

Dimension Reduction

Problem of high dimensional data, i.e., the number of observations is close or smaller than the number of features describing the objects (see James Gareth et al., pp 238-244):

- High likelihood of overfitting, i.e., the objects are perfectly predicted in the training sample.
- Models can no longer be meaningful interpreted (not so much an issue in ML)
- Several statistical methods like regression techniques will not work anymore due to multicollinearity.
- Features are highly redundant.
 - ⇒ The information in redundant features becomes highly inflated.
 - ⇒ Consequently, these features will substantially influence the outcome of supervised and unsupervised ML algorithms.
 - ⇒ Recall the `R`-script `ThinPlateOverfitReg.r`.

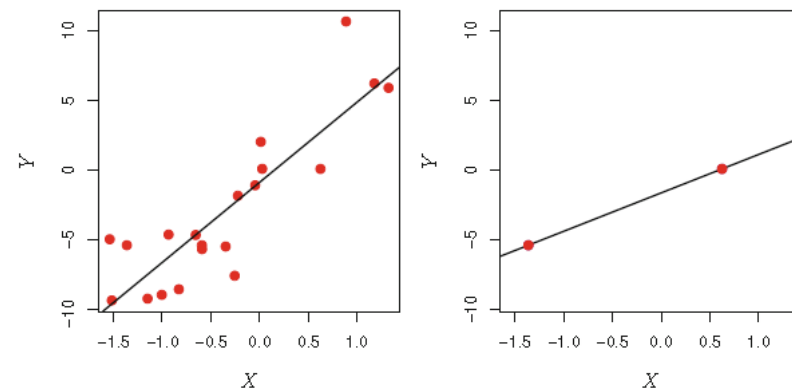


FIGURE 6.22. Left: Least squares regression in the low-dimensional setting. Right: Least squares regression with $n = 2$ observations and two parameters to be estimated (an intercept and a coefficient).

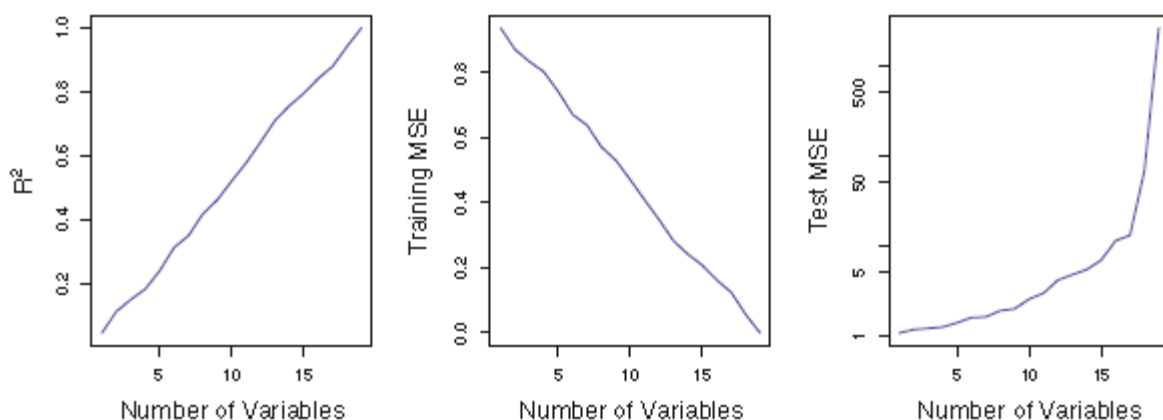


FIGURE 6.23. On a simulated example with $n = 20$ training observations, features that are completely unrelated to the outcome are added to the model. Left: The R^2 increases to 1 as more features are included. Center: The training set MSE decreases to 0 as more features are included. Right: The test set MSE increases as more features are included.

Methods of dimensionality reduction:

- Principal component analysis (PC) with applications in principal component regression and partial least squares regression.
- Uniform manifold approximation and projection (UMAP). See Burkov, pp 119-121.
- Autoencoders (Boehmke et al. Chapter 19)

Geometrical Interpretation of Principal Components and Eigenvalues

Review: z-transformation and the correlation matrix

- Let \mathbf{x} and \mathbf{y} be two random vectors with means \bar{x} and \bar{y} as well as standard deviations s_x^2 and s_y^2 , respectively.
- The z-transformation transforms the vector \mathbf{x} to \mathbf{z}_x and analog the vector \mathbf{y} to \mathbf{z}_y by

$$\mathbf{z} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \rightarrow \mathbf{z}_x = \begin{pmatrix} (x_1 - \bar{x}) / s_x \\ (x_2 - \bar{x}) / s_x \\ \vdots \\ (x_n - \bar{x}) / s_x \end{pmatrix}.$$

The new vectors \mathbf{z}_x and \mathbf{z}_y have a mean of zero and a standard deviation of one.

- The correlation between \mathbf{x} and \mathbf{y} is defined as

$$\begin{aligned} \text{corr}(\mathbf{x}, \mathbf{y}) &= \frac{1}{n} \cdot \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_x \cdot s_y} \\ &= \frac{1}{n} \cdot \sum_{i=1}^n ((x_i - \bar{x}) / s_x) \cdot ((y_i - \bar{y}) / s_y) \\ &= \frac{1}{n} \cdot \sum_{i=1}^n z_{xi} \cdot z_{yi} \\ &= \frac{1}{n} \cdot \mathbf{z}_x^T \cdot \mathbf{z}_y \end{aligned}$$

and can be written therefore in terms of the vectors of \mathbf{z}_x and \mathbf{z}_y .

- Let the p variables in a matrix $\mathbf{X}_{n \times p}$ be standardized by their means and standard deviations.

The resulting standardized matrix is $z(\mathbf{X}) \rightarrow \mathbf{Z}$.

- The **correlation matrix** \mathbf{R} among the p variables can then be easily calculated in terms of z-transformed variables by matrix multiplication:

$$\mathbf{R}_{p \times p} = \frac{1}{n} \cdot \mathbf{Z}^T \cdot \mathbf{Z}$$

- Usually, the correlation matrix provides the input to PC but alternatively also a covariance matrix can be used if the variance of individual features reflects their information content.

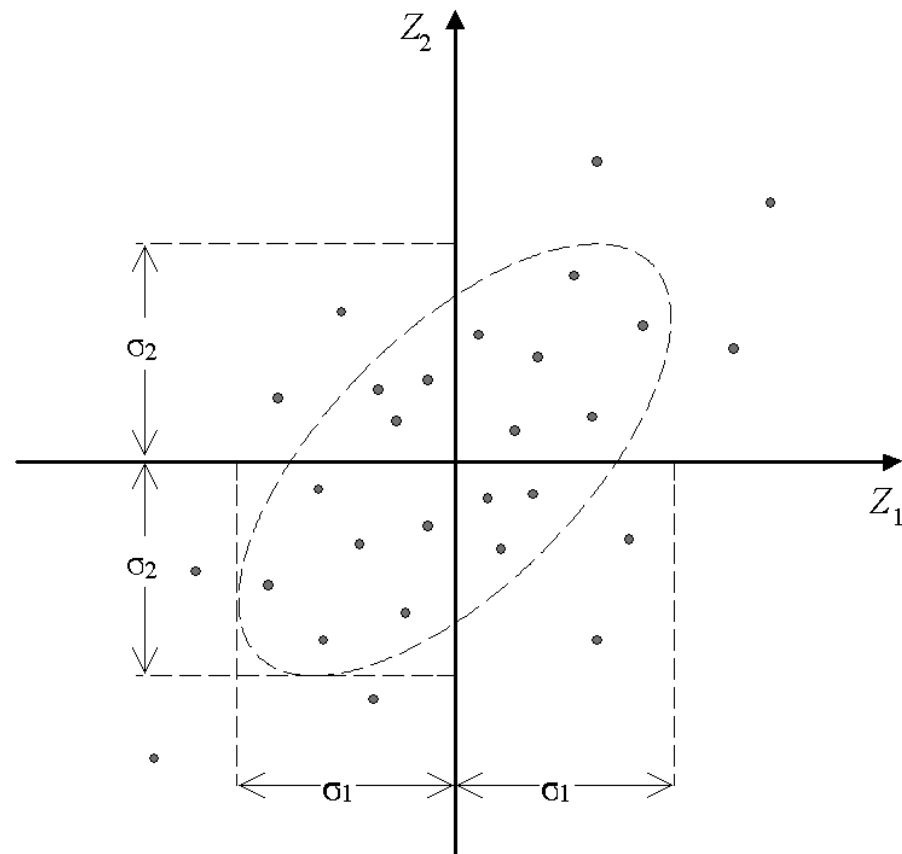


Figure 1: A scatter plot between two z-transformed variables

- Two positively correlated bivariate normal distributed variables
- Approximately 68% of the cases lie within the normal ellipse

- Due to the z-transformation:
 - The point (0,0) is in the center of the distribution
 - The spread along the Z_1 and Z_2 axes is $\sigma_1 = \sigma_2 = 1$.
- The Z_1 and Z_2 axes are orthogonal (both are at a rectangular or 90° angle).
- **Definition orthogonal:** let \mathbf{x} and \mathbf{y} be two vectors with identical number of components n .
 - If their cross-product $\mathbf{x}^T \cdot \mathbf{y}$ is zero, i.e., $\sum_{i=1}^n x_i \cdot y_i = 0$, then they are said to be orthogonal.
 - Corollary: If two z-transformed variables are orthogonal then they are also uncorrelated.
- The coordinates of the points (observations) are given by $\mathbf{Z} = \begin{pmatrix} z_{11} & \cdots & z_{n1} \\ z_{12} & \cdots & z_{n2} \end{pmatrix}^T$

Components as new Coordinate System by Orthogonal Rotation

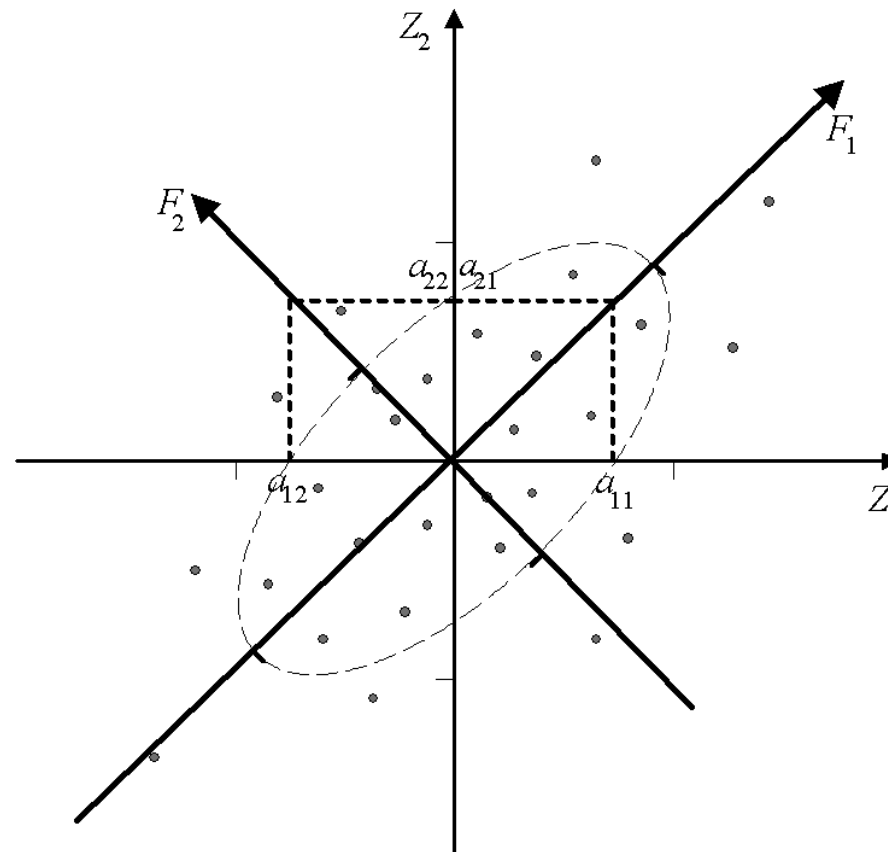


Figure 2: Rotation of the original reference system to new component system

- The first component axis fits the **main axis** of the ellipse. It captures **most of the variation** of the point cloud
- The second component axis is orthogonal to the first eigenvector axis and stretches along the **minor axis** of the ellipse. It captures the **remaining variation**.
- Scenario: What happens when both variables Z_1 and Z_2 are perfectly correlated?
 \Rightarrow One component axis captures all the variation and the other component axis captures none, i.e., it becomes irrelevant.
- The relationship between the old and the new coordinate system is indicated by the loading coefficients of the rotation matrix $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = (\mathbf{a}_1 | \mathbf{a}_2)$
- The loading coefficients have been standardized so that their lengths are $a_{11}^2 + a_{21}^2 = 1$ and $a_{12}^2 + a_{22}^2 = 1$
- Furthermore, the coefficients of \mathbf{a}_1 and \mathbf{a}_2 are orthogonal $a_{11} \cdot a_{12} + a_{21} \cdot a_{22} = 0$ or $\mathbf{a}_1^T \cdot \mathbf{a}_2 = 0$.
- Thus the rotation matrix \mathbf{A} is orthonormal: $\mathbf{A}^T \cdot \mathbf{A} = \mathbf{I}$

Coordinates of Points in New Coordinate System

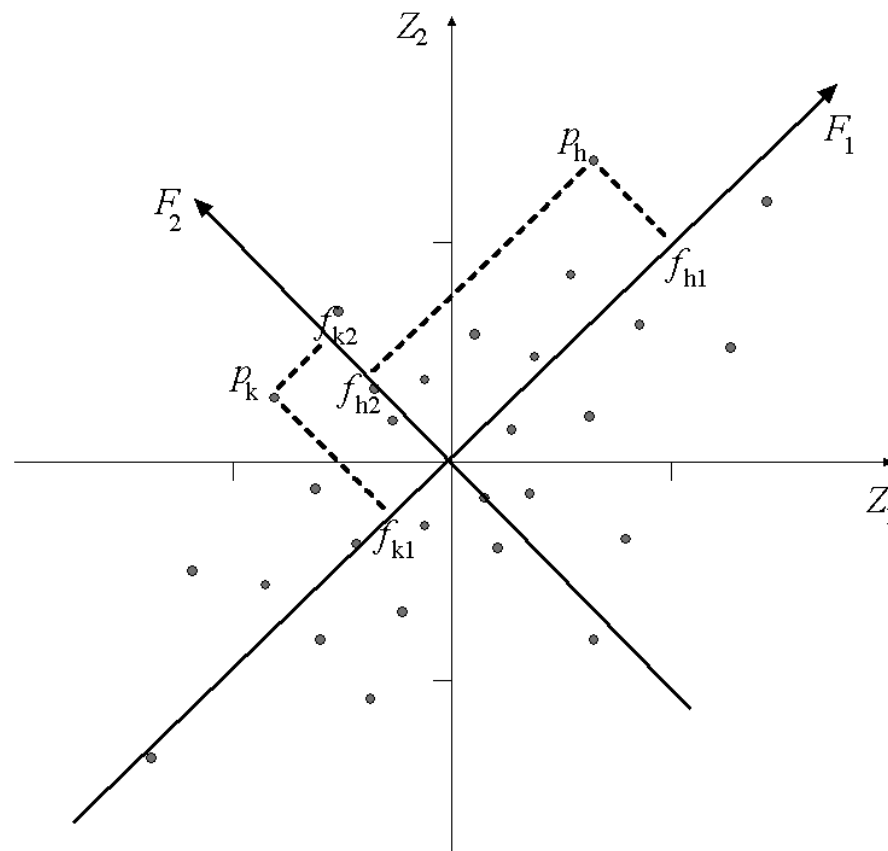


Figure 3: Coordinates (component scores) of points in the new component system

- Each data point $\mathbf{p}_i = (z_{i1}, z_{i2})^T$ in the **original** coordinate system has **new** but equivalent coordinates in the rotated component coordinate system $\mathbf{p}_i = (f_{i1}, f_{i2})^T$.
- The first component axis is expressed by $\mathbf{f}_1 = a_{11} \cdot \mathbf{z}_1 + a_{21} \cdot \mathbf{z}_2 = \mathbf{Z} \cdot \mathbf{a}_1$ and the second eigenvector axis is expressed by $\mathbf{f}_2 = a_{12} \cdot \mathbf{z}_1 + a_{22} \cdot \mathbf{z}_2 = \mathbf{Z} \cdot \mathbf{a}_2$, that is in matrix terms,

$$\begin{matrix} \mathbf{f}_1 & \mathbf{f}_2 & \mathbf{z}_1 & \mathbf{z}_2 \\ \begin{pmatrix} a_{11}z_{11} + a_{21}z_{12} & a_{12}z_{11} + a_{22}z_{12} \\ a_{11}z_{21} + a_{21}z_{22} & a_{12}z_{21} + a_{22}z_{22} \\ \vdots & \vdots \\ a_{11}z_{n1} + a_{21}z_{n2} & a_{12}z_{n1} + a_{22}z_{n2} \end{pmatrix} & = & \begin{pmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \\ \vdots & \vdots \\ z_{n1} & z_{n2} \end{pmatrix} \cdot \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \end{matrix}$$

- The two vectors \mathbf{f}_1 and \mathbf{f}_2 are uncorrelated because they are orthogonal $\mathbf{f}_1^T \cdot \mathbf{f}_2 = 0$ and centered around zero.

Identification of new Component System through Constrained Optimization

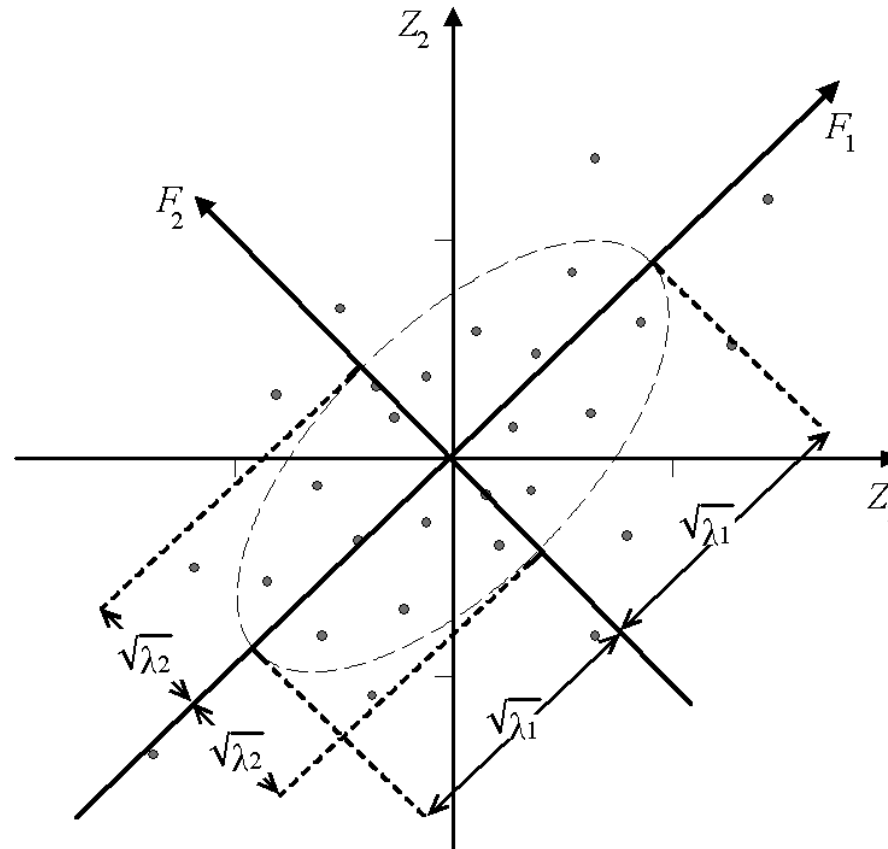


Figure 4: Eigenvalues and variance of along component axes

- The eigenvalues λ_1 and λ_2 measure the spread of the ellipse along both component axes, that is, the variance of the data points along the first component axis is $Var(\mathbf{f}_1) = \lambda_1$ and along the second component axis it is $Var(\mathbf{f}_2) = \lambda_2$.
- The eigenvalues are distinctly ordered with $\lambda_1 \geq \lambda_2 \geq 0$
- The eigenvectors is calculated such that
 - $\max_{a_{11}, a_{21}} Var(\mathbf{f}_1)$ for the first component and
 - the second component is calculate such that $\max_{a_{12}, a_{22}} Var(\mathbf{f}_2)$ subject to $\mathbf{f}_1^T \cdot \mathbf{f}_2 = 0$
- The eigenvalues sum to $\lambda_1 + \lambda_2 = Var(Z_1) + Var(Z_2) = 2$, which is equal to the **trace** of the 2×2 correlation matrix \mathbf{R} . Therefore, no variation is **lost or added** in the rotated component coordinate system.
- Because the correlation matrix \mathbf{R} between \mathbf{z}_1 and \mathbf{z}_2 is **positive definite**, both eigenvalues are greater than zero (or in an extreme case equal to zero).
- The **determinate** of the correlation matrix \mathbf{R} equals the product of the eigenvalues.

- For more than $p = 2$ original variables we would get the new coordinates $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_p$ with $Var(\mathbf{f}_1) = \lambda_1, Var(\mathbf{f}_2) = \lambda_2, \dots, Var(\mathbf{f}_p) = \lambda_p$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$

Geometrical Interpretation of Principal Components, Eigenvalues, and Eigenvectors

Objectives of Principal Component Analysis:

- Principal component analysis is an ***exploratory data driven*** method.
 - That is, it is not based on an ***underlying conceptional model*** and depends solely on the ***observed data***.
- It seeks a few ***underlying dimensions*** (the components) that explain the ***pattern of covariation*** in the correlation matrix.
 - ⇒ It follows the concept of parsimony (complexity reduction, suppression of random noise and simplification)
- The ***component loadings*** define the ***meaning*** of these underlying dimensions with regards to the original variables.

⇒ They actually are the **correlation** between the original variables \mathbf{z}_l with the component \mathbf{f}_k , i.e., $\text{corr}(\mathbf{z}_l, \mathbf{f}_k) = a_{lk}$.

- In order to make a component \mathbf{f}_k easy to interpret, it should load high on a few variables and zero on the remaining variables. Thus the component can be interpreted in terms of the factor.
- The underlying dimensions are generally supposed to be **independent** (uncorrelated) from each other.
- Consequently, they become useful as input to other methods such as regression analysis or cluster analysis.
Review questions: Why are uncorrelated variables in regression analysis preferred?
- Potential applications:
 - Analysis of **potentially redundancy in data** leading to dimensionality reduction.

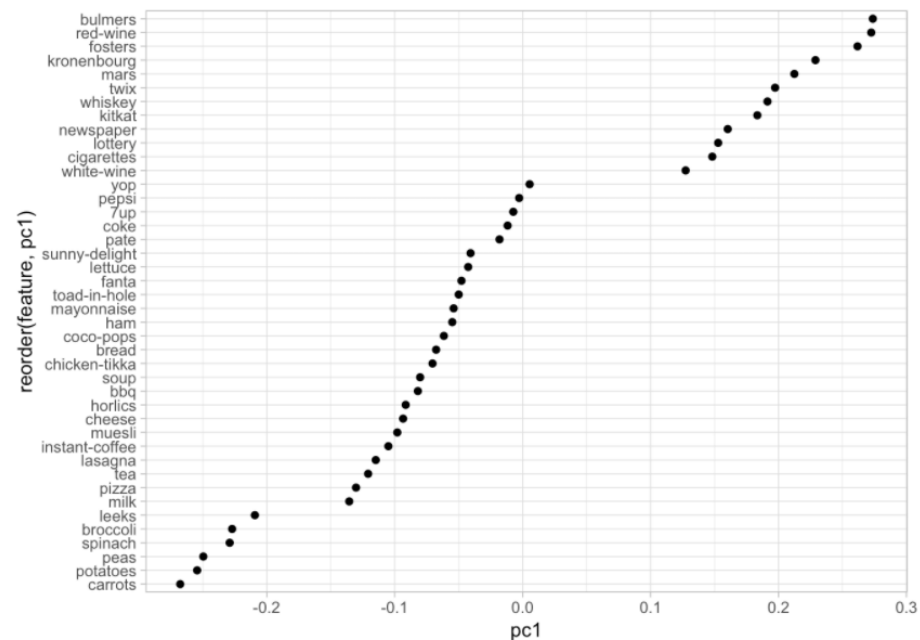


Figure 17.3: Feature loadings illustrating the influence that each variable has on the first principal component.

- Identification of ***underlying mechanism*** that has generated our observed data.
- In ***remote sensing*** to combine several redundant spectral bands, etc.

Model Structure:

- The correlation matrix among variables is in the center of the PC method.
The correlation coefficient reflects the ***relationship*** among the observed pairs of variables.
PC analysis ***evaluates the internal structure*** in the correlation matrix. (this allows to distinguish and label components)
- If the correlation matrix does not have an internal structure, such as for ***mutually independent variables***, dimensionality reduction become meaningless.
- The underlying dimensions (components) are derived as combinations of the observed variables.
Underlying dimensions can only ***indirectly*** be observed by investigating their relationships among the original variables.
- ***Reversal of the perspective***: In the geometrical introduction, the components were combinations of the variables $\mathbf{F} = \mathbf{Z} \cdot \mathbf{A}$;
in the reversal, the original variables are modeled by a set of components:
 $\Rightarrow \mathbf{Z} = \mathbf{F} \cdot \mathbf{A}^T$ because for rotation matrices $\mathbf{A}^T = \mathbf{A}^{-1}$

- Principal component analysis **resembles a regression model** of an observed variable being regressed on a set of underlying components.

Selection of the number of Components:

- No direct statistical guidelines (i.e., significance tests) are available because principal component analysis is not a statistical model based on a statistical distribution model (no direct distribution assumptions about the underlying distributions are made).
- Some guidelines are:
 - None of the components should explain **less than** the variation that a single variable captures.
⇒ Consequently, do not consider components that have an eigenvalue $\lambda_j < 1$.
 - Generate a **scree-plot**. For datasets with many variables, one may observe

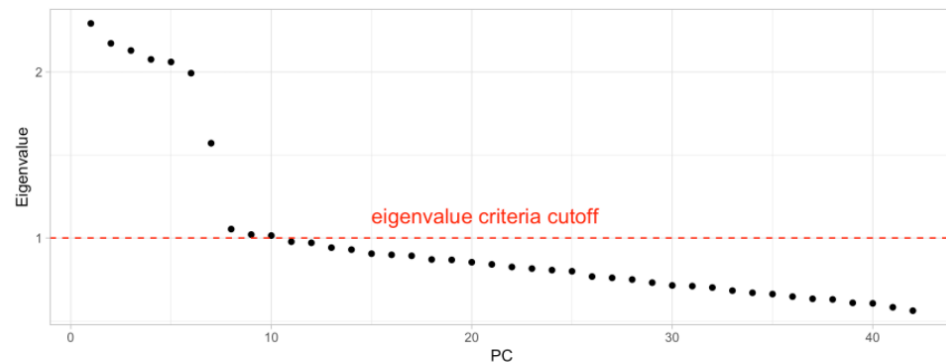


Figure 17.5: Eigenvalue criterion keeps all principal components where the sum of the eigenvalues are above or equal to a value of one.

a clear discontinuity in the decreasing sequence of the eigenvalues. Use all components prior to that discontinuity.

- See the example **basinFactorComp.r** as an example for a basic factor analysis using principal components.

Principal Components in Regression Analysis

PC is unsupervised by an outcome variable y .

- PC determines independent linear combinations of the original features that best capture the variability in a set of features.
- Since the outcome variable y is not involved in this process of generating linear combinations, there is no guarantee that the best combination set (these are the principal components) will be found that predict y .
- PC often turns out to be a reasonably good approximation of the relevant predictor features.
- PC dimension reduction lowers the **variance of the predictor function** in supervised learning and only **marginally increases the bias** by missing relevant predictors.

- The ***number of variables*** used or the ***number of their associated principal components*** are hyper- or tuning parameters selected by the analyst.

Partial least squares regression

- Partial least squares regression overcomes the unsupervised nature of principal component regression by defining the components with the fit to the target variable in mind.
- The first component is defined by maximizing the explanatory power. The loading coefficients of the individual variables are proportional to the correlation between the features and the target variable.
- To identify the second component first the residuals for each feature regressed on the first component are calculated. These residuals are uncorrelated with the first component. Then the feature coefficients for the second component are again proportional to the residual correlations and the target variable.
- In subsequent steps residuals are calculated for all features using the identified components as regressors.

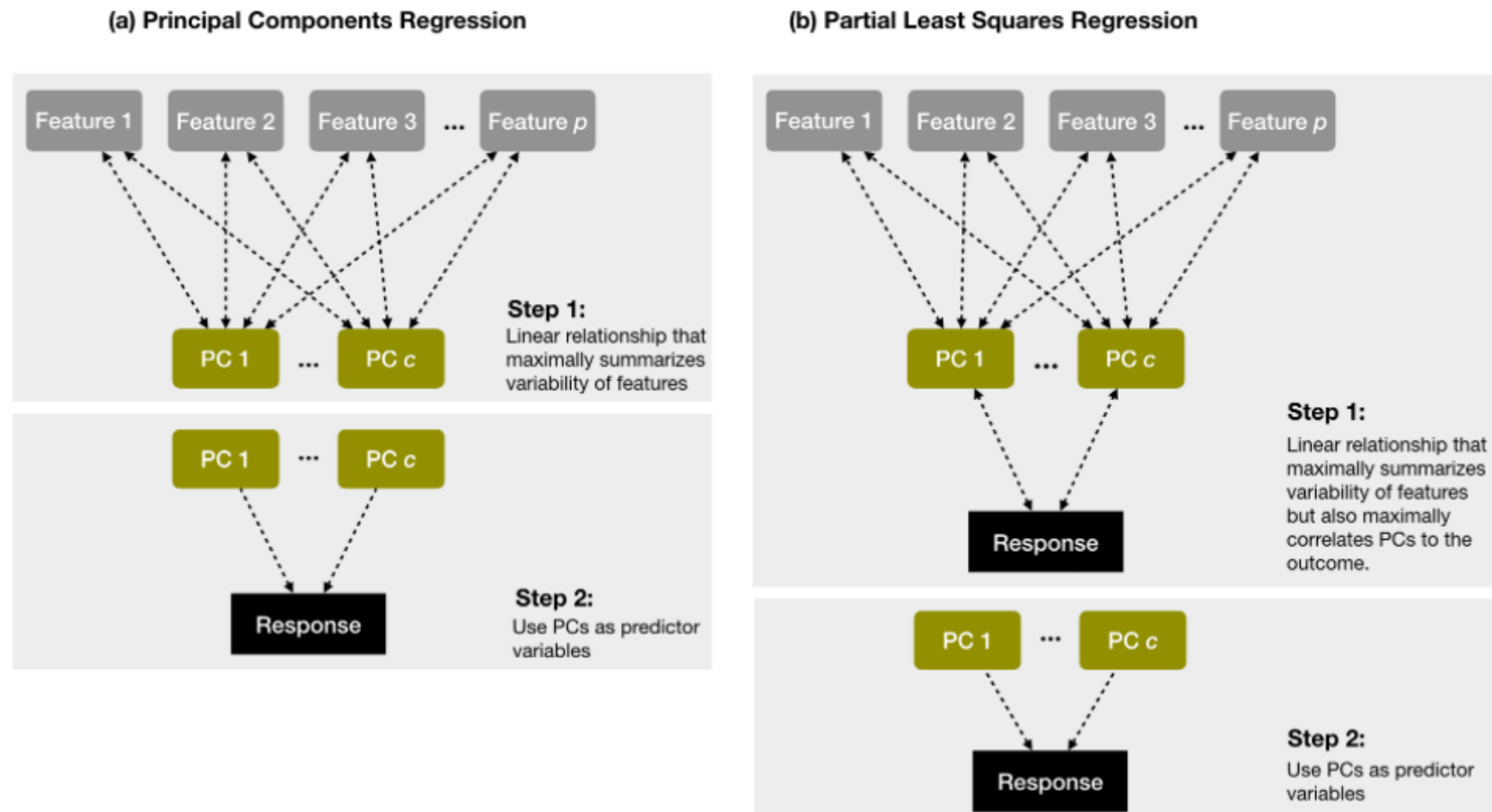


Figure 4.8: A diagram depicting the differences between PCR (left) and PLS (right). PCR finds principal components (PCs) that maximally summarize the features independent of the response variable and then uses those PCs as predictor variables. PLS finds components that simultaneously summarize variation of the predictors while being optimally correlated with the outcome and then uses those PCs as predictors.

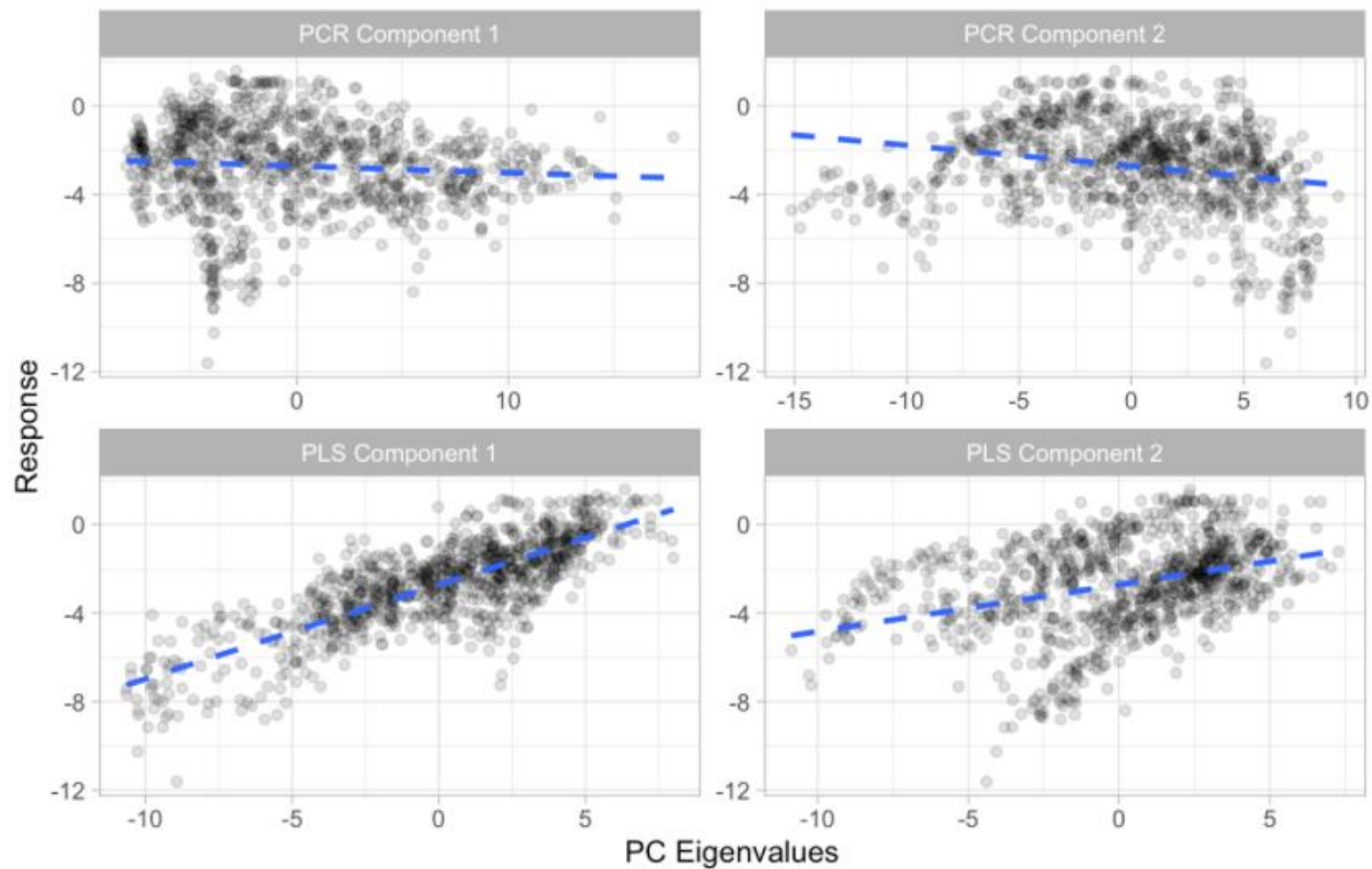


Figure 4.9: Illustration showing that the first two PCs when using PCR have very little relationship to the response variable (top row); however, the first two PCs when using PLS have a much stronger association to the response (bottom row).