CENTRALE LYON

# Statistical Learning

## Course 3 - Principal Component and PLS Regression

Ecole Centrale de Lyon - Bachelor 2nd Year
2024-2025

Come back on linear models of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

Aims of selecting sparser models which are simpler to interpret and also have better predictive performance in comparison to the model that includes all $p$ predictors. Specifically, we saw

- Model-search methods: discrete procedures which require full or partial enumeration of all possible models and select one best model.
- Coefficient shrinkage methods: continuous procedures which shrink the coefficients of the full model to zero (penalised regression eg Lasso and Ridge regression).

Both approaches effectively reduce the dimensionality of the problem.

# Dimension Reduction

There is another alternative strategy to standard LS regression. It is quite different than subset selection and penalised regression. The idea is to reduce dimensionality before applying LS. Instead of utilizing the $p$ original predictors, this approach utilises $q$ transformed variables which are linear combinations of the original predictors, where $q < p$.

Idea: Transform the predictor variables and then fit a Least Squares model using the transformed variables. The number of transformed variables is smaller than the number of predictors.

# Dimension reduction-continued

1. Define $q < p$ linear transformations $Z_k, k = 1, \cdots, q$ of $X_1, X_2, \cdots, X_p$ as

$$Z_k = \sum_{j=1}^{p} \phi_{jk} X_j, \tag{1}$$

for some constants $\phi_{1k}, \phi_{2k}, \cdots, \phi_{pk}$.

2. Fit a LS regression model of the form

$$y_i = \theta_0 + \sum_{k=1}^{q} \theta_k z_{ik} + \epsilon_i, \ i = 1, \cdots, n. \tag{2}$$

# Dimension reduction-continued

- Notice that from definition 1,

$$\sum_{m=1}^{q} \theta_m z_{im} = sum_{m=1}^{q} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{q} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij} \tag{3}$$

where

$$\beta_j = \sum_{m=1}^{q} \theta_m \phi_{jm}. \tag{4}$$

- Hence model 2 can be thought of as a special case of the original linear regression model.
- Dimension reduction serves to constrain the estimated $\beta_j$ coefficients, since now they must take the form 4.
- Can win in the bias-variance tradeoff.

Question: How should we choose the $\phi$'s?

Well, we need to find transformations $Z_k = \sum_{j=1}^{p} \phi_{jk} X_j$ which are meaningful. Trying to directly approach the problem with respect to the constants $\phi_{1k}, \phi_{2k}, \cdots, \phi_{qk}$ is difficult.

It is easier to approach it with respect to the $Z$ variables. Specifically, to consider what properties we would like the $Z$'s to have.

Desired properties:

- If the new variables $Z$ are uncorrelated this would give a solution to the problem of multicollinearity.
- If the new variables $Z$ capture most of the variability of the $X$'s, and assuming that this variability is predictive of the response, the $Z$ variables will be reliable predictors of the response.

This is exactly what principal component regression (PCR) does!

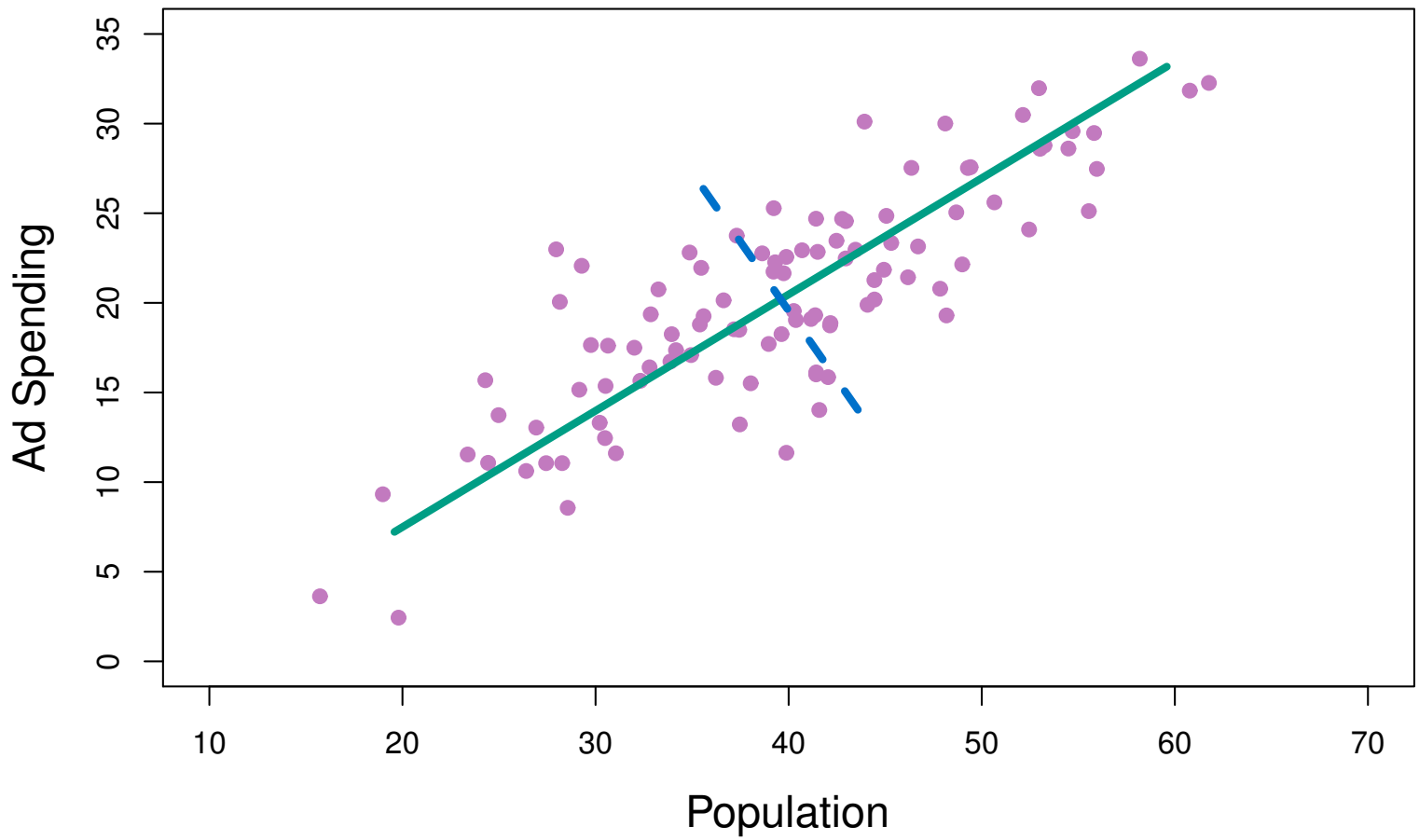
# Principal Component Analysis

PCR utilises Principal Component Analysis (PCA) of the data matrix $X$.

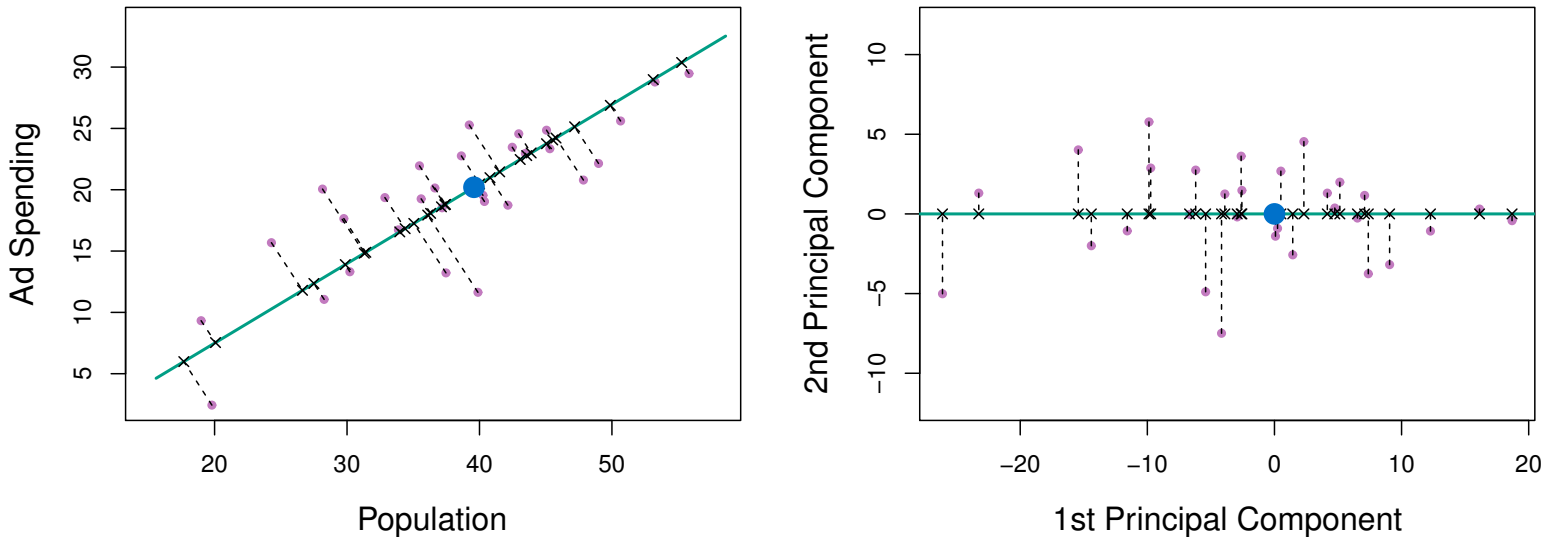
- The first principal component is that (normalized) linear combination of the variables with the largest variance.
- The second principal component has largest variance, subject to being uncorrelated with the first.
- And so on.
- Hence with many correlated original variables, we replace them with a small set of principal components that capture their joint variation.

Technical note: If the predictors are not on the same scale, the columns of matrix X should be standardised before implementing PCA.

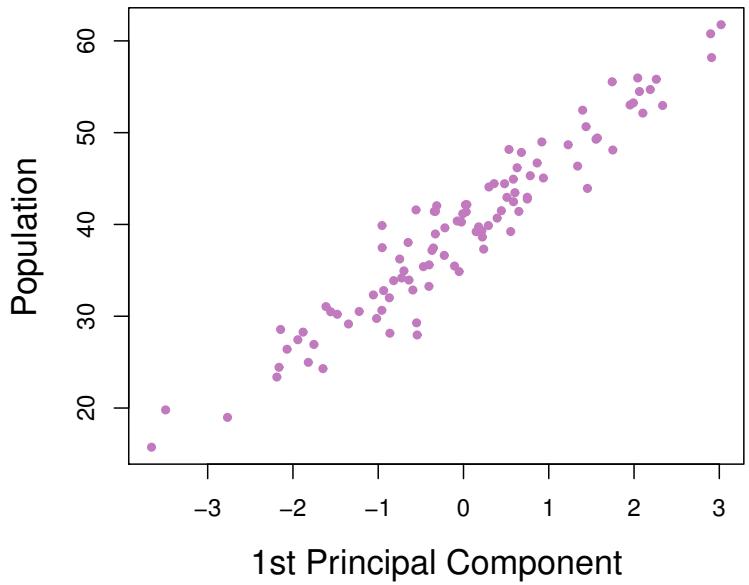
# Example: Advertisement Data

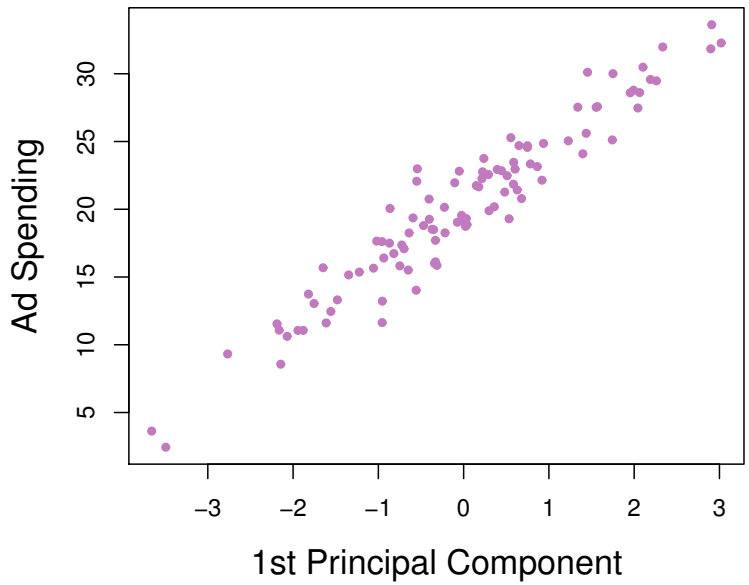



The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component

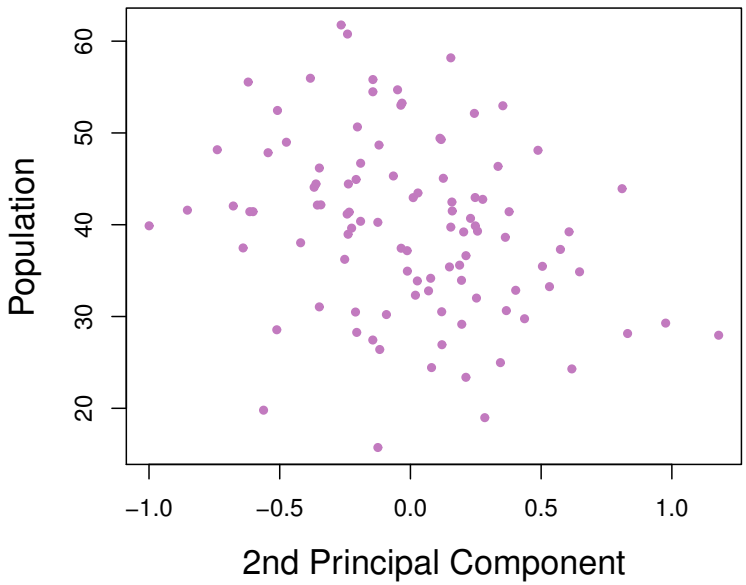


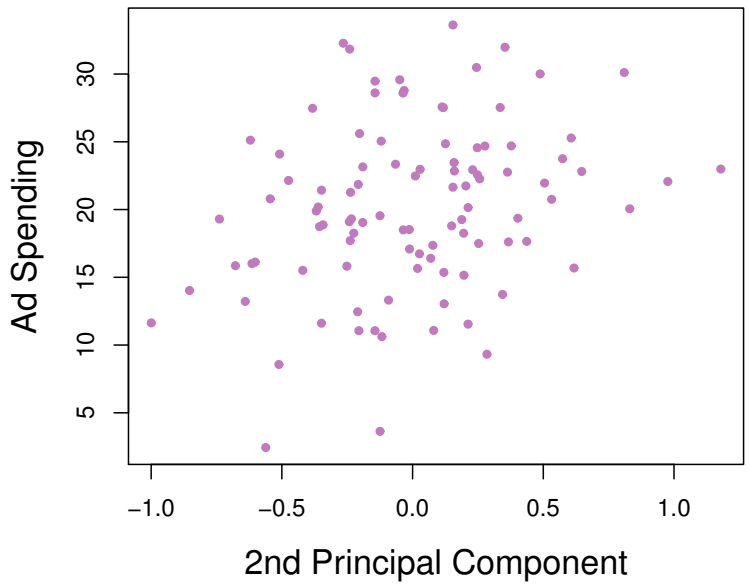

A subset of the advertising data. Left: The first principal component, chosen to minimize the sum of the squared perpendicular distances to each point, is shown in green. These distances are represented using the black dashed line segments. Right: The left-hand panel has been rotated so that the first principal component lies on the x-axis.




Plots of the first principal component scores zi1 versus pop and ad. The relationships are strong.




Plots of the second principal component scores zi2 versus pop and ad. The relationships are weak.

# Principal Component Regression

Principal Component Regression (PCR) involves constructing the first $q$ principal components $Z_1, ..., Z_q$, and then using these components as the predictors in a linear regression model that is fit using least squares.

The key idea is that often a small number of principal components suffice to explain most of the variability of the data, as well as the relationship with the response. In other words, we assume that the directions in which $X_1, ..., X_p$ show the most variation are the directions that are associated with $Y$. This assumption is not guaranteed, but it is often reasonable enough to give good results.

# How Many Components?

Once again we have a tuning problem. In ridge and lasso the tuning parameter ($\lambda$) was continuous - in PCR the tuning parameter is the number of components ($q$), which is discrete.

- Some methods calculate the total variance of $X$ explained as we add further components. When the incremental increase is negligible, we stop. One such popular method is the scree plot (see the practical demonstration extension).
- Alternatively, we can use cross-validation!

# What about Interpretability?

One drawback of PCR is that it lacks interpretability, because the estimates $\hat{\theta}_k$

for $k = 1, \cdots, q$ are the coefficients of the principal components, which have no meaningful interpretation. Well, that is partially true, but there is one thing we can do: once we fit LS on the transformed variables we can use the constraint in order to obtain the corresponding estimates of the original predictors $\hat{\beta}_j = \sum_{k=1}^{q} \hat{\theta}_k \phi_{jk}$, for $j = 1, \cdots, p$ However, for $q < p$ these will not correspond to the LS estimates of the full model. In fact, the PCR estimates get shrunk in a discrete step-wise fashion as $q$ decreases (this is why PCR is still a shrinkage method).

# PCR - a Summary

- Advantages:
  - Similar to ridge and lasso, it can result in improved predictive performance in comparison to Least Squares by resolving problems caused by multicollinearity and/or large $p$.
  - It is a simple two-step procedure which first reduces dimensionality from $p+1$ to $q+1$ and then utilizes LS.
- Disadvantages:
  - It does not perform feature selection.
  - The issue of interpretability.
  - That is, the response does not supervise the identification of the principal components.
  - there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.

## Final Comments:

- In general, PCR will tend to do well in cases when the first few principal components are sufficient to capture most of the variation in the predictors as well as the relationship with the response.
- We note that even though PCR provides a simple way to perform regression using $q < p$ predictors, it is not a feature selection method. This is because each of the $q$ principal components used in the regression is a linear combination of all $p$ of the original features.
- In general, dimension-reduction based regression methods are not very popular nowadays, mainly due to the fact that they do not offer any significant advantage over penalised regression methods.
- However, PCA is a very important tool in unsupervised learning where it is used extensively in a variety of applications!

# Partial Least Squares

- Like PCR, PLS is a dimension reduction method, which first identifies a new set of features $Z_1, \cdots, Z_M$ that are linear combinations of the original features, and then fits a linear model via OLS using these $M$ new features.
- But unlike PCR, PLS identifies these new features in a supervised way - that is, it makes use of the response $Y$ in order to identify new features that not only approximate the old features well, but also that are related to the response.
- Roughly speaking, the PLS approach attempts to find directions that help explain both the response and the predictors

# Details of Partial Least Squares

- After standardizing the $p$ predictors, PLS computes the first direction $Z_1$ by setting each $\phi_{1j}$ in 1 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 $X_j$ and the loading vecto $\phi$ maximize

$$\phi_1 = argmax_{||\phi||=1}(Cov(X\phi, y))$$

- Hence, in computing $Z_1 = \sum_{j=1}^{p} \phi_{1j} X_j$ , PLS places the highest weight on the variables that are most strongly related to the response.
- Regess the outcome $Y$ on $Z_1$ to obtain $\theta_1$.
- Orthogonalize $X_1, \cdots X_p$ with respect to $Z_1$.

# Details of Partial Least Squares

- Subsequent directions are found by taking residuals and then repeating the above prescription until $M \leq p$ directions are obtained.
- PLS produces a sequence orthogonal directions $Z_1, \cdots, Z_M$.

# Algorithm Partial Least Squares

- Standardize the variables $x_j$ for $j = 1, \cdots, p$ . Initialize $\hat{y}^{(0)} = \bar{y}1$ and $x_j^{(0)} = xj, j = 1, \cdots, p$.
- For $k = 1, 2, \cdots, p$
  - $$z_k = \sum_{j=1}^{p} \hat{\phi}_{kj} x_j^{k-1}$$

    where où $\hat{\phi}_{kj} =< x_j^{k-1}, y >$,
  - $$\hat{\theta}_{(}k) = \frac{< z_k, y >}{< z_k, z_k >}$$
  - Orthogonalization of $x_j$ with respect to $z_k$

$$x_j^{(k)} = x_j^{(k-1)} - \frac{< z_k, x_j^{(k-1)} >}{< z_k, z_k >} z_k$$