \cdot u_{q,j}}$$

$$\boxed{V(\hat{\beta}_j) = \frac{\sigma^2}{n} \cdot \sum_{i=1}^{q} \frac{u_{ij}^2}{\lambda_i}}$$

Remark that small values of $\lambda_i \simeq$

$\simeq$ multicollinearity

$\Downarrow$

explosion of the variance of $\hat{\beta}_j$'s.

Idea: Keep only the principal comps. with large variances (λ_i's).

But: the least squares criterion maximizes R^2 and the most explanatory c.p. are not necessarily correlated to Y!

(18) Since the c_i's are uncorrelated

$$R^2(Y, \underbrace{\gamma_1 c_1 + \cdots + \gamma_q c_q}_{\hat{Y}}) = \underline{R^2(Y, c_1)} + \cdots + \underline{R^2(Y, c_q)}.$$

- See the example on cars.

  - use stepwise selection of p.c.'s

  - use the first q components selected by cross-validation procedure (package PLS, function pcr)

19

So, ideally, one wants "principal components" which:

- are highly correlated with Y (R^2)
- explain large amount of information from X. (λ)

This is not always the case and a compromise between the goodness of fit (R^2) and the stability of the regression coefficients ($V(\hat{\beta}_j)$) must be done!

It is a difficult task!

- The PLS regression builds such components maximizing the covariance instead the R^2.

20 Let look for a component

$$t = u_1 X_1 + u_2 X_2 + \cdots + u_p X_p$$

Such that

$$\boxed{\mathrm{Cov}^2(t, Y) \text{ is maximized}}$$

Tucker criterion

under the constraint $\sum_{i=1}^{p} u_i^2 = 1$

Remark that, because

$$\boxed{\mathrm{Cov}^2(t,Y) = R^2(t,Y) \cdot V(t) \cdot V(Y)}$$

one maximizes simultaneously both

$R^2(Y,t)$ and $V(t)$

$\downarrow$ link with Y $\qquad$ $\downarrow$ link with X

This component which maximizes a such criterion (Tucker) is called

first PLS component

(21) Let denote by t_1 this <u>first PLS component</u>:

$$t_1 = X_1 \cdot u_{11} + X_2 \cdot u_{12} + \dots + X_p u_{1p}$$

The weights $\{u_{1i}\}_{i=1\dots p}$ which satisfy the Tucker criterion are given by:

$$u_{1,i} = \frac{\mathbb{E}(Y \cdot X_i)}{\sqrt{\sum_{j=1}^{p} \mathbb{E}^2(Y \cdot X_j)}}$$

Of course, t_1 "explains" $\underline{X}$, in general, <u>less</u> than the first principal component c_1, but <u>is better</u> for the prediction of $\underline{Y}$:

de Jong (1993): $\underline{R^2(Y, t_1) \geq R^2(Y, c_1)}$

(22) We are looking now for a <u>second PLS component</u>, $\underline{t_2}$.

For this purpose we perform the following simple linear regressions:

Put $X^{(0)} = X$ and $Y^{(0)} = Y$

$$\begin{cases} X_i^{(0)} = \underline{p_{1,i}}\, t_1 + \underbrace{X_i^{(1)}}_{\text{residual}} , \quad i = 1, \dots, p. \\ Y^{(0)} = \underline{c_1 t_1} + \underbrace{Y^{(1)}}_{\text{residual}} \end{cases}$$

t_2 is built in the same way as t_1 but using $\underbrace{X^{(1)} \text{ and } Y^{(1)}}_{\text{residuals after regression on } t_1}$.

$$\begin{cases} X^{(1)} = (X_1^{(1)}, X_2^{(1)}, \dots X_p^{(1)}) \\ Y^{(1)} \end{cases}$$

23

$$t_2 = u_{21} X_1^{(1)} + u_{22} X_2^{(1)} + \cdots + u_{2p} X_p^{(1)}$$

with $u_2 = (u_{21}, \ldots u_{2p})$ such that the Tucker criterion for $X^{(1)}$ and $Y^{(1)}$

$$\boxed{\max \operatorname{Cov}^2\left(t_2, Y^{(1)}\right) = \max \operatorname{Cov}^2\left(\sum_{i=1}^{p} u_{2i} X_i^{(1)}, Y^{(1)}\right)}$$

$$u_{2i} = \frac{\mathbb{E}\left(Y^{(1)} \cdot X_i^{(1)}\right)}{\sqrt{\sum_{j=1}^{p} \mathbb{E}^2\left(Y^{(1)} \cdot X_j^{(1)}\right)}}, \quad i = 1, \ldots, p.$$

Notice that:

$$\left\{\begin{array}{ll} \cdot\ X_i^{(1)} \perp t_1 & \left(\mathbb{E}\left(X_i^{(1)} \cdot t_1\right) = 0\right) \\ \cdot\ Y^{(1)} \perp t_1 & \left(\mathbb{E}\left(Y^{(1)} t_1\right) = 0\right) \end{array}\right.$$

$$\to \quad t_1 \perp t_2 \qquad \left(\mathbb{E}(t_1, t_2) = 0\right).$$

Remark also that

- t_1 is linear combination of X_i's
- t_2 is linear combination of $X_i^{(1)}$'s <u>but</u> also of X_i's

$$t_2 = \sum_{i=1}^{p} u_{2i} X_i^{(1)} = \underbrace{\sum_{i=1}^{p} u_{2i} \cdot (X_i - p_{1,i} t_1)}_{\text{linear comb of } X_i\text{'s.}}$$

In the similar way as for t_1, we compute new residuals using t_2:

$$\begin{cases} X_i^{(1)} = \underline{p_{2,i} t_2} + \underbrace{X_i^{(2)}}_{\text{residual.}} \\ Y^{(1)} = \underline{r_2 t_2} + \underbrace{Y^{(2)}}_{\text{residual.}} \end{cases}$$

The computation procedure of t_h and of $X^{(h)}$ and $Y^{(h)}$ is called the h^{th} step of the PLS regression of Y on X. (In our example $h=2$).

Expansion formulas and linear approximation

$h=2$:

$$\begin{cases} X_i = p_{1,i}\, t_1 + X_i^{(1)} \\ Y = c_1 t_1 + Y^{(1)} \end{cases} \quad \text{and} \quad \begin{cases} X_i^{(1)} = p_{2,i}\, t_2 + X_i^{(2)} \\ Y^{(1)} = c_2 t_2 + Y^{(2)} \end{cases}$$

$$\begin{cases} X_i = \underbrace{p_{1,i}\, t_1 + p_{2,i}\, t_2} + \underbrace{X_i^{(2)}} \\ Y = \underbrace{c_1 t_1 + c_2 t_2}_{\substack{\text{prediction} \\ f(X)}} + \underbrace{Y^{(2)}}_{\text{residual}} \end{cases}$$

26

PLS regression is an iterative method.

At step $\underline{h}$ we have:

$\{t_1, t_2, \dots, t_h\}$: the PLS components (the first h)

⚠ $\forall i \neq j$ t_i and t_j are uncorrelated!

We have also the following expansion formulas:

$$\begin{cases} X_i = p_{1i} t_1 + p_{2,i} t_2 + \dots + p_{h,i} t_h + X_i^{(h+1)} \\ Y = \underbrace{c_1 t_1 + c_2 t_2 + \dots + c_h t_h}_{\text{PLS prediction at step } h} + \underbrace{Y^{(h+1)}}_{\text{residual}} \end{cases}$$

A question: how large is h?

- of course, because $\{t_i\}_{i=1,\dots,h}$ are uncorrelated

$$\underline{h \leq \dim(L(X))}$$

More precisely, if $L_Y(X)$ is the smallest subspace of $L(X)$ which contains $\hat{Y} = P_X(Y)$, then

$$h \leq \dim(L_Y(X))$$

In practice, h is chosen by cross-validation

Root Mean Squared Error of Prediction: (RMSEP)

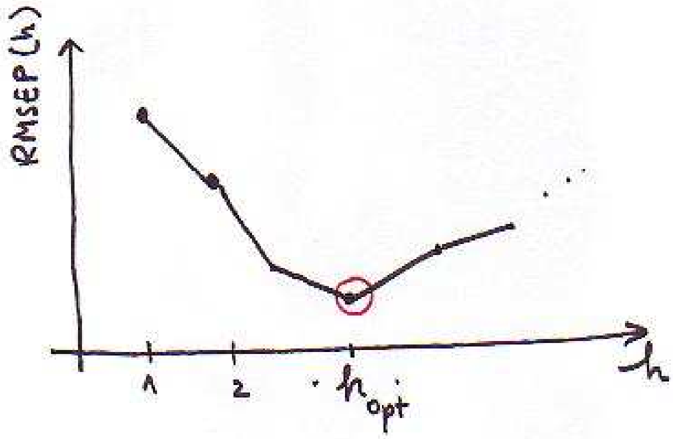
$$RMSEP = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

In general $\hat{Y}_i$ is computed with the model built without the observation "i" (Leave-one-out cross-validation).

For fixed h, one has

RMSEP(h).

(28) One can chose h which minimizes the function RMSEP(h)



Tenenhaus proposed to stop the iterative process when the $(h+1)$ PLS component does not contribute significantly to the prediction:

- if $\alpha \in (0,1)$ - typically 0.9 or 0.95, the process stops at step $\underline{h}$ if

$$RMSEP(h+1) \geq \alpha \cdot RMSEP(h)$$

$$\Leftrightarrow \boxed{\frac{RMSEP(h) - RMSEP(h+1)}{RMSEP(h)} \leq \underline{1-\alpha}.}$$

(the relative contribution is less than $1-\alpha$).