The Elements of Statistical Learning

Ch3: Liner Methods for Regression - PLS and PCR

Philip Lin

Data Science in Hsinchu

2015.09.02

DSHC is a non-profit studying group.

This slide is created to help us to discuss the content of "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning" and is not used in any profit-oriented activity.

Framework

Today, we talk about (traditional, not state of the art) prediction methods in multicollinearity problem.

- Understanding eigenvalue and eigenvector
- Principal Component Regression (Massy 1965)
  - Some cautionary notes on PCR (Hadi, 1998)
- Partial Least Square (Wold, 1975; Paul H. Garthwaite, 1994)
- an unified framework for OLS, PCR, PLS, and RR (Stone and Brooks, 1990; Frank and Friedman, 1993)

Keywords:

Ordinary Least Square (OLS), Principal Component Regression (PCR), Partial Least Square (PLS), Ridge regression (RR), shrinkage factor.

Understanding eigenvalue and eigenvector

Questions

- What are eigenvalue and eigenvector ?
- Traditional motivation for PCA:

  choose δ according to

  $$\max_{\delta:\|\delta\|=1} \mathsf{Var}(\delta^T X) = \max_{\delta:\|\delta\|=1} \delta^T \mathsf{Var}(X)\delta$$

  then, the direction δ is given by the eigenvector corresponding to the largest eigenvalue of the covariance matrix.
- Why we always use first few components with largest eigenvalues ? is there a more intuitive explanation ?

$\boxed{\text{Theorem}}$ (Karhunen Loeve decomposition)

$X(t)$ is a random process, $X(t) \in L^2$ and
$EX(t) = 0, \quad \text{cov}(X(s), X(t)) = EX(s)X(t), \quad \forall t \in [a, b]$
then

$$X(t) = \sum_{k=1}^{\infty} Z_k e_k(t) \ \text{ converge in } L^2 \text{ and uniformly in } t$$

$$Z_k = \int_a^b X(t) e_k(t) \mathrm{d}t \sim (0, \lambda_k)$$

where

- Z_k's are uncorrelated random variables
- $e_k(t)$'s are continuous real-valued function on $[a, b]$ and pairwise orthogonal on $[a, b]$
- usually, λ_k and $e_k(t)$ are called eigenvalue and eigenfunction
- of all such approximations, the KL approximation is the one that minimizes the total mean square error (provided we have arranged the eigenvalues in decreasing order)

Remarks:

- compare to Taylor expansion of $f(x)$ on point a

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x - a)^n$$

PCA is a special case of Karhunen Loeve

- Let $\boldsymbol{x} = \{x_1, \ldots, x_p\}$ be $p-$dimensional random vector

  (note that $\boldsymbol{x}$ is regarded as a random process if p goes to infinity)
- $A = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_m)$: eigenvectors of $\Sigma_{\boldsymbol{x}}$, $m < p$
- $\boldsymbol{z} = (z_1, z_2, \ldots, z_m)'$, $z_k = \boldsymbol{x}'\boldsymbol{e}_k$

then, by Karhunen-Loeve

$$\begin{aligned}\|\boldsymbol{x} - \hat{\boldsymbol{x}}_m\|^2 =& \|\boldsymbol{x} - \sum_{k=1}^{m} z_k \boldsymbol{e}_k\|^2 \\ =& \|\boldsymbol{x} - A\boldsymbol{z}\|^2 \\ \geq& \|\boldsymbol{x}\|^2 - \|A\boldsymbol{z}\|^2\end{aligned}$$

where

$$\|A\boldsymbol{z}\|^2 = (A\boldsymbol{z})'(A\boldsymbol{z}) = \boldsymbol{z}'A'A\boldsymbol{z} = \boldsymbol{z}'\boldsymbol{z} = \mathrm{Var}(\boldsymbol{z})$$

thus, KL minimize $\|\boldsymbol{x} - \hat{\boldsymbol{x}}_m\|^2$ (also maximize $\mathrm{Var}(\boldsymbol{z})$).

Remarks

- That is, KL criterion (with finite dimensional random vector case) is equivalent to the criterion in PCA. (check with page 3)
- whet does it means for $m = p$?

Principal Component Regression

Suppose that the columns of $\boldsymbol{X}$ are highly multicollinear, but we want to keep all the varaibles in $\boldsymbol{X}$.

steps for PCR

1. compute the standardized version of $\boldsymbol{X}_{n\times p}$ and denote it as $\boldsymbol{Z}$
2. compute the principal components:
   Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ be the eigenvalues of $\boldsymbol{Z}^T\boldsymbol{Z}$ and $\boldsymbol{V}$ be the corresponding eigenvectors. Let $\boldsymbol{W} = \boldsymbol{Z}\boldsymbol{V}$
3. regress Y on the first m PCs, $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_m$, where $m \leq p$

Reason for using PCR

- $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_m$, are orthogonal, the problem of multicollinearity disappears completely.
- All of the variables in $\boldsymbol{X}$ are used.
- PCR presumably improves the prediction accuracy. (we will explain it later.)

Hadi, A.S. & Ling, R.F., 1998. Some Cautionary Notes on the Use of Principal Components Regression. The American Statistician, 52(1), pp.15-19.

causionary note: Hold's Data ($p = 4$ with response $\boldsymbol{U}$)

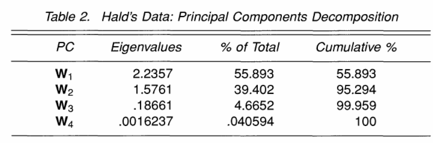
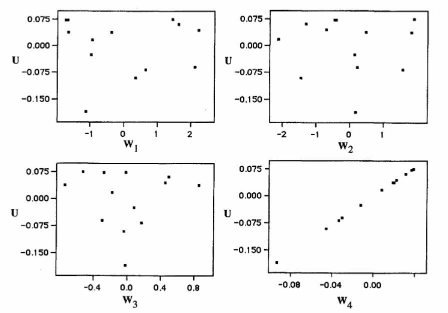


Table 2. Hald's Data: Principal Components Decomposition

PC	Eigenvalues	% of Total	Cumulative %
$\mathbf{W}_1$	2.2357	55.893	55.893
$\mathbf{W}_2$	1.5761	39.402	95.294
$\mathbf{W}_3$	.18661	4.6652	99.959
$\mathbf{W}_4$	.0016237	.040594	100



For usual linear regression $Y = \boldsymbol{X}\beta + \epsilon$, we denote

$\boldsymbol{Z}$ is standardization of $\boldsymbol{X}$

$\boldsymbol{V}$ is the eigenvector of $\boldsymbol{Z}^T\boldsymbol{Z}$

Let $\boldsymbol{W} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_p)$, $\boldsymbol{W} = \boldsymbol{Z}\boldsymbol{V}$

then

$$\begin{aligned} Y =& \boldsymbol{Z}\beta + \epsilon \\ =& \boldsymbol{Z}\boldsymbol{V}\boldsymbol{V}^T\beta + \epsilon \quad (\text{ where } \boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}) \\ =& \boldsymbol{W}\theta + \epsilon \end{aligned}$$

where $\theta = \boldsymbol{V}^T\beta$

if the true vector of regression coefficient β is in the direction of the jth eigenvector of $\boldsymbol{Z}^T\boldsymbol{Z}$, i.e.,

$$\boldsymbol{V}_j = \alpha\beta, \quad \alpha \text{ is an nonzero scalar}$$

then

$$\begin{aligned} \theta_j =& \boldsymbol{V}_j^T\beta = \alpha\beta^T\beta \\ \theta_k =& \boldsymbol{V}_k^T\beta = \frac{1}{\alpha}\boldsymbol{V}_k^T\boldsymbol{V}_j = 0, \quad \forall k \neq j \end{aligned}$$

- the jth PC, $\boldsymbol{W}_j$, alone will contribute everything to the fit while the remaining PCs will contribute nothing.

Question

- ▷ Although PCR is able to construct a set of orthogonal surrogate for predictions, this construction depends only on the information of design matrix and does not guarantee to improve the regression fit.
- ▷ Is it possible to take the information of repsonse into consideration in PCR ? or find a construction which satisfies
  1. predictors are uncorrelated.
  2. the construction wll acturally improves the regression fit.

Partial Least Square

$$\hat{Y} = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \cdots + \beta_p T_p$$

- $Y, X_1, \ldots, X_m$
- each component T_k is a linear combination of the X_j's and the sample correlation for any pair of components in 0.
- Y and X_j are centered to give U_1 and V_{1j}
  for $j = 1, \ldots, m$

$$\begin{aligned} U_1 =& Y - \bar{y} \\ V_{1j} =& X_j - \bar{x}_j \end{aligned}$$

Garthwaite, P.H., 1994. An Interpretation of Partial Least Squares. Journal of the American Statistical Association, 89(425), p.122.

We construct the components sequentially

1. The first component, T_1, is intended to be useful for predicting U_1 and is constructed as a linear combination of V_{1j}'s
   (a) Regress U_1 against each V_{1j} in turn: $\hat{U}_{1j} = b_{1j}V_{1j}, \quad j = 1, \dots, m$
   (b) take a weighted average over these univariate fit.

$$T_1 = \sum_{j=1}^{m} w_{1j} b_{1j} V_{1j}$$

   note that $\sum_j w_{1j} = 1$ is not essential.

2. The information in V_1 that is not in T_1 may be estimated by the residuals from a regression of V_1 on T_1 (i.e., $\hat{V}_{1j} = d_{1j}T_1$)

$$V_{2j} = V_{1j} - d_{1j}T_1, \quad j = 1, \dots, m$$

   The variability in Y that is not explained by T_1 can be estimated by residuals from a regression of U_1 on T_1 (i.e., $\hat{U}_{1j} = g_{1j}T_1$)

$$U_{2j} = U_{1j} - g_{1j}T_1, \quad j = 1, \dots, m$$

   then

$$\hat{U}_{2j} = b_{2j}V_{2j} \quad \text{and} \quad T_2 = \sum_{j=1}^{m} w_{2j} b_{2j} V_{2j}$$

3. repeat 1 and 2 to construct $T_1, \dots, T_p$

Remarks

- $V_{(i+1)j}$ is uncorrelated with T_i for all j (Because the residuals from a regression are uncorrelated with a regressor)
- each of components $T_{i+1}, \ldots, T_p$ is a linear combination of $V_{(i+1)j}$'s so they are uncorrelated with T_i
- When to stop ? (cross validation... etc.). Usually, we only need the first few components, that is, $p < m$. When $p = m$, OLS is obtained.
- Compared to PCR, we also construct uncorrelated predictors to fit the regression and our contruction use the information of response.
- Note that this interpretation is not the original idea of PLS, however, the interpretation of Garthwaite is pretty closed to the statistican.

An unified framework

$$\{y_i, x_{1i}, \ldots, x_{pi}\}_1^N$$

Let variables are standardized

$$y \leftarrow (y - \bar{y})/[\text{ave}(y - \bar{y})^2]^{\frac{1}{2}}$$

$$x_k \leftarrow (x_k - \bar{x}_k)/[\text{ave}(x_k - \bar{x}_k)^2]^{\frac{1}{2}}$$

$$\text{ave}(\eta) = \frac{1}{N}\sum_{i=1}^{N} \eta_i$$

consider

$$V = \text{ave}(\boldsymbol{x}\boldsymbol{x}^T) = \sum_{k=1}^{p} e_k^2 \boldsymbol{v}_k \boldsymbol{v}_k^T$$

$\{e_k^2\}_1^p$ are the eigenvalues of V arranged in $e_1 \geq e_2 \geq \cdots \geq e_p$

$\{\boldsymbol{v}_k\}_1^p$ are their corresponding eigenvectors.

Frank, I.E. & Friedman, J.H., 1993. A Statistical of Some Chemometrics View Regression Tools. Technometrics, 35(2), pp.109-135.

two-step process

OLS

$$\boldsymbol{c}_{\mathsf{OLS}} = \underset{\boldsymbol{c}^T\boldsymbol{c}=1}{\arg\max}\ \ \mathsf{corr}^2(y, \boldsymbol{c}^T\boldsymbol{x})$$

the OLS solution is then a simple least square regression of y on $\boldsymbol{c}_{\mathsf{OLS}}^T\boldsymbol{x}$

PCR

$$\boldsymbol{c}_k(\mathsf{PCR}) = \underset{\substack{\{\boldsymbol{c}^T V\boldsymbol{c}_l=0\}_1^{k-1}\\ \boldsymbol{c}^T\boldsymbol{c}=1}}{\arg\max}\ \ \mathsf{var}(\boldsymbol{c}^T\boldsymbol{x})$$

the Kth PCR model is given by a least square regression of the response on the K linear combinations $\{\boldsymbol{c}_k^T\boldsymbol{x}\}_1^K$

PLS

$$\boldsymbol{c}_k(\mathsf{PLS}) = \underset{\substack{\{\boldsymbol{c}^T V\boldsymbol{c}_l=0\}_1^{k-1}\\ \boldsymbol{c}^T\boldsymbol{c}=1}}{\arg\max}\ \ \mathsf{corr}^2(y, \boldsymbol{c}^T\boldsymbol{x})\mathsf{var}(\boldsymbol{c}^T\boldsymbol{x})$$

the Kth PLS model is given by a least square regression of the response on the K linear combinations $\{\boldsymbol{c}_k^T\boldsymbol{x}\}_1^K$

Stone, M. & Brooks, R.J., 1990. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. Royal Statistical Society. Series B, 37(12), p.1130.

RR

$$\boldsymbol{c}_{\text{RR}} = \arg\max_{\boldsymbol{c}^T\boldsymbol{c}=1} \ \text{corr}^2(y, \boldsymbol{c}^T\boldsymbol{x}) \frac{\text{var}(\boldsymbol{c}^T\boldsymbol{x})}{\text{var}(\boldsymbol{c}^T\boldsymbol{x}) + \lambda}$$

the ridge solution is then taken to be a ridge regression of y on $\{\boldsymbol{c}_{\text{RR}}^T\boldsymbol{x}\}$

$$\hat{y}_{\text{RR}} = \left[\frac{\text{ave}(y\boldsymbol{c}_{\text{RR}}^T\boldsymbol{x})}{\text{ave}(y\boldsymbol{c}_{\text{RR}}^T\boldsymbol{x}) + \lambda}\right] \boldsymbol{c}_{\text{RR}}^T\boldsymbol{x}$$

Remarks

- The criterion associated with RR, PCR and PLS all involve the scale of $\boldsymbol{c}^T\boldsymbol{x}$ through its sample variance, thereby producing biased estimates.
- The effect of this bias is to pull the solution coefficient vector away from the OLS solution toward directions in which the projected data (predictors) have larger spread, i.e., larger var$(\boldsymbol{c}^T\boldsymbol{x})$.

for PCR and PLS

- ▷ the degree of bias is introduced by K. when $K = R$ (rank of V), one obtains an unbiased OLS solution. For $K < R$, bias is introduced. The smaller the value of K, the larger the bias.
- ▷ In PCR, we constrain $\boldsymbol{c}$ to lie in the subspace spanned by the first K eigenvectors of V which places a lower bound on the sample variance of $\boldsymbol{c}^T\boldsymbol{x}$,

  $$\mathrm{var}(\boldsymbol{c}^T\boldsymbol{x}) \geq e_K^2$$

  Since the eigenvectors are ordered on decreasing values of e_K^2, increasing K has the effect of easing this restriction., thereby reducing the bias.
- ▷ In PLS, there is no sharp lower bound on $\mathrm{var}(\boldsymbol{c}^T\boldsymbol{x})$ for a given K.

for RR

- ▷ $\lambda > 0$ introduces increasing bias toward larger values of $\mathrm{var}(\boldsymbol{c}^T\boldsymbol{x})$ and increased shrinkage of the length of the solution coefficient vector.
- ▷ The control of degree of bias in RR is more continuous and smoother than that in PCR and PLS.

Questions

When can these estimators substantially improve (prediction) performance and which one can do it best ?

consider a highly idealized situation with i.i.d. homoscendastic error

$$y = \boldsymbol{\alpha}^T \boldsymbol{x} + \epsilon, \qquad \epsilon \sim (0, \sigma^2)$$

and we further assume

- ▷ the predictors are uncorrelated

  $$V = \mathsf{diag}(e_1^2, \ldots, e_p^2)$$

  e_j^2 is the sample predictor variance

- ▷ $\boldsymbol{a}$ is an estimate of $\boldsymbol{\alpha}$. Then the MSE of prediction at $\boldsymbol{x}$ is

  $$\mathsf{MSE}[\hat{y}(\boldsymbol{x})] = E_\epsilon[\boldsymbol{\alpha}^T \boldsymbol{x} - \boldsymbol{a}^T \boldsymbol{x}]^2$$

- ▷ In the view of Bayesian, consider all coefficient vector directions $\boldsymbol{\alpha}/|\boldsymbol{\alpha}|$ equally likely; that is, the prior distribution depends only on its norm $\boldsymbol{\alpha}^T \boldsymbol{\alpha}$

  $$\pi(\boldsymbol{\alpha}) = \pi(\boldsymbol{\alpha}^T \boldsymbol{\alpha})$$

- ▷ consider a simple linear shrinkage of the form

  $$a_j = f_j \hat{\alpha}_j, \quad j = 1, \ldots, p$$

  where $\hat{\boldsymbol{\alpha}}$ is the OLS estimate and the $\{f_i\}_1^p$ are shrinkage factors taken to be independent of the sample response values.

In this case, the mean squared prediction error becomes

$$\begin{aligned}\mathsf{MSE}[\hat{y}(\boldsymbol{x})] =& E_{\boldsymbol{\alpha}} E_{\epsilon}\left[\sum_{j=1}^{p}(\alpha_j - f_j\hat{\alpha}_j)x_j\right]^2 \\ =& \sum_{j=1}^{p}\left[(1-f_j)^2 E_{\boldsymbol{\alpha}}|\boldsymbol{\alpha}|^2\frac{1}{p} + f_j^2\frac{\sigma^2}{Ne_j^2}\right]x_j^2\end{aligned}$$

- The first term depends on true $\boldsymbol{\alpha}$ and is independent of the error vairance or the predictor distribution. (squared bias of estimate)
- The second term is independent of the nature of the true $\boldsymbol{\alpha}$ and depends on error variance and predictor design. (variance of estimate)
- Setting $\{f_j = 1\}_1^p$ yields the OLS, which is unbiased.

  Reducing any (or all) of the $\{f_j\}_1^p$ to a value less than 1 causes an increase in bias but decrease the variance.

  Setting any (or all) of the $\{f_j\}_1^p$ to a value greater than 1 increase both the bias squared and the variance.
- Directions with small spread in the predictor variables give rise to high variance in the model estimate. (see e_j^2)

The values of $\{f_j\}_1^p$ that minimize the MSE are

$$f_j^* = \frac{e_j^2}{e_j^2 + \lambda}, \quad j = 1, \ldots, p$$

with

$$\lambda = \frac{p}{N} \frac{\sigma^2}{E_{\boldsymbol{\alpha}} |\boldsymbol{\alpha}|^2}$$

and optimal estimates

$$a_j = \hat{\alpha}_j \cdot \frac{e_j^2}{e_j^2 + \lambda}$$

- ▷ The degree of improvement (over OLS) will increase with decreasing signal-to-noise ratio and training-sample size and increasing collinearity as reflected by the disparity in the eigenvalues of the predictor covariance matrix.
- ▷ Under equal direction prior, the optimal estimate among all linear shrinkage estimators is ridge. (note that PLS is not a linear shrinkage estimator so it cannot be compared here, but PCR is. $a_j(\text{CPR}) = \hat{\alpha}_j I(e_j^2 \geq e_k^2))$
- ▷ The set of situations that favor PLS and PCR would involve $\boldsymbol{\alpha}$'s that have small projections on the subspace spanned by the eigenvectors corresponding to the smallest eigenvalues.

In the discussion of Stone and Brooks 1990, Frank and Friedman said

> *"A large simulation study (Frank, 1989) shows that RR, PLS and PCR behave quite similarly, all with vastly superior performance to OLS, and with RR dominating PLS and PCR (sometimes only slighly) in all situations considers."*

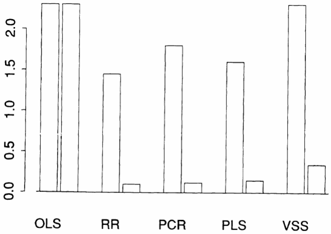


Figure 8. Performance Comparisons Conditioned on Low and High Collinearity Situations.

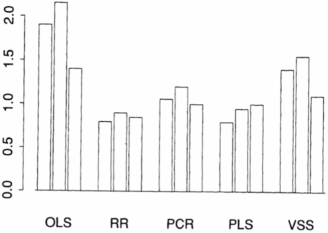
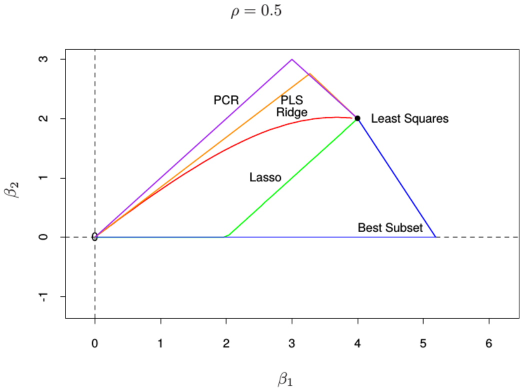


Figure 10. Performance Comparisons Conditioned on High, Medium, and Low Signal-to-Noise Ratio.

Frank, I.E. & Friedman, J.H., 1993. A Statistical of Some Chemometrics View Regression Tools. Technometrics, 35(2), pp.109-135.



The figure is cited from "Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning"

The end