SF 2930 Regression Analysis

Lecture 11.1

Principal Components Regression (PCR)

Timo Koski

KTH Royal Institute of Technology

2023

Content – Lecture

- Principal Components Analysis (PCA)
- Principal Components Regression (PCR)
- Partial Least Squares (PLS)

Dimension reduction methods

The methods that we have discussed so far have controlled variance in two different ways,

- using a subset of the original variables, or
- shrinking their coefficients toward zero.

All of these methods are defined using the original predictors,

$$x_1, x_2, \ldots, x_p.$$

Dimension reduction methods

The methods that we have discussed so far have controlled variance in two different ways,

- using a subset of the original variables, or
- shrinking their coefficients toward zero.

All of these methods are defined using the original predictors,

$$x_1, x_2, \ldots, x_p.$$

We now explore a class of approaches that transform the predictors and then fit a least squares model using the transformed variables.
We will refer to these techniques as *dimension reduction methods* such as

- Principal component regression, and
- Partial least squares.

Principal component analysis (PCA)

A *principal component analysis (PCA)* is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables.

Its general objectives are

- data reduction, and
- interpretation.

Although p components are required to reproduce the total system variability, often much of this variability can be accounted for by a small number k of the principle components.

Let $\mathbf{X} = (X_1, ..., X_p)^T$ be a random vector with the covariance matrix Σ.

Algebraically, the principal components are particular linear combinations of the p random variables.

Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with $x_1, ..., x_p$ as the coordinate axes.

The first principal component direction of the data is that along which the observations vary the most.

Consider the linear combinations

$$y_1 = \mathbf{a}_1^T\mathbf{x}, \ldots, y_p = \mathbf{a}_p^T\mathbf{x}, \tag{1}$$

with the variances and covariances

$$\begin{aligned} \mathrm{Var}(y_i) \quad &= \mathbf{a}_i^T\Sigma\mathbf{a}_i, \quad i = 1, \ldots p, \\ \mathrm{Cov}(y_i, y_k) \quad &= \mathbf{a}_i^T\Sigma\mathbf{a}_k, \quad i, k = 1, \ldots p. \end{aligned} \tag{2}$$

The **principal components** are those *uncorrelated* linear combinations $y_1, \ldots, y_p$ whose variances above are as large as possible.The p **principal components** are defined as follows.

Definition

- *First principal component = linear combination* $y_1 = \mathbf{a}_1^T\mathbf{x}$ *that maximizes* $\mathrm{Var}(\mathbf{a}_1^T\mathbf{x})$ *subject to* $\mathbf{a}_1^T\mathbf{a}_1 = 1$.
- *ith principal component = linear combination* $y_i = \mathbf{a}_i^T\mathbf{x}$ *that maximize* $\mathrm{Var}(\mathbf{a}_i^T\mathbf{x})$ *subject to* $\mathbf{a}_i^T\mathbf{a}_i = 1$ *and* $\mathrm{Cov}(\mathbf{a}_i^T\mathbf{x}, \mathbf{a}_k^T\mathbf{x}) = 0$ *for* $k < i$, $i = 1, \ldots, p$.

Let $\lambda_1, ..., \lambda_p > 0$ be the eigenvalues of the matrix Σ and let $D = (\mathbf{d}_1, ..., \mathbf{d}_p)$ be an $m \times m$ orthogonal matrix such that

$$D^T \Sigma D = \Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & \ddots & \vdots & \dots & 0 \\ 0 & 0 & 0 & \dots & \lambda_n \end{pmatrix} \tag{3}$$

so that $\mathbf{d}_i$ is an eigenvector of Σ corresponding to the eigenvalue λ_i.

Proposition

For $k = 1, \ldots, p$

$$\lambda_k = \max_{\mathbf{a}^T\mathbf{a}=1,\ \mathbf{a}^T\mathbf{d}_i=0,\ i=1,\ldots,k-1} \mathbf{a}^T\Sigma\mathbf{a} = \mathbf{d}_k^T\Sigma\mathbf{d}_k.$$

Since $\mathbf{d}_i$ and $\mathbf{d}_j$ are orthogonal, the linear combinations $y_i = \mathbf{d}_i^T\mathbf{x}$ and $y_j = \mathbf{d}_j^T\mathbf{x}$ are uncorrelated.

Measures of total variation

Note that in transforming to principal components the measures $\operatorname{Tr}\Sigma$ and $|\Sigma|$ of total variation are unchanged, for

$$\operatorname{Tr}\Sigma = \operatorname{Tr} D^T \Sigma D = \operatorname{Tr}\Lambda = \sum_{i=1}^{p} \lambda_i,$$

$$|\Sigma| = |D^T \Sigma D| = |\Lambda| = \prod_{i=1}^{p} \lambda_i.$$

Note also that $\sum_{i=1}^{k} \lambda_i$ is the variance of the first k principal components.

Measures of total variation

Note that in transforming to principal components the measures $\operatorname{Tr}\Sigma$ and $|\Sigma|$ of total variation are unchanged, for

$$\operatorname{Tr}\Sigma = \operatorname{Tr} D^T \Sigma D = \operatorname{Tr}\Lambda = \sum_{i=1}^{p} \lambda_i,$$

$$|\Sigma| = |D^T \Sigma D| = |\Lambda| = \prod_{i=1}^{p} \lambda_i.$$

Note also that $\sum_{i=1}^{k} \lambda_i$ is the variance of the first k principal components.

- In principal component analysis the hope is that for some small k, this variance is close to $\operatorname{Tr}\Sigma$, i.e.,
- the first k principle components explain most of the variation in $\mathbf{X}$,
- the remaining $q = p - k$ principal components contribute little.

Sample principle component analysis

Assume that $\mathbf{X}_1, ..., \mathbf{X}_n$ are independently distributed as $N_p(\boldsymbol{\mu}, \Sigma)$. The MLE of $\boldsymbol{\mu}$ and Σ are given by

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \tag{4}$$

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T. \tag{5}$$

MLEs of the λ_i's, are the ordered eigenvalues of $\widehat{\Sigma}$. The $\hat{\lambda}_i$'s are distinct with probability one, since $n > p$.

Definition

The ith sample principal component is defined as the linear combination $\hat{y}_i = \mathbf{a}_i^T \mathbf{x}$ *that maximize the sample variance* $\mathbf{a}_i^T S \mathbf{a}_i$ *subject to* $\mathbf{a}_i^T \mathbf{a}_i = 1$ *and* $\mathbf{a}_i^T S \mathbf{a}_k = 0$ *for* $k < i$, $i = 1, ..., p$.

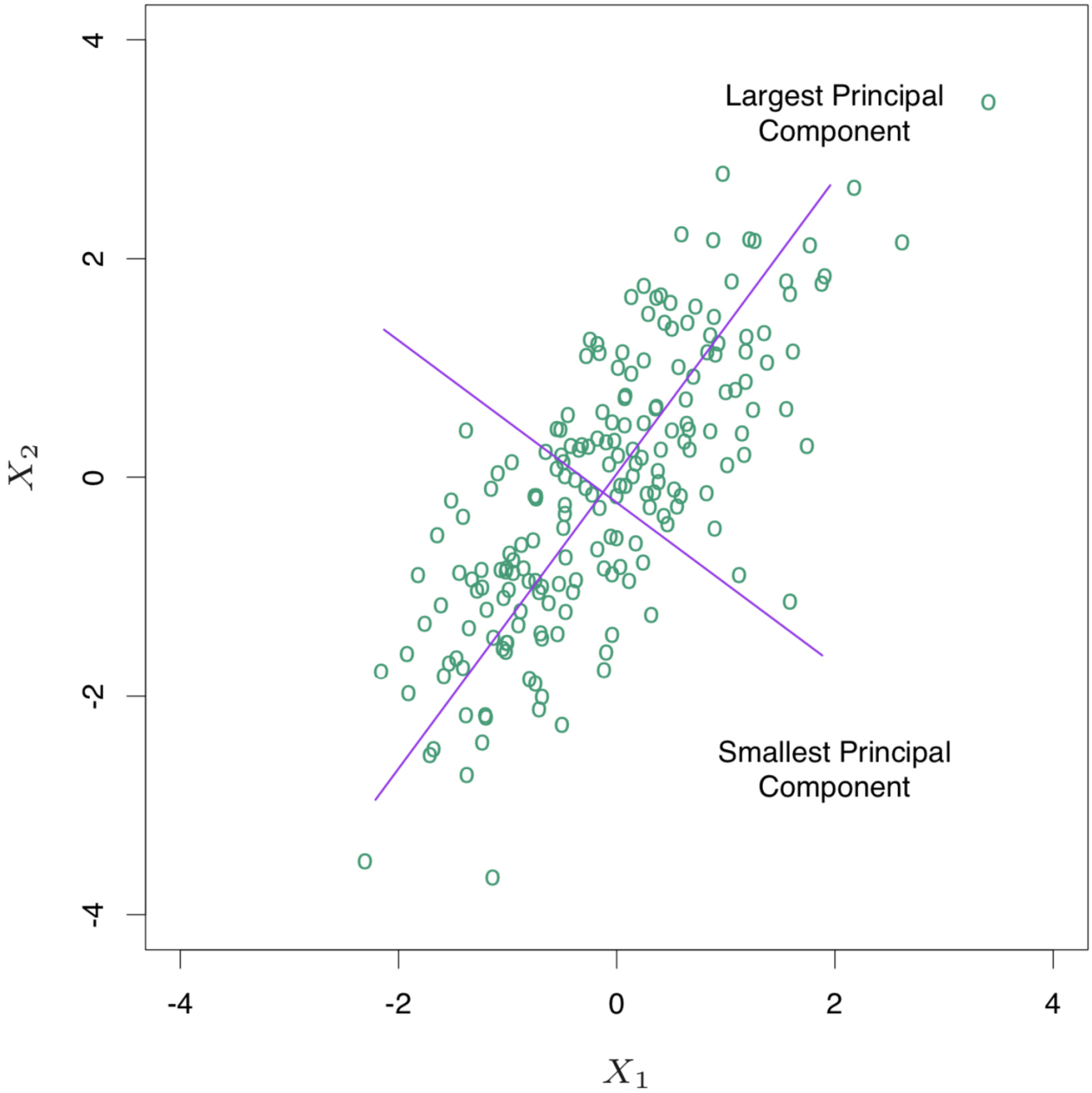
We now have a similar theorem as above.

Proposition

For $k = 1, ..., p$

$$\hat{\lambda}_k = \max_{\mathbf{a}^T\mathbf{a}=1,\quad \mathbf{a}^T\hat{\mathbf{d}}_i=0,\ i=1,...,k-1} \mathbf{a}^T S \mathbf{a} = \hat{\mathbf{d}}_k^T S \hat{\mathbf{d}}_k, \tag{6}$$

where $\hat{\mathbf{d}}_k$ *is an eigenvector of the MLE* $\widehat{\Sigma}$ *corresponding to the eigenvalue* $\hat{\lambda}_k$.


Largest Principal
Component
Smallest Principal
Component
X_1
X_2
-4
-2
0
2
4
-4
-2
0
2
4

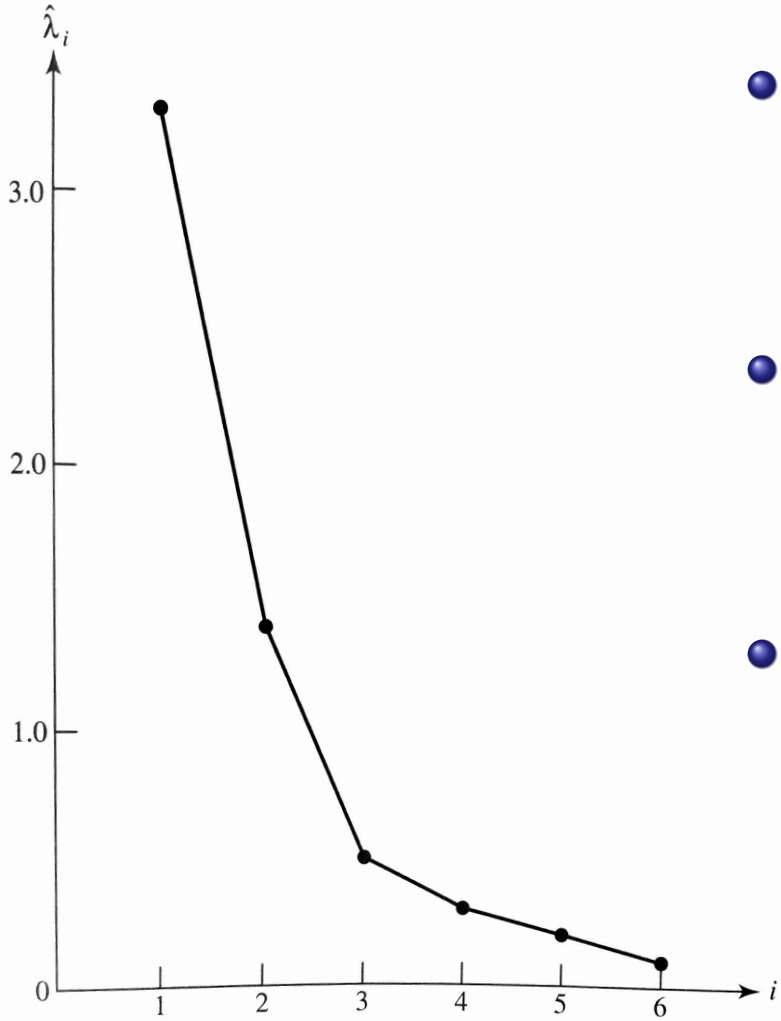
The number of principal components to use

- There is always the question of how many components to retain - and there is no definitive answer to this question.
- Things to consider include
  - the amount of total sample variance explained,
  - the relative sizes of the eigenvalues (i.e., the variance of the sample components),
  - the subject-matter interpretations of the components.

The number of principal components to use

- There is always the question of how many components to retain - and there is no definitive answer to this question.
- Things to consider include
  - the amount of total sample variance explained,
  - the relative sizes of the eigenvalues (i.e., the variance of the sample components),
  - the subject-matter interpretations of the components.
- However, a component associated with an eigenvalue near zero and, hence, deemed unimportant, may indicate an unsuspected linear dependency in the data.
- A useful visual aid to determining an appropriate number of principal components is a *scree plot*.

Scree plot



- A scree plot is the eigenvalues orded from the largest to the smallest.
- To determine the appropriate number of components, we look for an elbow (bend) in the plot.
- The number of components is taken to be the point at which the remaining eigenvalues are relatively small and all about the same size. In this case keep two components.

Principal components regression (PCR)

In **principal components regression (PCR)**,

- Center your model: X is $n \times k$.
- we first perform PCA on the original predictors,
- then perform dimension reduction by selecting the number of principal components (M) using cross-validation or test set error,
- finally conduct regression using the first M dimension reduced principal components.

Principal components regression (PCR)

In **principal components regression (PCR)**,

- Center your model: X is $n \times k$.
- we first perform PCA on the original predictors,
- then perform dimension reduction by selecting the number of principal components (M) using cross-validation or test set error,
- finally conduct regression using the first M dimension reduced principal components.

Hence, the PCR uses the most important principal components as the predictors in a linear regression model.

The fitting process for obtaining the PCR estimator involves regressing the response vector on the derived data matrix Z, since the principal components are mutually orthogonal to each other. Thus in the regression step, performing a multiple linear regression jointly on the M selected principal components as covariates is equivalent to carrying out M independent simple linear regressions separately on each of the M selected principal components as a covariate.

A more detailed description: The Multiple Centered Linear Regression Model

$\beta \in \mathbb{R}^k, n \geq k.$

$$\mathbf{Y} = X\beta + \varepsilon. \tag{7}$$

1) $E[\varepsilon] = \mathbf{0} \in \mathbb{R}^n$
2) $\mathbf{C}_\varepsilon = E\left[\varepsilon\varepsilon^T\right] = \sigma^2\mathbf{I}_n$
3) X^TX *is invertible*

Real symmetric matrices are diagonalizable by orthogonal matrices. $X^T X$ is a a real symmetric matrix, hence there is an orthogonal $k \times k$ matrix V such that

$$V^T X^T X V = \Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & \ddots & \vdots & \dots & 0 \\ 0 & 0 & 0 & \dots & \lambda_k \end{pmatrix}$$

V is an orthogonal (or orthonormal matrix): columns and rows are orthonormal vectors and thus

$$V^T V = V V^T = \mathbb{I}_k$$

Take $M \in \{1, \ldots, k\}$ and let V_M be the $k \times M$ matrix consisting of the first M columns of $k \times 1$ vectors $\mathbf{v}_i$ of V. Set

$$Z = XV_M = (X\mathbf{v}_1 \ldots, X\mathbf{v}_M)$$

$n \times M$ matrix. $X\mathbf{v}_i$ are the **principal components**. The $n \times 1$ column vectors of W_M are orthogonal, since

$$(X\mathbf{v}_i)^T X\mathbf{v}_j = \mathbf{v}_i X^T X\mathbf{v}_j$$

is the element on position (i, j) in $V^T X^T XV$ and thus $' = 0$ for $i \neq j$ and $= \lambda_i$ for $i = j$.
Then we can find at least one $M \times 1$ vector $\boldsymbol{\alpha}$ such that

$$\boldsymbol{\beta} = V_M \boldsymbol{\alpha}$$

(This can be shown by the properties of generalized inverses.)
Then we get in (7) that

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = XV_M\boldsymbol{\alpha} + \boldsymbol{\varepsilon} = Z\boldsymbol{\alpha} + \boldsymbol{\varepsilon}.$$

$$\mathbf{Y} = Z\boldsymbol{\alpha} + \boldsymbol{\varepsilon}.$$

Then LSE of $\boldsymbol{\alpha}$ is

$$\widehat{\boldsymbol{\alpha}} = (Z^T Z)^{-1} Z^T \mathbf{y} \in \mathbb{R}^M$$

This approach reduces the problem of estimating the k coefficients $\beta_1, ..., \beta_k$ to the simpler problem of estimating the M coefficients $\theta_1, ..., \theta_M$, where $M < k \Rightarrow$ the dimension of the problem has been reduced from k to M.

PCR: Properties

The final PCR estimator of β based on the first M principal components is $\widehat{\beta}_{PCR} = V_M \widehat{\alpha}$. Then

$$C_{\widehat{\beta}_{PCR}} = \sigma^2 V_M (Z^T Z)^{-1} V_M^T = \sigma^2 \, V_M \Lambda_M^{-1} V_M^T = \sigma^2 \sum_{j=1}^{M} \frac{\mathbf{v}_j \mathbf{v}_j^T}{\lambda_j}. \qquad (8)$$

When $M = k$, we have $Z = XV$ and

$$\widehat{\beta}_{PCR} = V\widehat{\alpha} =$$

$$= V(Z^T Z)^{-1} Z^T \mathbf{y} = V(XV)^T (XV))^{-1} (XV)^T \mathbf{y}$$

$$= V(V^T X^T XV)^{-1} (XV)^T \mathbf{y} = V\Lambda^{-1} V^T X^T \mathbf{y}$$

But since V is orthogonal, $V\Lambda^{-1} V^T = (V^T \Lambda V)^T = (X^T X)^{-1}$. Thus

$$\widehat{\beta}_{PCR} = \widehat{\beta}$$

the PCR estimator is the ordinary LSE.

PCR: Properties

But then by (8)

$$C_{\widehat{\boldsymbol{\beta}}} = \sigma^2 \sum_{j=1}^{k} \frac{\mathbf{v}_j \mathbf{v}_j^T}{\lambda_j}. \tag{9}$$

For any i

$$\mathbf{Var}\left[\widehat{\beta}_j\right] = \sigma^2 \sum_{l=1}^{k} \frac{v_{il}^2}{\lambda_l^2}.$$

Here we see that small eigenvalues can destroy the precision of LSE.

PCR: Collinearity

In addition, since $Z = XV$,

$$Z_i = \sum_{l=1}^{k} \mathbf{v}_{li} \mathbf{x}_l,$$

where $\mathbf{x}_i$ is a column of X, and $\mathbf{v}_{li}$ are the elements of the ith column of an eigenvector of $X^T X$ Small variance of a pricipal component indicates that there are linear components of the original regressors that are almost constant.

Principal components regression (PCR)

Let $z_1, ..., z_M$ represent $M < k$ linear combinations of our original k regressors. That is

$$z_m = \sum_{j=1}^{p} \phi_{jm} x_j,$$

for some constants ϕ_{jm}, $j = 1, ..., p$, $m = 1, ..., M$. We can then fit the linear regression model

$$y_i = \sum_{m=1}^{M} \theta_m z_{im} + \mathrm{Var}epsilon_i, \quad i = 1, ..., n,$$

using least squares.
The key idea is that often a small number of principal components suffice to explain most of the variability in the data, as well as the relationship with the response.

If the constants ϕ_{jm} are chosen wisely, then such dimension reduction approaches can often outperform least squares regression.

However, the directions are identified in an unsupervised way, since the response is not used to help determine the principal component directions.

Consequently, PCR suffers from a drawback: there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.

Note that

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij},$$

where

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}.$$

Hence, we have a special case of the original linear regression model, with constrains on the β_j's coefficients.

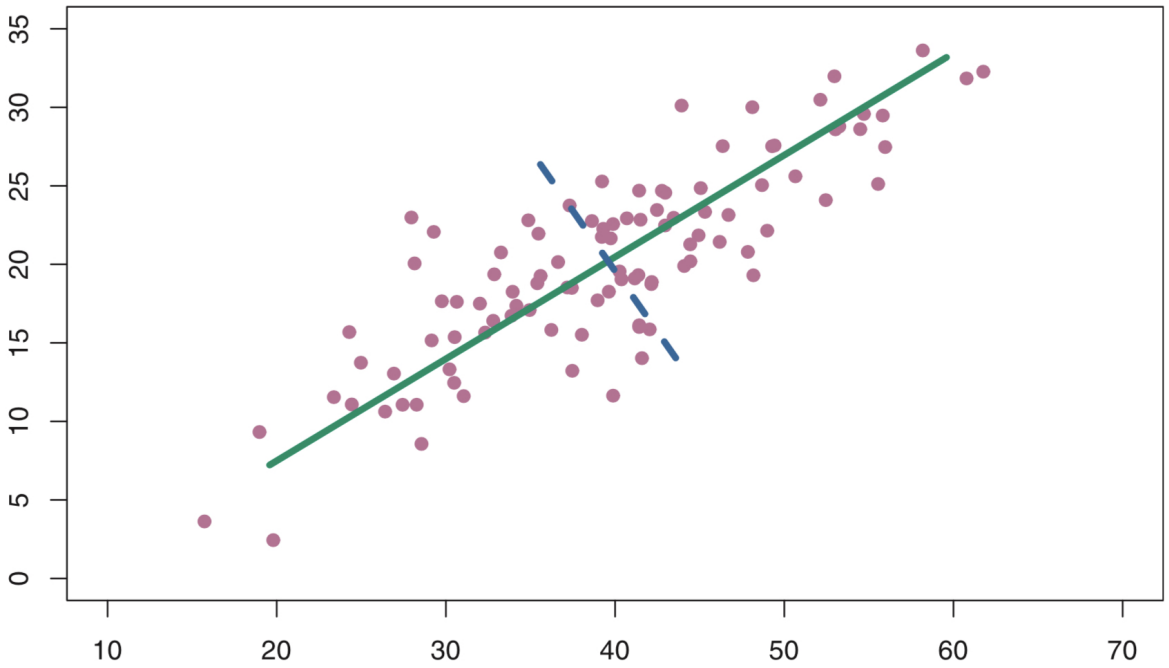
Note that

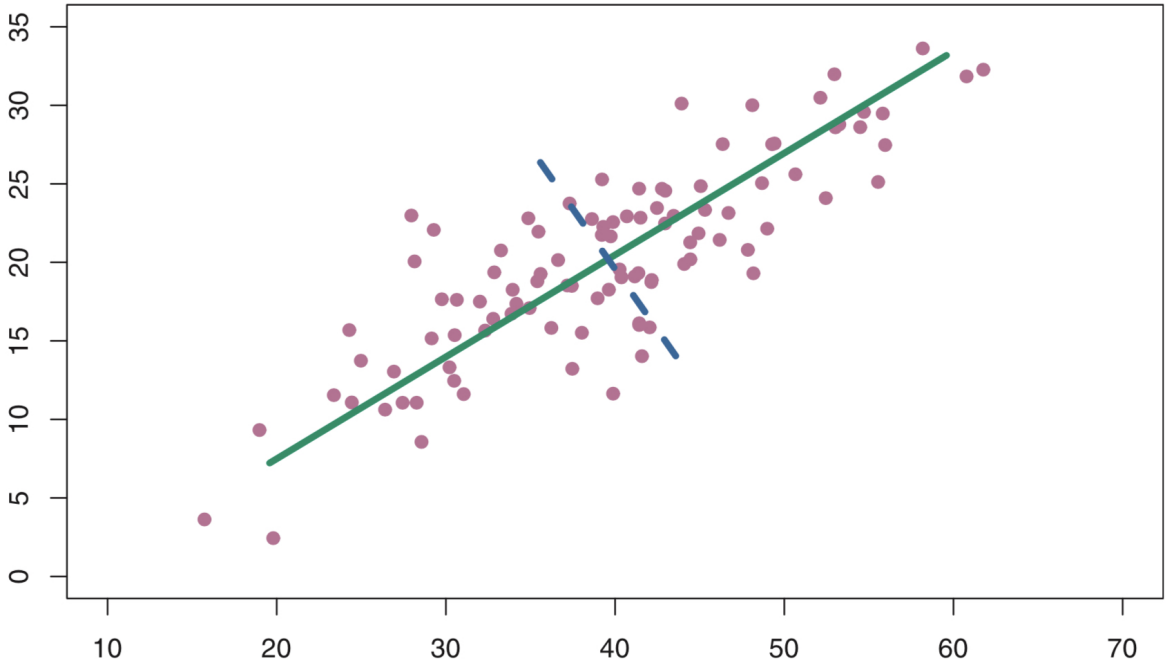
$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij},$$

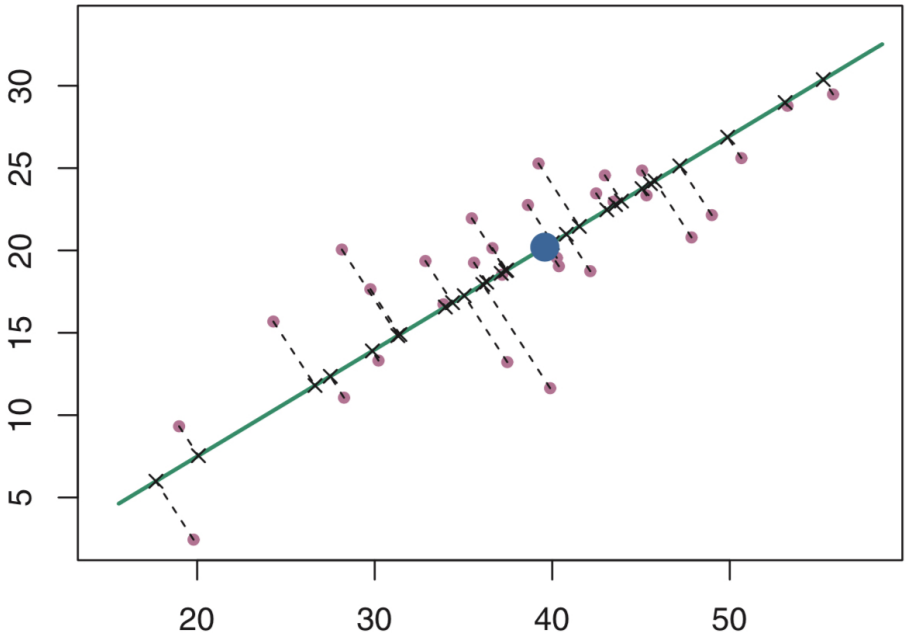
where

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}.$$

Hence, we have a special case of the original linear regression model, with constrains on the β_j's coefficients.
This constraint on the form of the coefficients has the potential to bias the coefficient estimates. However, in situations where p is large relative to n, selecting a value of $M \ll p$ can significantly reduce the variance of the fitted coefficients.


0
5
10
15
20
25
30
35
10
20
30
40
50
60
70


0
5
10
15
20
25
30
35
10
20
30
40
50
60
70


5
10
15
20
25
30
20
30
40
50

Choosing principal components

In PCR, the number of principal components, M, is typically chosen by cross-validation.
We note that even though PCR provides a simple way to perform regression using $M < p$ predictors, it is not a feature selection method.
This is because each of the M principal components used in the regression is a linear combination of all p of the original features.

Choosing principal components

In PCR, the number of principal components, M, is typically chosen by cross-validation.
We note that even though PCR provides a simple way to perform regression using $M < p$ predictors, it is not a feature selection method.
This is because each of the M principal components used in the regression is a linear combination of all p of the original features.
Therefore, while PCR often performs quite well in many practical settings, it does not result in the development of a model that relies upon a small set of the original features.
In this sense, PCR is more closely related to ridge regression than to the lasso.
In fact, one can show that PCR and ridge regression are very closely related. One can even think of ridge regression as a continuous version of PCR.

Standardizing the predictors

When performing PCR (also ridge regression), one generally recommend standardizing each predictor,

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \mathbf{a}rx_{\cdot j})^2}},$$

prior to generating the principal components. This standardization ensures that all variables are on the same scale.

Standardizing the predictors

When performing PCR (also ridge regression), one generally recommend standardizing each predictor,

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \mathbf{a}rx_{\cdot j})^2}},$$

prior to generating the principal components. This standardization ensures that all variables are on the same scale. In the absence of standardization, the high-variance variables will tend to play a larger role in the principal components obtained, and the scale on which the variables are measured will ultimately have an effect on the final PCR model. However, if the variables are all measured in the same units (say, kilograms, or inches), then one might choose not to standardize them.

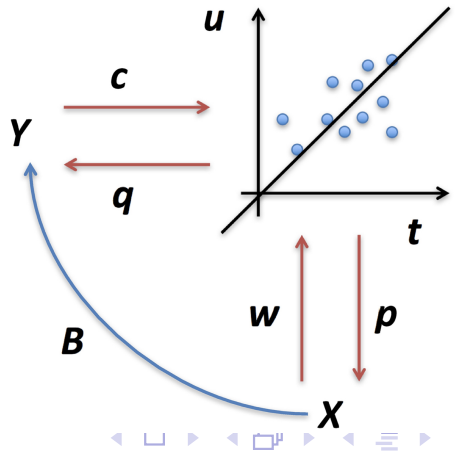
Partial least squares (PLS)

Partial least squares (PLS), is a supervised alternative to PCR. Like PCR, PLS is a dimension reduction method, which first identifies a new set of features that are linear combinations of the original features, and then fits a linear model via least squares using these new features.

Partial least squares (PLS)

Partial least squares (PLS), is a supervised alternative to PCR. Like PCR, PLS is a dimension reduction method, which first identifies a new set of features that are linear combinations of the original features, and then fits a linear model via least squares using these new features.
Unlike PCR, PLS identifies these new features in a supervised way, i.e., it makes use also of the response in order to identify new features that not only approximate the old features well, but also that are related to the response.

The PLS approach attempts to find directions that help explain both the response and the predictors.



Partial least squares (PLS) is a method for constructing **predictive** models when the factors are many and highly collinear.
Note that the emphasis is on predicting the responses and not necessarily on trying to understand the underlying relationship between the variables.
For example, PLS is **not** usually appropriate for screening out factors that have a negligible effect on the response.

Partial least squares (PLS) is a method for constructing **predictive** models when the factors are many and highly collinear.
Note that the emphasis is on predicting the responses and not necessarily on trying to understand the underlying relationship between the variables.
For example, PLS is **not** usually appropriate for screening out factors that have a negligible effect on the response.
However, when prediction is the goal and there is no practical need to limit the number of measured factors, PLS can be a useful tool.
The PLS algorithm works in the same fashion whether y is single response or multi-response.

We now describe how the first PLS direction is computed. After standardizing the p predictors, PLS computes the first direction z_1 by setting each ϕ_{j1} in

$$z_1 = \sum_{j=1}^{p} \phi_{j1} x_j,$$

equal to the coefficient from the simple linear regression of y onto x_j.

We now describe how the first PLS direction is computed.
After standardizing the p predictors, PLS computes the first direction z_1 by setting each ϕ_{j1} in

$$z_1 = \sum_{j=1}^{p} \phi_{j1} x_j,$$

equal to the coefficient from the simple linear regression of y onto x_j.
One can show that this coefficient is proportional to the correlation between y and x_j.
Hence, in computing z_1, PLS places the highest weight on the variables that are most strongly related to the response.

We now describe how the first PLS direction is computed.
After standardizing the p predictors, PLS computes the first direction z_1 by setting each ϕ_{j1} in

$$z_1 = \sum_{j=1}^{p} \phi_{j1} x_j,$$

equal to the coefficient from the simple linear regression of y onto x_j.
One can show that this coefficient is proportional to the correlation between y and x_j.
Hence, in computing z_1, PLS places the highest weight on the variables that are most strongly related to the response.
Often, the PLS direction does not fit the predictors as closely as does PCA, but it does a better job explaining the response.

To identify the second PLS direction we first adjust each of the variables for z_1, by regressing each variable on z_1 and taking residuals.
These residuals can be interpreted as the remaining information that has not been explained by the first PLS direction.
We then compute z_2 using this orthogonalized data in exactly the same fashion as z_1 was computed based on the original data.

To identify the second PLS direction we first adjust each of the variables for z_1, by regressing each variable on z_1 and taking residuals.
These residuals can be interpreted as the remaining information that has not been explained by the first PLS direction.
We then compute z_2 using this orthogonalized data in exactly the same fashion as z_1 was computed based on the original data.
This iterative approach can be repeated M times to identify multiple PLS components $z_1, ..., z_M$.
Finally, at the end of this procedure, we use least squares to fit a linear model to predict y using $z_1, ..., z_M$ in exactly the same fashion as for PCR.

As with PCR, the number M of partial least squares directions used in PLS is a tuning parameter that is typically chosen by cross-validation.
We generally standardize the predictors and response before performing PLS.

As with PCR, the number M of partial least squares directions used in PLS is a tuning parameter that is typically chosen by cross-validation.
We generally standardize the predictors and response before performing PLS.
PLS is popular in the field of chemometrics, where many variables arise from digitized spectrometry signals.
In practice it often performs no better than ridge regression or PCR.
While the supervised dimension reduction of PLS can reduce bias, it also has the potential to increase variance, so that the overall benefit of PLS relative to PCR is a wash.