Hastie, Tibshirani and Friedman (ESL), Section 3.5.1

☛ The principal components (PCs) are also often used as a dimension reduction tool to perform other statistical analyses.

☛ For example, it is frequently used in regression. It is also used in classification and clustering.

☛ In these problems, the analyses are performed with respect to s subset of the PCs of the data, instead of the original data.

☛ Consider a linear regression problem for a response variable Z and a p-vector X:

$$Z = m(X) + \epsilon,$$

where

- $m(x) = E(Z|X = x) = \alpha + \beta^T x$.
- ϵ is a random error with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$.
- X and ϵ are independent.

☛ For simplicity, assume that Z and X are **centered**:

$$E(X) = 0, \quad Z = m(X) + \epsilon, \quad m(x) = E(Z|X = x) = \beta^T x, \tag{1}$$

where $\beta \in \mathbb{R}^p$ is unknown $\Rightarrow$ no intercept α in the model.

☛ $m(x)$ is called the "regression function", or the "mean function".

☛ From (1), suppose we observe independent and identically distributed (i.i.d.) data

$$(X_1, Z_1), \ldots, (X_n, Z_n).$$

☛ Goal: estimate β

$\Rightarrow$ provides an estimator of the regression function $m(\cdot)$
$\Rightarrow$ makes prediction of Z for new values of X.

☛ The least squares (LS) estimator of β is

$$\hat{\beta}_{LS} = \text{argmin}_{b \in \mathbb{R}^p} \sum_{i=1}^{n} (Z_i - b^T X_i)^2 = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathcal{Z}$$

where

- $\mathcal{X} = (X_1 | \cdots | X_n)^T$ is the $n \times p$ data matrix. Usually known as the "design matrix" in regression context.
- $\mathcal{Z} = (Z_1, \ldots, Z_n)^T$ is an n-vector of the observed responses.

☛ When $p > n$, $\mathcal{X}^T\mathcal{X}$ is non-invertible. So $\hat{\beta}_{LS}$ is impossible to obtain:

- $rank(\mathcal{X}) \leq min(n, p)$.
- $rank(\mathcal{X}^T\mathcal{X}) = rank(\mathcal{X}) \leq min(n, p) = n < p$
- Since $\mathcal{X}^T\mathcal{X}$ is a $p \times p$ matrix, it can not be of full rank

☛ Even if $p < n$, $\mathcal{X}^T\mathcal{X}$ can also be non-invertible when some of its columns can be written a linear combination of the others.

☛ More generally, $\mathcal{X}^T\mathcal{X}$ may be nearly non-invertible, when one or more of the columns of $\mathcal{X}$ can be approximately written as a linear combination of the others. This is known as *near collinearity* among the columns of $\mathcal{X}$ (or among the components of X).

☛ In the latter case, $(\mathcal{X}^T\mathcal{X})^{-1}$ can also be tricky to obtain numerically in practice.

☛ Moreover, conditional on $X_1, \ldots, X_n$

$$Cov(\hat{\beta}_{LS}) = \sigma^2(\mathcal{X}^T\mathcal{X})^{-1}.$$

– If $\mathcal{X}^T\mathcal{X}$ is nearly non-invertible, some of its eigenvalues are very small.

– This means that some of the eigenvalues of $(\mathcal{X}^T\mathcal{X})^{-1}$ are huge.

– which in turns implies that $\hat{\beta}_{LS}$ is highly variable.

☛ Precisely, there can exist some unit vectors $a \in \mathbb{R}^p$ (e.g. the first eigenvector of $(\mathcal{X}^T\mathcal{X})^{-1}$) such that

$$Var(a^T\hat{\beta}_{LS}) = \sigma^2 a^T(\mathcal{X}^T\mathcal{X})^{-1}a$$

is large.

☛ Then $\hat{\beta}_{LS}$ is not very useful: it is very "noisy" $\Rightarrow$ bring us some misleading information about the relationship between X and Z.

☛ When we can't estimate the linear model (1) in a reliable way, we may instead take a modified version of it that can estimated reasonably well from the data.

☛ Suggestion: perform the linear regression on some PCs of X instead of on X itself.

☛ To motivate this, note that if some of the "X" variables can be written as a linear combination of the others, then some of the eigenvalues of $\Sigma_X = var(X)$ will be equal to zero.

☛ E.g. If $X = (X_1, X_2)^T$ and $X_2 = cX_1$ (*exact collinearity*) where c is a constant, then we have

$$\Sigma_X = \begin{pmatrix} var(X_1) & c \cdot var(X_1) \\ c \cdot var(X_1) & c^2 \cdot var(X_1) \end{pmatrix}$$

so that the second column of Σ is equal to c times the first column. Thus $|\Sigma| = 0$ and Σ has one zero eigenvalue.

☛ More generally, suppose that only the first q eigenvalues of Σ_X are non-zero so that the variance of the PCs $Y_1, \ldots, Y_p$ of X satisfy

$$var(Y_j) = \lambda_j, \quad \text{for } j = 1, \ldots, q$$

and

$$var(Y_j) = 0 \quad \text{for } j = q+1, \ldots, p.$$

☛ Since the Y_j's are uncorrelated, we have

$$var(Y) = \Sigma_Y = \Lambda$$

where

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix}$$

is a block matrix with the upper-left non-trivial block

$$\Lambda_1 = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & \ldots & 0 & \lambda_q \end{pmatrix}.$$

☛ In other words,

$$Y \sim (0, \Lambda),$$

i.e. $E[Y] = 0$ and the last $p - q$ components of Y have variance 0.

☛ If we write Y as

$$Y^T = (Y_{(1)}^T, Y_{(2)}^T)$$

where

$$Y_{(1)} = (Y_1, \ldots, Y_q)^T$$

and

$$Y_{(2)} = (Y_{q+1}, \ldots, Y_p)^T,$$

then

$$Y_{(1)} \sim (0, \Lambda_1), \quad Y_{(2)} \sim (0, 0),$$

i.e. $Y_{(2)}$ is a degenerate vector of both mean and variance equal to 0.

☛ Now since

$$Y = \Gamma^T X, \tag{2}$$

where Γ is the matrix with the eigenvectors of Σ_X as columns, we have

$$X = \Gamma Y = \Gamma_{(1)} Y_{(1)} + \Gamma_{(2)} Y_{(2)},$$

where

$$\Gamma_{(1)} = (\gamma_1 | \ldots | \gamma_q), \quad \Gamma_{(2)} = (\gamma_{q+1} | \ldots | \gamma_p)$$

and

$$\Gamma_{(2)} Y_{(2)} \sim (0, 0).$$

(Recall that we have assumed $E[X] = 0$, so $Y = \Gamma^T (X - E[X])$ in (2).)

☛ Thus X can just be expressed as

$$X = \Gamma_{(1)} Y_{(1)}.$$

$\Rightarrow$ X contains a lot of redundant information and we can just work with $Y_{(1)}$, the vector of the first q PCs.

☛ Getting back to our linear regression problem

$$Z = m(X) + \epsilon = \beta^T X + \epsilon,$$

we are interested in estimating $m(\cdot)$, when Z and X have mean zero and there is a level of collinearity among the components of X.

☛ With exact collinearity, Σ had only $q < p$ non-zero eigenvalues and

$$X = \Gamma_{(1)} Y_{(1)}$$

where $Y_{(1)}$ is the vector of the first q PCs. In this case, we can write

$$Z = m_{\text{PC}}(Y_{(1)}) + \epsilon, \tag{3}$$

where

$$m_{\text{PC}}(Y_{(1)}) = {\beta_{\text{PC}}}^T Y_{(1)} \quad \text{and} \quad \beta_{\text{PC}} = \underbrace{\Gamma_{(1)}^T \beta}_{q \times 1}$$

☛ To estimate β_{PC} we can use the least squares estimator

$$\hat{\beta}_{\text{PC}} = argmin_{b \in \mathbb{R}^q} \sum_{i=1}^{n} (Z_i - b^T Y_{(1),i})^2 = (\mathcal{Y}_{(1)}^T \mathcal{Y}_{(1)})^{-1} \mathcal{Y}_{(1)}^T \mathcal{Z} \,,$$

where $Y_{(1),i} = (Y_{i1}, \ldots, Y_{iq})^T$ denotes the version of $Y_{(1)}$ for the ith individual and

$$\mathcal{Y}_{(1)} = \begin{pmatrix} Y_{11} & \ldots & Y_{1q} \\ & \ldots & \\ Y_{n1} & \ldots & Y_{nq} \end{pmatrix} .$$

☛ More often, we only have *near collinearity*, so that all the eigenvalues of Σ are non-zero but some may be near zero.

☛ Suppose that $\lambda_{q+1}, \ldots, \lambda_p$ are small, and let

$$Y^T = (Y_{(1)}^T, Y_{(2)}^T), \quad Y_{(1)} = (Y_1, \ldots, Y_q)^T, \quad Y_{(2)} = (Y_{q+1}, \ldots, Y_p)^T.$$

Then

$$var(Y_{q+1}) = \lambda_{q+1}, \ldots, var(Y_p) = \lambda_p$$

are small.

☛ In this case, getting rid of the PCs with small eigenvalues will give rise to an approximation:

$$X = \Gamma Y = \Gamma_{(1)} Y_{(1)} + \Gamma_{(2)} Y_{(2)} \approx \Gamma_{(1)} Y_{(1)}.$$

☛ For centered response Z and regressor vector X, we can approximate the full regression model

$$\begin{aligned} Z &= \beta^T X + \epsilon \\ &= \beta^T \Gamma Y + \epsilon \\ &= \beta^T \underbrace{\Gamma_{(1)}}_{p\times q} \underbrace{Y_{(1)}}_{q\times 1} + \beta^T \underbrace{\Gamma_{(2)}}_{p\times(p-q)} \underbrace{Y_{(2)}}_{(p-q)\times 1} + \epsilon \end{aligned}$$

with the simpler model

$$Z = \beta_{PC}^T Y_{(1)} + \eta$$

where $E(\eta) = 0$, *believing that the variability from* $Y_{(2)}$ *is negligible*, i.e.

$$X \approx \Gamma_{(1)} Y_{(1)}.$$

☛ The mean function of the simpler model is defined as

$$m_{\text{PC}}(y_{(1)}) = E(Z|Y_{(1)} = y_{(1)}) = {\beta_{\text{PC}}}^T y_{(1)},$$

which is easier to estimate than $m(\cdot)$ because $dim(\beta_{\text{PC}}) < dim(\beta)$.

☛ Again, the least squares estimator of β_{PC} is

$$\hat{\beta}_{\text{PC}} = argmin_{b \in \mathbb{R}^q} \sum_{i=1}^{n} (Z_i - b^T Y_{(1),i})^2 = (\mathcal{Y}_{(1)}^T \mathcal{Y}_{(1)})^{-1} \mathcal{Y}_{(1)}^T \mathcal{Z} \, .$$

☛We hope that the approximation

$$m(x) \approx m_{\text{PC}}(y_{(1)})$$

is good so that we can estimate $m(x)$ by

$$\hat{m}_{\text{PC}}(y_{(1)}) = \hat{\beta}_{\text{PC}}^T \, y_{(1)}.$$

☛ Of course, in practice, one can only use an empirical version of Γ based on spectral-decomposing an empirical estimate of Σ.

☛ One can then declare the "PCR estimate" of β as

$$\hat{\beta}^{pcr}(q) = \underbrace{\hat{\Gamma}_{(1)} \hat{\beta}_{PC}}_{p \times 1},$$

where $\hat{\Gamma}_{(1)}$ is an empirical version of $\Gamma_{(1)}$.

- It can be shown that when $q < p$, conditional on $\mathcal{X}$,

$$Var(\hat{\beta}_{LS}) - Var(\hat{\beta}^{pcr}(q)) \geq 0,$$

i.e. $Var(\hat{\beta}_{LS}) - Var(\hat{\beta}^{pcr}(q))$ is positive semidefinite.

- Hence, for any vector $a \in \mathbb{R}^p$, conditional on $\mathcal{X}$

$$Var(a^T \hat{\beta}_{LS}) \geq Var(a^T \hat{\beta}^{pcr}(q))$$

- But be aware that $\hat{\beta}^{pcr}(q)$ is biased, unlike $\hat{\beta}_{LS}$.

- We've reduced the variance in exchange for some bias; an example of *bias-variance tradeoff*.

Proof for variance reduction: Under our simplifying assumption that $E(X) = 0$, suppose we form the PCs based on $\hat{\Sigma} = n^{-1}\mathcal{X}^T\mathcal{X}$.

1. Let $\hat{\Sigma} = n^{-1}\mathcal{X}^T\mathcal{X} = \hat{\Gamma}\hat{\Lambda}\hat{\Gamma}^T$, the spectral decomposition of $\hat{\Sigma}$.

2. Let $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$, where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p > 0$ are all non-zero, and $\hat{\Gamma} = (\hat{\gamma}_1 | \ldots | \hat{\gamma}_p)$ with $\hat{\gamma}_1 \ldots \hat{\gamma}_p$ as columns.

3. We have:

$$\begin{aligned}
Var(\hat{\beta}^{pcr}(q)) &= \hat{\Gamma}_{(1)} Var(\hat{\beta}_{PC}) \hat{\Gamma}_{(1)}^T \\
&= \sigma^2 \hat{\Gamma}_{(1)} (\mathcal{Y}_{(1)}^T \mathcal{Y}_{(1)})^{-1} \hat{\Gamma}_{(1)}^T \\
&= \sigma^2 \hat{\Gamma}_{(1)} (\hat{\Gamma}_{(1)}^T \mathcal{X}^T \mathcal{X} \hat{\Gamma}_{(1)})^{-1} \hat{\Gamma}_{(1)}^T \\
&= \frac{\sigma^2}{n} \hat{\Gamma}_{(1)} (\hat{\Gamma}_{(1)}^T \hat{\Gamma} \hat{\Lambda} \hat{\Gamma}^T \hat{\Gamma}_{(1)})^{-1} \hat{\Gamma}_{(1)}^T \\
&= \frac{\sigma^2}{n} \hat{\Gamma}_{(1)} \text{diag}(\hat{\lambda}_1^{-1}, \ldots, \hat{\lambda}_q^{-1}) \hat{\Gamma}_{(1)}^T = \frac{\sigma^2}{n} \sum_{j=1}^{q} \frac{\hat{\gamma}_j \hat{\gamma}_j^T}{\hat{\lambda}_j},
\end{aligned}$$

where we've used $\underbrace{\hat{\Gamma}_{(1)}^T \hat{\Gamma}}_{q \times p} = (\underbrace{I_q}_{\substack{q \times q \\ \text{identity}}} | \underbrace{0, \ldots, 0}_{\substack{p-q \text{ many} \\ \text{columns of } 0\text{'s}}})$ in the last equality.

4. We also know that

$$Var(\hat{\beta}_{LS}) = \sigma^2(\mathcal{X}^T\mathcal{X})^{-1} = \frac{\sigma^2}{n}\hat{\Gamma}\hat{\Lambda}^{-1}\hat{\Gamma}^T = \frac{\sigma^2}{n}\sum_{j=1}^{p}\frac{\hat{\gamma}_j\hat{\gamma}_j^T}{\hat{\lambda}_j}.$$

5. Hence,

$$Var(\hat{\beta}_{LS}) - Var(\hat{\beta}^{pcr}(q)) = \frac{\sigma^2}{n}\sum_{j=q+1}^{p}\frac{\hat{\gamma}_j\hat{\gamma}_j^T}{\hat{\lambda}_j},$$

which is positive definite because for any test vector $a \in \mathbb{R}^p$,

$$\frac{\sigma^2}{n}a^T\left(\sum_{j=q+1}^{p}\frac{\hat{\gamma}_j\hat{\gamma}_j^T}{\hat{\lambda}_j}\right)a = \frac{\sigma^2}{n}\sum_{j=q+1}^{p}\underbrace{\frac{a^T\hat{\gamma}_j\hat{\gamma}_j^Ta}{\hat{\lambda}_j}}_{\geq 0} \geq 0 \text{ for all } j.$$

6. In particular, if the $\hat{\lambda}_{q+1}, \ldots, \hat{\lambda}_p$ are very close to zero, the quadratic-form quantities $\frac{a^T\hat{\gamma}_j\hat{\gamma}_j^Ta}{\hat{\lambda}_j}$, $j = q+1, \ldots p$, can be very large $\Rightarrow$ *PCR buys a lot of variance reduction*.

☛ How to choose q? Choosing the minimizer of the overall *residual sum of squares*

$$q = \text{argmin}_{k=1,\dots p}\, RSS(k),$$

does NOT work. Here

$$RSS(k) = \sum_{i=1}^{n} (Z_i - Y_{(1),i}^T \hat{\beta}_{\text{PC},k})^2,$$

and

$$\hat{\beta}_{\text{PC},k} = (\mathcal{Y}_{(1)}^T \mathcal{Y}_{(1)})^{-1} \mathcal{Y}_{(1)}^T \mathcal{Z}$$

is the least squares estimator of β_{PC} with the first k PCs, i.e.

$$\mathcal{Y}_{(1)} \text{ is } n\text{-by-}k.$$

(but its dependence on k is suppressed to simplify notation).

☛ It must be the case that

$$RSS(p) \leq \cdots \leq RSS(1)$$

(why?), so one will always end up with $q = p$.

☛ This is the problem of *overfitting*.

☛ Cross-Validation (CV) is often used to choose the number q of PCs.

☛ In its simplest form, it consists in choosing

$$q = argmin_{k=1,\dots,p}\, CV(k)\,,$$

where $CV(k)$ is the *leave-one-out* CV (LOOCV) error based on k PCs.

☛ Here,

$$CV(k) = \sum_{i=1}^{n} (Z_i - Y_{(1),i}^T \hat{\beta}_{\text{PC},k}^{-i})^2\,,$$

where $\hat{\beta}_{\text{PC},k}^{-i}$ denotes the least squares estimator of β_{PC} computed based on the first k PCs of X, but without using the ith data point, i.e

$$\hat{\beta}_{\text{PC},k}^{-i} = (\mathcal{Y}_{(1)}^{-i\,T} \mathcal{Y}_{(1)}^{-i})^{-1} \mathcal{Y}_{(1)}^{-i\,T}\, \mathcal{Z}^{-i}$$

where

- $\mathcal{Z}^{-i}$ is $\mathcal{Z}$ with its ith component Z_i removed, and
- $\mathcal{Y}_{(1)}^{-i}$ is $\mathcal{Y}_{(1)}$ with its ith row $Y_{(1),i}^T$ removed, i.e. it is $(n-1) \times k$.

☛ Consider the centered linear regression model again:

$$Z = m(X) + \epsilon, \quad m(x) = E(Z|X = x) = \beta^T x$$

where

- $X = (X_1, \ldots, X_p)^T$ and x are p-vectors
- $\beta \in \mathbb{R}^p$ is the parameter vector
- $E(X) = E(Z) = E(\epsilon) = 0$.

☛ *Issues with PCR*: The decomposition into successive PCs $Y_1, \ldots, Y_p$ ignores information about Z completely.

☛In regression, we aim to explain the relationship between Z and X; selecting the first few PCs of X, which focus on the variability *within* X alone, is not necessarily the best thing to do.

☛Partial least squares (PLS): An alternative to PCR for reducing the dimension of X that focuses on the relationship between X and Z.

☛ In PLS, we choose a decomposition of X into successive linear projections

$$T_1, T_2, \ldots, T_p$$

that strive to keep most information about the relationship between Z and X.

☛ The first component T_1 is constructed as

$$T_1 = \phi_1^T X,$$

where $\phi_1 \in \mathbb{R}^p$ is a coefficient vector chosen such that

$$|cov(Z, T_1)|$$

is maximized, under the constraint that it has unit length, i.e.

$$\|\phi_1\| = 1.$$

☛ For $k = 2, \ldots, p$, the k-th component T_k is constructed recursively as

$$T_k = \phi_k^T X,$$

where $\phi_k \in \mathbb{R}^p$ is chosen such that

$$|cov(Z, T_k)|$$

is maximized, under the constraints $\|\phi_k\| = 1$ AND

$$cov(T_k, T_j) = \phi_k^T \Sigma \phi_j = 0 \text{ for } j = 1, \ldots, k-1,$$

i.e. the k-th component T_k must be uncorrelated with the prior $k-1$ components.

☛ These T_j's are constructed as linear combinations of X that explain most linear relationship between Z and X.

☛ Collecting $\phi_1, \ldots, \phi_p$ as columns into a $p \times p$ matrix as

$$\Phi = (\phi_1 | \ldots | \phi_p);$$

the hope is that the first few components of the p-vector

$$T = (T_1, \ldots, T_p)^T = \Phi^T X$$

contain the most useful information for estimating $m(x)$.

- Recall the linear model is

$$m(x) = E(Z|X = x) = \beta^T x.$$

- Define the p-vector t as

$$t = (t_1, \ldots, t_p)^T = \Phi^T x$$

and for $q < p$ define the q-vectors

$$T_{(1)} = (T_1, \ldots, T_q)^T \text{ and } t_{(1)} = (t_1, \ldots, t_q)^T.$$

- We use the approximation

$$m(x) = \beta^T x \approx E(Z|T_{(1)} = t_{(1)}) \equiv m_{\text{PLS}}(t_{(1)}) = \beta_{\text{PLS}}^T t_{(1)},$$

and regress our response on $T_{(1)}$ instead.

☛ The least squares estimator of β_{PLS} is equal to

$$\hat{\beta}_{\text{PLS}} = argmin_{b \in \mathbb{R}^q} \sum_{i=1}^{n} (Z_i - b^T T_{(1),i})^2 = (\mathcal{T}_{(1)}^T \mathcal{T}_{(1)})^{-1} \mathcal{T}_{(1)}^T \mathcal{Z} \,,$$

where $\mathcal{Z} = (Z_1, \ldots, Z_n)^T$ is an n-vector,

$$T_{(1),i} = (T_{i1}, \ldots, T_{iq})^T$$

collects the first q components for the ith individual, and

$$\mathcal{T}_{(1)} = \begin{pmatrix} T_{11} & \ldots & T_{1q} \\ & \ldots & \\ T_{n1} & \ldots & T_{nq} \end{pmatrix} .$$

☛ As with PCR, we can choose q by cross-validation:

$$q = argmin_{k=1,\ldots,p}\, CV(k)\,,$$

where

$$CV(k) = \sum_{i=1}^{n} \{ Z_i - T^T_{(1),i} \hat{\beta}^{-i}_{\text{PLS},k} \}^2\,,$$

and $\hat{\beta}^{-i}_{\text{PLS},k}$ denotes the least squares estimator of β_{PLS} based on k PLS components, but computed without the i-th sample of $\mathcal{Z}$ and $\mathcal{T}_{(1)}$.

☛ We have presented PLS at the population level. Of course, in practice, we can't get the exact ϕ_1 that maximizes $|cov(Z, T_1)|$ (a population quantity), nor the $\phi_2, \ldots, \phi_p$'s.

☛ We instead compute an empirical version $\hat{\phi}_1$ as illustrated next, given the response vector $\mathcal{Z}$ $(n \times 1)$ and data matrix $\mathcal{X}(n \times p)$.

☛ In what follows, $\langle a, b \rangle$ denotes the dot product between $a, b \in \mathbb{R}^n$.

Computing the 1st PLS component in practice:

1. Let $\bar{Z} = n^{-1} \sum_i Z_i$ and $\tilde{\mathcal{Z}} = \mathcal{Z} - (\bar{Z}, \ldots, \bar{Z})^T$, the centered version of $\mathcal{Z}$.

2. Similarly, let $\tilde{\mathcal{X}} = (\tilde{\mathcal{X}}_1 | \ldots | \tilde{\mathcal{X}}_p)$ be $\mathcal{X}$ centered, i.e. each column $\tilde{\mathcal{X}}_j$ of $\tilde{\mathcal{X}}$ is
$$\tilde{\mathcal{X}}_j = (X_{1j} - \bar{X}_j, \ldots, X_{nj} - \bar{X}_j)^T,$$
where $\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}$.

3. For any candidate projection vector $\phi_1 = (\phi_{11}, \ldots, \phi_{1p})^T$,
$$\widehat{cov}(Z, {\phi_1}^T X) = n^{-1} \langle \tilde{\mathcal{Z}}, \tilde{\mathcal{X}} \phi_1 \rangle = \frac{\langle \tilde{\mathcal{Z}}, \tilde{\mathcal{X}}_1 \rangle \phi_{11} + \cdots + \langle \tilde{\mathcal{Z}}, \tilde{\mathcal{X}}_p \rangle \phi_{1p}}{n}$$
is the *empirical* analog of $\text{cov}(Z, {\phi_1}^T X)$.

4. Up to n^{-1}, $\widehat{cov}(Z, {\phi_1}^T X)$ is a dot product between ϕ_1 and the vector
$$(\langle \tilde{\mathcal{Z}}, \tilde{\mathcal{X}}_1 \rangle, \ldots, \langle \tilde{\mathcal{Z}}, \tilde{\mathcal{X}}_p \rangle)^T.$$
By Cauchy's inequality,
$$\hat{\phi}_1 = \text{argmax}_{\|\phi_1\|=1} \widehat{cov}(Z, {\phi_1}^T X)$$
must be the unit vector $\frac{1}{\sqrt{\sum_{j=1}^p \langle \tilde{\mathcal{Z}}, \tilde{\mathcal{X}}_j \rangle^2}} (\langle \tilde{\mathcal{Z}}, \tilde{\mathcal{X}}_1 \rangle, \ldots, \langle \tilde{\mathcal{Z}}, \tilde{\mathcal{X}}_p \rangle)^T$.

Computing the 2nd PLS component in practice: Orthogonalizing $\tilde{\mathcal{X}}_1, \ldots, \tilde{\mathcal{X}}_p$ with respect to the first PLS component

$$\tilde{\mathcal{T}}_1 \equiv \tilde{\mathcal{X}}\hat{\phi}_1,$$

and then maximizing the empirical covariance. In detail:

1. For $j = 1, \ldots, p$, define $\tilde{\mathcal{X}}_j^{(2)} = \tilde{\mathcal{X}}_j - \frac{\langle \tilde{\mathcal{T}}_1, \tilde{\mathcal{X}}_j \rangle}{\langle \tilde{\mathcal{T}}_1, \tilde{\mathcal{T}}_1 \rangle} \tilde{\mathcal{T}}_1$. These $\tilde{\mathcal{X}}_1^{(2)}, \ldots, \tilde{\mathcal{X}}_p^{(2)}$ span a space orthogonal to that of $\tilde{\mathcal{T}}_1$.

2. Collect $\tilde{\mathcal{X}}_1^{(2)}, \ldots, \tilde{\mathcal{X}}_p^{(2)}$ as columns in the $n \times p$ matrix

$$\tilde{\mathcal{X}}^{(2)} = (\tilde{\mathcal{X}}_1^{(2)} | \ldots | \tilde{\mathcal{X}}_p^{(2)}).$$

3. Now for *any* candidate projection vector $\phi_2^{(2)} = (\phi_{21}^{(2)}, \ldots, \phi_{2p}^{(2)})^T$, the linear combination

$$\tilde{\mathcal{X}}^{(2)}\phi_2^{(2)} = \phi_{21}^{(2)}\tilde{\mathcal{X}}_1^{(2)} + \cdots + \phi_{2p}^{(2)}\tilde{\mathcal{X}}_p^{(2)}$$

must be orthogonal to $\tilde{\mathcal{T}}_1$, so that the empirical covariance

$$n^{-1}\langle \tilde{\mathcal{X}}^{(2)}\phi_2^{(2)}, \tilde{\mathcal{T}}_1 \rangle$$

between $\tilde{\mathcal{X}}^{(2)}\phi_2^{(2)}$ and $\tilde{\mathcal{T}}_1$ must be zero.

4. As before, we let $\hat{\phi}_2^{(2)}$ be the normalized version of $(\langle \tilde{\mathcal{Z}}, \tilde{\mathcal{X}}_1^{(2)} \rangle, \ldots, \langle \tilde{\mathcal{Z}}, \tilde{\mathcal{X}}_p^{(2)} \rangle)^T$, i.e.

$$\hat{\phi}_2^{(2)} = \frac{1}{\sqrt{\sum_{j=1}^p \langle \tilde{\mathcal{Z}}, \tilde{\mathcal{X}}_j^{(2)} \rangle^2}} (\langle \tilde{\mathcal{Z}}, \tilde{\mathcal{X}}_1^{(2)} \rangle, \ldots, \langle \tilde{\mathcal{Z}}, \tilde{\mathcal{X}}_p^{(2)} \rangle)^T$$

and defines the 2nd PLS component as

$$\tilde{\mathcal{T}}_2 = \tilde{\mathcal{X}}^{(2)} \hat{\phi}_2^{(2)}.$$

This maximizes the empirical covariance with $\tilde{\mathcal{Z}}$.

☛ **Important note**: The normalized $\hat{\phi}_2^{(2)}$ is <u>NOT</u> the $\hat{\phi}_2$ that we might have originally wanted, because $\hat{\phi}_2^{(2)}$ is with respect to the new design matrix $\tilde{\mathcal{X}}^{(2)}$, while $\hat{\phi}_2$ is with respect to the original design matrix $\tilde{\mathcal{X}}$.

☛ The second PLS component returned by the `scores` object of the `plsr` function in `R` is the same as the $\tilde{\mathcal{T}}_2$ computed by the recipe above.

☛ Getting the remaining PLS components follows iteratively, by first orthogonalizing the feature vectors with respect to the previous PLS components and maximizing the empirical covariance with $\tilde{\mathcal{Z}}$.

☛ Note: When $\tilde{\mathcal{X}}$ has orthogonal columns, i.e. when $\tilde{\mathcal{X}}^T\tilde{\mathcal{X}} = I$, then PLS will coincide with the usual linear regression. (ESL. Ex 3.14)

☛ Read Algorithm 3.3 in ESL (p.81) (Caveat: Their notation is different; "$\mathbf{z}$" represents the PLS components, and "$\mathbf{y}$" represents the responses in regression).

Algorithm 3.3 *Partial Least Squares.*

1. Standardize each $\mathbf{x}_j$ to have mean zero and variance one. Set $\hat{\mathbf{y}}^{(0)} = \bar{y}\mathbf{1}$, and $\mathbf{x}_j^{(0)} = \mathbf{x}_j,\ j = 1, \ldots, p$.

2. For $m = 1, 2, \ldots, p$

   (a) $\mathbf{z}_m = \sum_{j=1}^{p} \hat{\varphi}_{mj} \mathbf{x}_j^{(m-1)}$, where $\hat{\varphi}_{mj} = \langle \mathbf{x}_j^{(m-1)}, \mathbf{y} \rangle$.

   (b) $\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$.

   (c) $\hat{\mathbf{y}}^{(m)} = \hat{\mathbf{y}}^{(m-1)} + \hat{\theta}_m \mathbf{z}_m$.

   (d) Orthogonalize each $\mathbf{x}_j^{(m-1)}$ with respect to $\mathbf{z}_m$: $\mathbf{x}_j^{(m)} = \mathbf{x}_j^{(m-1)} - [\langle \mathbf{z}_m, \mathbf{x}_j^{(m-1)} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle] \mathbf{z}_m,\ j = 1, 2, \ldots, p$.

3. Output the sequence of fitted vectors $\{\hat{\mathbf{y}}^{(m)}\}_1^p$. Since the $\{\mathbf{z}_\ell\}_1^m$ are linear in the original $\mathbf{x}_j$, so is $\hat{\mathbf{y}}^{(m)} = \mathbf{X}\hat{\beta}^{\text{pls}}(m)$. These linear coefficients can be recovered from the sequence of PLS transformations.

☛ PCA and PLS are also used for nonlinear regression, where we approximate a possibly nonlinear mean function $m(x) = E(Z|X = x)$ by

$$m_{\text{PLS}}(t_{(1)}) = E(Z|T_{(1)} = t_{(1)}) \text{ or } m_{\text{PC}}(y_{(1)}) = E(Z|Y_{(1)} = y_{(1)}).$$

☛ For regression, PLS often works better than PCA.

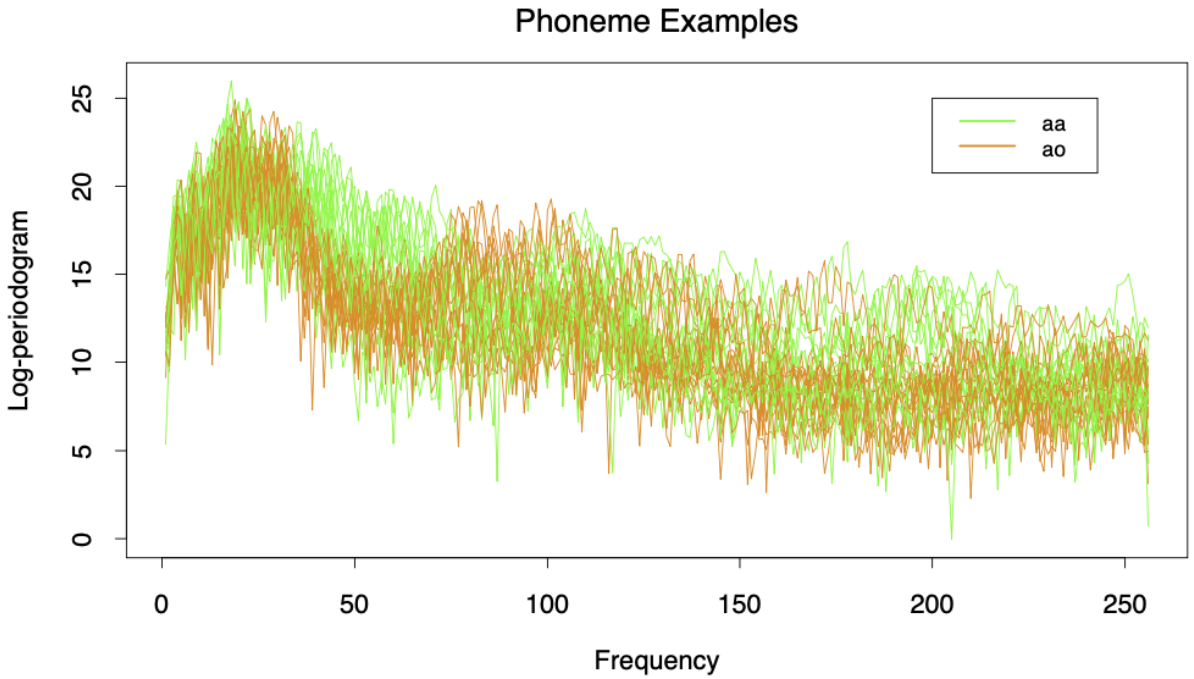
☛ Typically, either $m_{\text{PLS}}(t_{(1)})$ approximates $m(x)$ better than $m_{\text{PC}}(y_{(1)})$ does, or if they achieve the same approximation, then the number q of components needed in PLS is smaller than that needed by PC $\Rightarrow$ more computationally efficient.

☛ Example: Phoneme data available at `www-stat.stanford.edu/ElemStatLearn`.

☛ The dataset $\{X_i\}_{i=1}^N$ has a total of $N = 1717$ observations.

☛ Phoneme: a unit of sound that distinguishes one word from another in English.

☛ Each $X_i \in \mathbb{R}^{256}$ represents a log-periodogram constructed from recordings of one of two phonemes, "aa" as in "dark" and "ao" as in "water".



- For each i, we simulated Z_i values by taking

$$Z_i = m(X_i) + \epsilon_i = \beta^T X_i + \epsilon_i,$$

where $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ and $\beta \in \mathbb{R}^{256}$.

- Recall $X_i = \Gamma Y_i$. By letting $\beta_{PC} = \Gamma^T \beta$, we can write

$$Z_i = m(X_i) + \epsilon_i = \beta_{PC}^T Y_i + \epsilon_i.$$

- Given Γ is invertible, the coefficient vector β is chosen to make β_{PC} contain exactly 5 non-zero entries, i.e. all the interaction between X and Z depends on 5 PCs only:

  Case (i): Information contained in the first 5 PCs (i.e., first 5 entries in β_{PC} are non-zero)

  Case (ii): Information contained in PC 6 to 10.

  Case (iii): Information contained in PC 11 to 15.

  Case (iv): Information contained in PC 16 to 20.

☛ Randomly created 200 subsamples of size n ($<< N$), and compute the PLS and PCA regression estimators from each of these subsamples.

☛ n is the size of the *training sample*.

☛ In each instance, for each i in the remaining $N - n$ observations not used in training, we computed the predictors

$$\hat{Z}_i^{\text{PLS}} = \hat{m}_{\text{PLS}}(T_{(1),i})$$

or

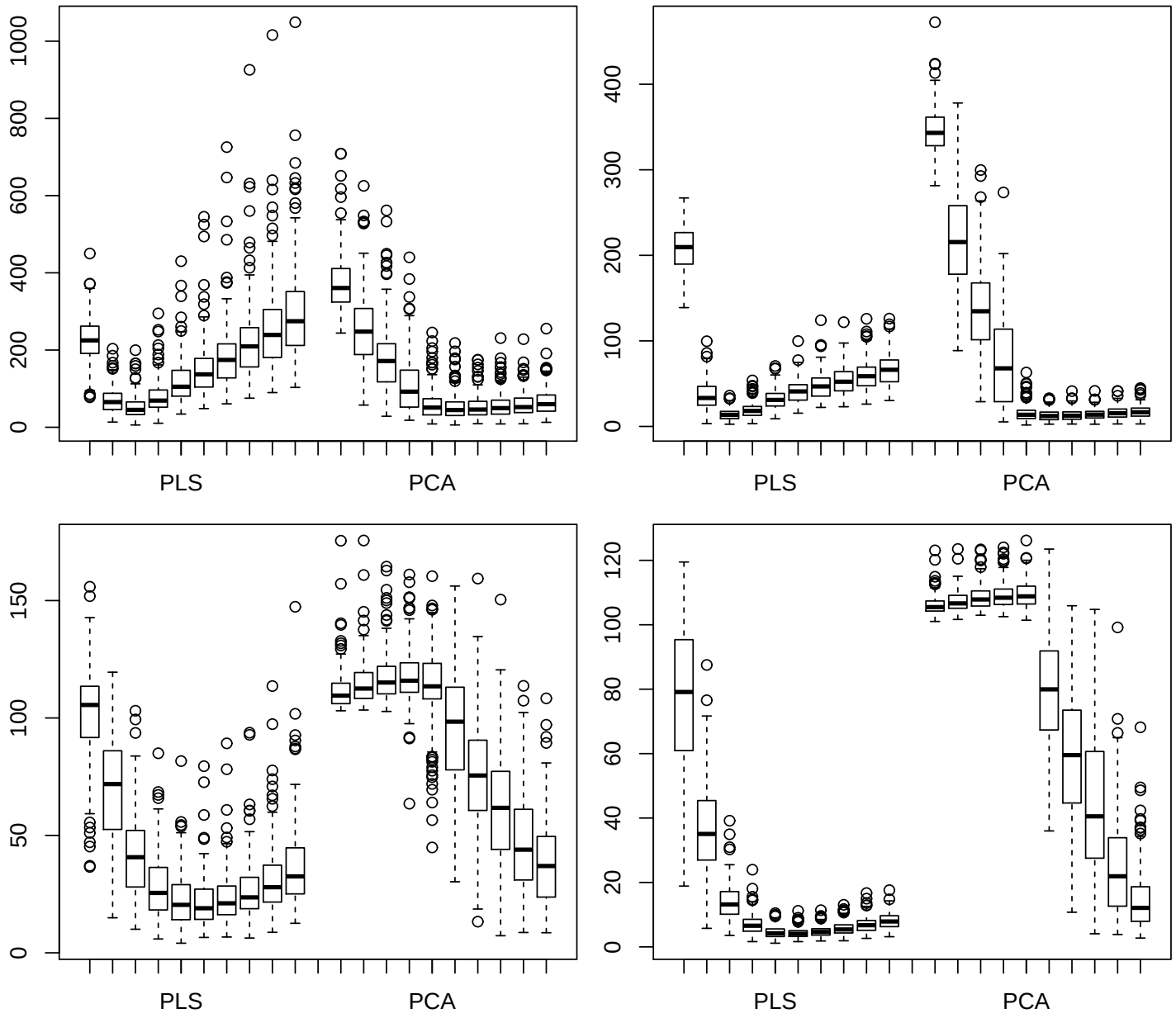
$$\hat{Z}_i^{\text{PC}} = \hat{m}_{\text{PC}}(Y_{(1),i})$$

for $q = 1, 2, \ldots, 10$.

☛ For $\hat{Z}_i$ denoting $\hat{Z}_i^{\text{PLS}}$ or $\hat{Z}_i^{\text{PC}}$, we computed the average squared **prediction error**

$$PE = \frac{1}{N-n} \sum_{i=1}^{N-n} (Z_i - \hat{Z}_i)^2$$

Cases (i) (top) and (ii) (bottom) when $n = 30$ (left) or $n = 100$ (right). Boxplots of the PE's for $q = 1$ to 10 from left to right.



Cases (iii) (top) and (iv) (bottom) when $n = 30$ (left) or $n = 100$ (right).