Principal Components Analysis

Assume the data matrix $\mathbf{X}_{n\times p}$ (without the intercept column) has been standardized (each column has mean zero and variance 1[a]).

Question: Find a linear combination of the p features, such that the projection of the data along that direction $\mathbf{a} \in \mathbb{R}^p$ has the largest variation. That is,

$$\max_{\mathbf{a}\in\mathbb{R}^p} \text{Sample Var}(\mathbf{Xa}), \text{ subj to } \|\mathbf{a}\|^2 = 1.$$

Since Sample Var$(\mathbf{Xa}) \propto \mathbf{a}^t\mathbf{X}^t\mathbf{X}\mathbf{a}^t$, the solution of $\mathbf{a}$ is equal to the top eigenvector (i.e., the one with the largest eigenvalue) of matrix $\mathbf{X}^t\mathbf{X}$.

[a] In some applications, columns of $\mathbf{X}$ are just centered, but won't be scaled.

Recall the SVD of $\mathbf{X}$: $\mathbf{X} = \mathbf{U}_{n\times p}\mathbf{D}_{p\times p}\mathbf{V}_{p\times p}$, where $\mathbf{D}$ is a diagonal matrix $\text{diag}(d_j)_{j=1}^p$ with d_j's being the singular values of $\mathbf{X}$ and arranged in decreasing order $d_1 \geq d_2 \geq \cdots$.

It turns out the solution is the 1st column of $\mathbf{V}$, known as the first principal component (PC) direction, along which the data vary the most (i.e., has the largest variation). The projection of the data onto the 1st PC direction, $F_1 = \mathbf{X}\mathbf{V}[, 1]$ is called the first principal component.

The variance explained by the 1st PC direction is equal to d_1^2.

We can then find the next optimal direction (i.e., it optimizes the variance and is orthogonal to the 1st direction), which turns out to be the 2nd column of $\mathbf{V}$.

In general, one can construct up to q PC directions, which is the same as the rank of the design matrix $\mathbf{X}_{n\times p}$. If we assume $\mathbf{X}$ is of full rank, then $p = q$.

We usually graph PVE or Cumulate PVE vs the number of PCs. PVE (proportion of variance explained) by the m-th PC is

$$\frac{d_m^2}{d_1^2 + \cdots + d_m^2 + \cdots + d_p^2};$$

Cumulative PVE by the top m PCs

$$\frac{d_1^2 + \cdots + d_m^2}{d_1^2 + \cdots + d_m^2 + \cdots + d_p^2}.$$

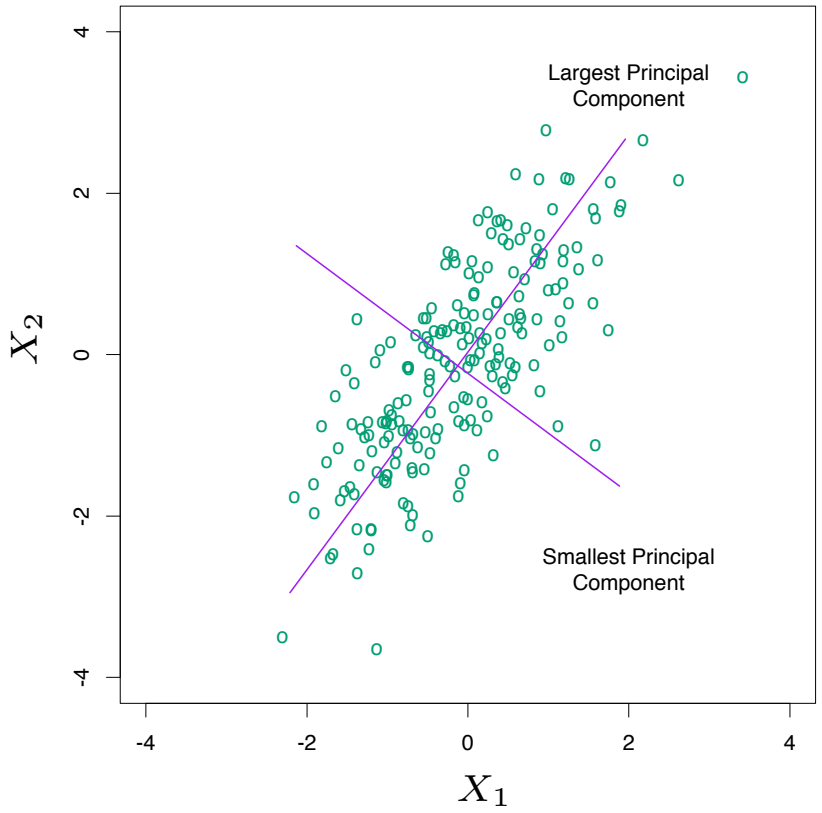

FIGURE 3.9. *Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects* $\mathbf{y}$ *onto these components, and then shrinks the coefficients of the low-variance components more than the high-variance components.*

Another View of PCA: Best Low-Rank Approximation

Recall that we have n measurements on each of the p variables, $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, which are the n rows of the design matrix $\mathbf{X}_{n \times p}$.

Suppose we want to find an approximation of the original set of n points in $\mathbb{R}^p$ by a m-dimensional linear subspace. WLOG, denote a set of orthogonal basis vectors for this m-dim subspace by $\mathbf{v}_1, \ldots, \mathbf{v}_m$.

$$\min_{\mathbf{A},\mathbf{V}} \sum_{i=1}^{n} \|\mathbf{x}_i - (a_{i1}\mathbf{v}_1 + a_{i2}\mathbf{v}_2 + \cdots a_{im}\mathbf{v}_m)\|^2,$$

$$\Longrightarrow \quad \min_{\mathbf{A},\mathbf{V}} \|\mathbf{X} - \mathbf{A}\mathbf{V}^t\|^2$$

where $\mathbf{A}_{n \times m}$ with its (i,j)th entry being a_{ij}, and $\mathbf{V}_{p \times m}$ with its j-th column being $\mathbf{v}_j$ is an orthogonal matrix.

It turns out that the the solution is: $\mathbf{V}_{p\times m}$ is the first m PC directions and $\mathbf{A}$ is the first m PCs.

So in other words, the top m PCs can be viewed as the best rank m approximation of the original data, and the variance explained by the top m PCs can be interpreted as the approximation error.

Principal Components Regression

Create a new design matrix $\mathbf{F} = \mathbf{XV}[, 1 : m]$; Column of F are the top m PCs. And the new regression model is

$$\mathbf{y} = \text{intercept} + \mathbf{F}\boldsymbol{\alpha}_{m\times 1} + \text{ err.}$$

Each new variable (corresponding to one column of $\mathbf{F}$) is a linear combination of all the p original variables. Also note that even if each column of the original design matrix $\mathbf{X}$ has been scaled to have the same variance, the columns of the new design matrix $\mathbf{F}$ will not have the same variance.

If we use all the PCs, the LS fitting from the new design matrix $\mathbf{F}$ should be the same the original regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \text{ err.}$$

Advantages of PCR

- Necessary dimension reduction when $p \geq n$.
- Orthogonal design matrix makes computation easier and reduces correlations among the new predictors.

  Recall that the key issue for subset selection for linear models: total number of candidate models is 2^p, which grows exponentially fast with p.

  But if the design matrix is orthogonal, then we can reduce the number of candidate models to p. Why?

Criticism Against PCR

- How many components? The number of PCs are often chosen by eye-balling the plot of PVE vs. the number of PCs (scree plot) to find the so-called elbow point, at which the slope of the curve goes from "steep" to "flat", which is quite *ad hoc*.
- The primary goal of regression is to predict Y. The information of Y is never used in extracting PCs. How do we know the top 3 PCs are relevant to the prediction of Y?
- If the goal is to find a sparse model involving only a small fraction of the original p features, one cannot use PCR, since each newly derived feature (Z_j) depends on all the original p features.

Evidence supporting PCR

- Work very well in practice (e.g., on image data).

In practice

- Useful for applications where we believe that a considerable amount of variation in $\mathbf{X}$ is due to Y.
- Do not just set $m = 5$ or use the elbow point.
- Pick top m PCs which explain at least 95% of the variation, and then use some model selection criteria or CV to select m.
- Try a combination of PCR and other variable selection/shrinkage methods.