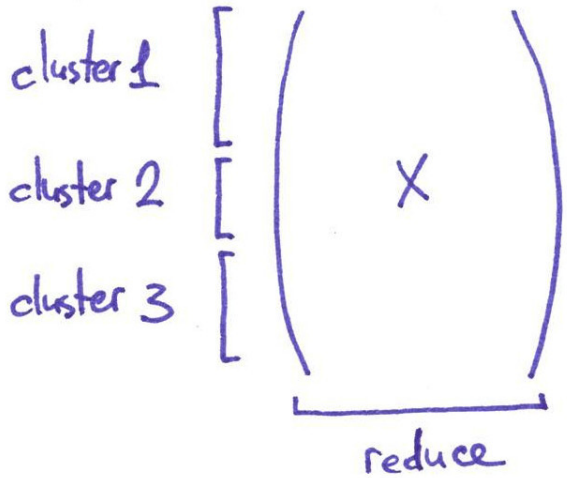
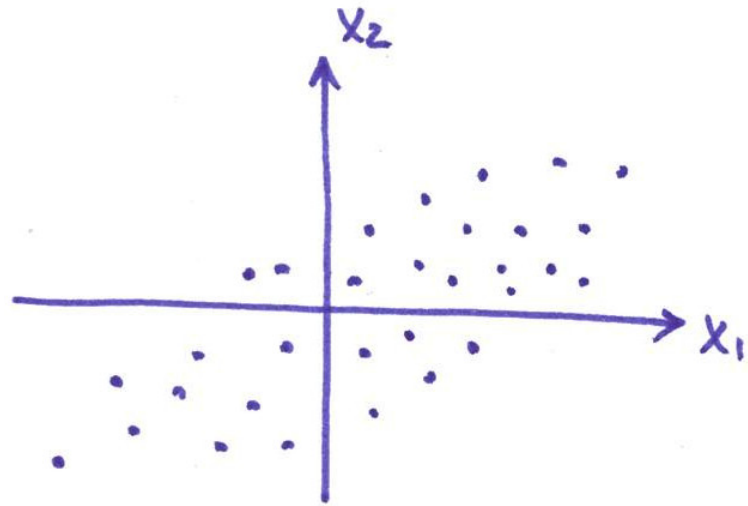
Dimensionality reduction



What for?

- To obtain some insight into the data;
- As a preprocessing step.

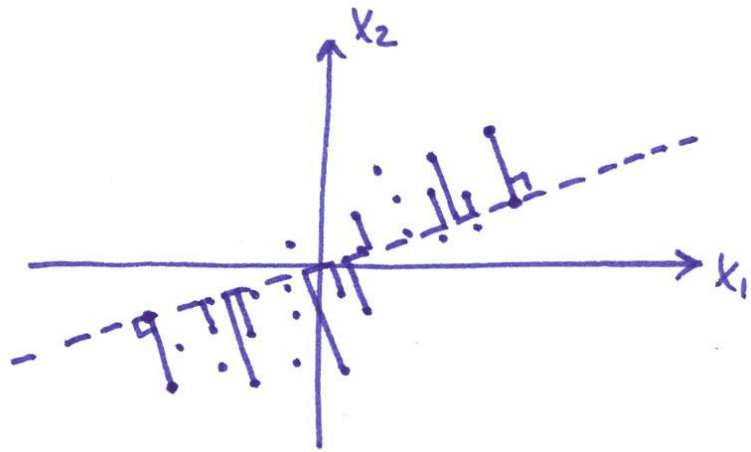
Principal component analysis (PCA)



Linear dimensionality reduction to 1 dimension: turns $\mathbf{X}$ into $\mathbf{Xw}$.

It is enough to consider only unit vectors, $\|\mathbf{w}\| = 1$.

Principal component analysis (PCA)



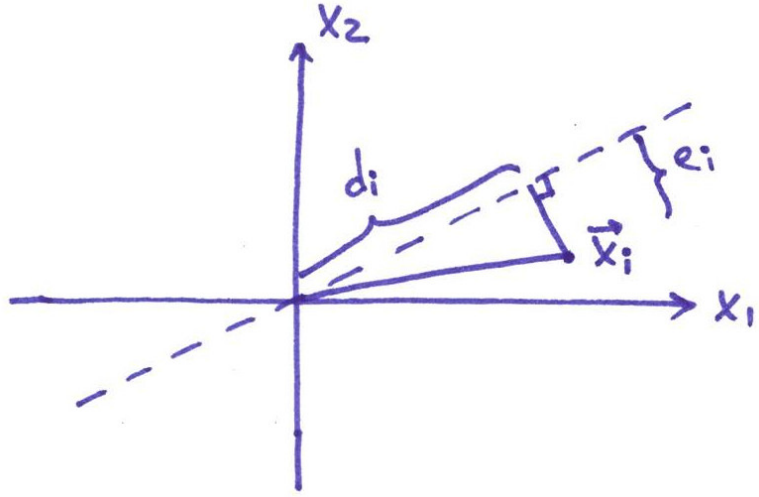
How to choose $\mathbf{w}$?

1. To minimize the reconstruction error.
2. To maximize the variance.

Surprising fact: these are equivalent and PCA does both!

Maximizing variance $\Leftrightarrow$ minimizing error

Assume all features are centered:

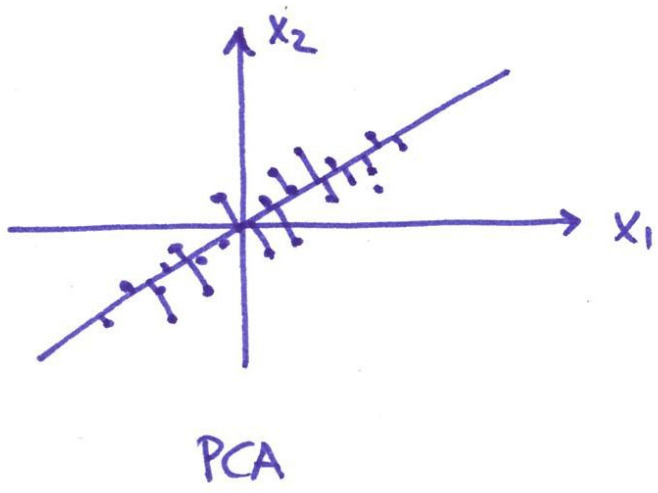


$$d_i^2 + e_i^2 = \|\mathbf{x}_i\|^2$$

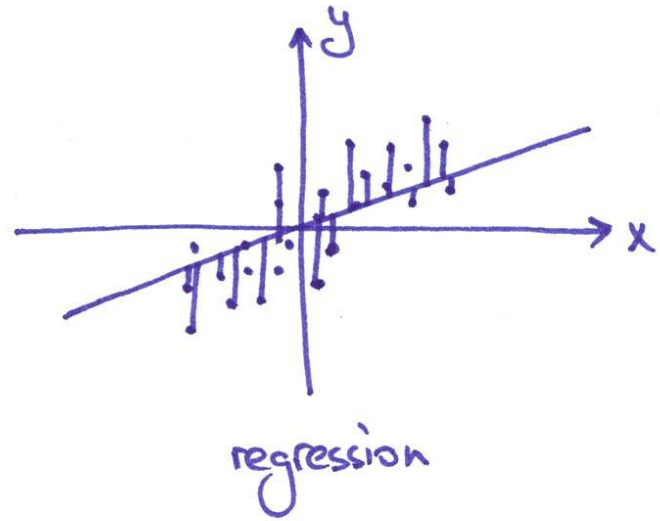
$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 + \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i\|^2 = \text{const}$$

$$\text{Variance} + \text{Mean squared error} = \text{const}$$

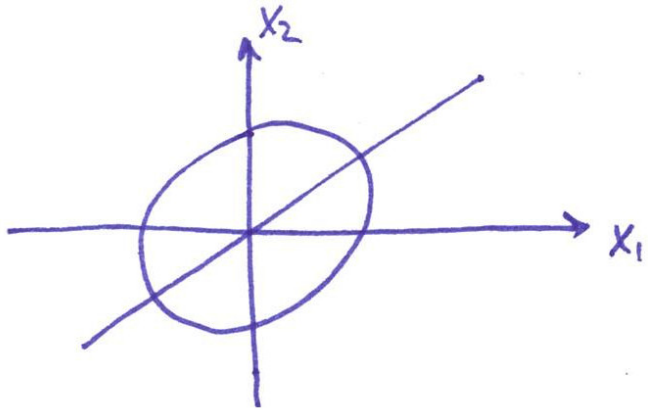
PCA vs. regression

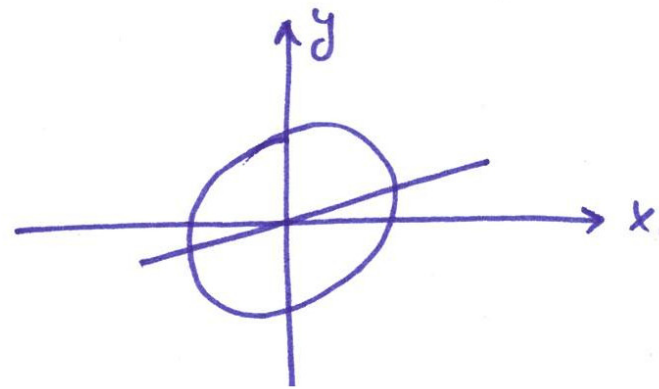


PCA



regression





PCA loss function

Minimizing reconstruction error:

$$\mathcal{L} = \|\mathbf{X} - \mathbf{X}\mathbf{w}\mathbf{w}^\top\|^2.$$

Maximizing variance:

$$-\mathcal{L} = \frac{1}{n}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{w}^\top\mathbf{C}\mathbf{w}, \quad \text{s.t. } \|\mathbf{w}\|^2 = 1.$$

Here $\mathbf{C} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ is the sample covariance matrix.

Maximizing $\mathbf{w}^\top \mathbf{C}\mathbf{w}$

We can use Lagrange multiplier to solve this problem (see Lecture 4):

$$-\mathcal{L} = \mathbf{w}^\top \mathbf{C}\mathbf{w} - \lambda(\mathbf{w}^\top \mathbf{w} - 1).$$

Setting $\partial \mathcal{L}/\partial \mathbf{w} = 0$, we get:

$$\mathbf{C}\mathbf{w} = \lambda \mathbf{w}$$

This means that $\mathbf{w}$ should be an *eigenvector* of $\mathbf{C}$.

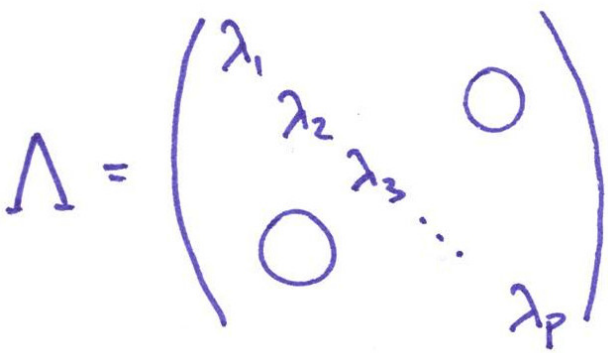
To maximize $\mathbf{w}^\top \mathbf{C}\mathbf{w} = \lambda$, choose the eigenvector with the largest *eigenvalue* λ.

Spectral theorem

$\mathbf{C}$ is a symmetric $p \times p$ matrix. One can prove that it has p eigenvectors that are all orthogonal to each other.

If $\mathbf{w}_1^\top \mathbf{w}_2 = 0$ for eigenvectors $\mathbf{w}_1$ and $\mathbf{w}_2$, then $\mathbf{w}_1^\top \mathbf{C} \mathbf{w}_2 = 0$, i.e. projections on two eigenvectors have correlation zero.

This implies that in the eigenvector basis, the covariance matrix becomes diagonal:

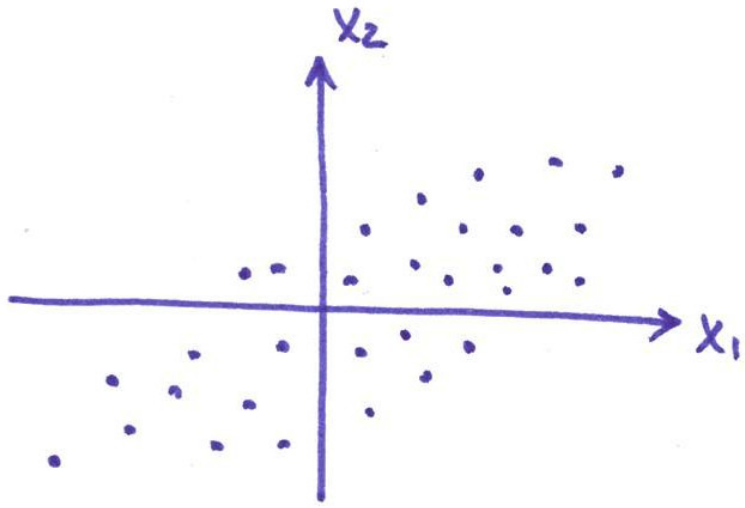


Rotated data: $\mathbf{XV}$.

Covariance:
$\frac{1}{n}\mathbf{V}^\top \mathbf{X}^\top \mathbf{X} \mathbf{V} = \mathbf{V}^\top \mathbf{C} \mathbf{V} = \mathbf{\Lambda}$.

Equivalently: $\mathbf{C} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^\top$.

Max. variance ⇔ min. error ⇔ diag. covariance



Subsequent eigenvectors correspond to the subsequent principal components.

Relationship to SVD

Consider singular value decomposition $\mathbf{X} = \mathbf{USV}^\top$. Then

$$\mathbf{C} = \frac{1}{n}\mathbf{VSU}^\top\mathbf{USV}^\top = \mathbf{V}\frac{\mathbf{S}^2}{n}\mathbf{V}^\top = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top.$$

This is eigendecomposition!

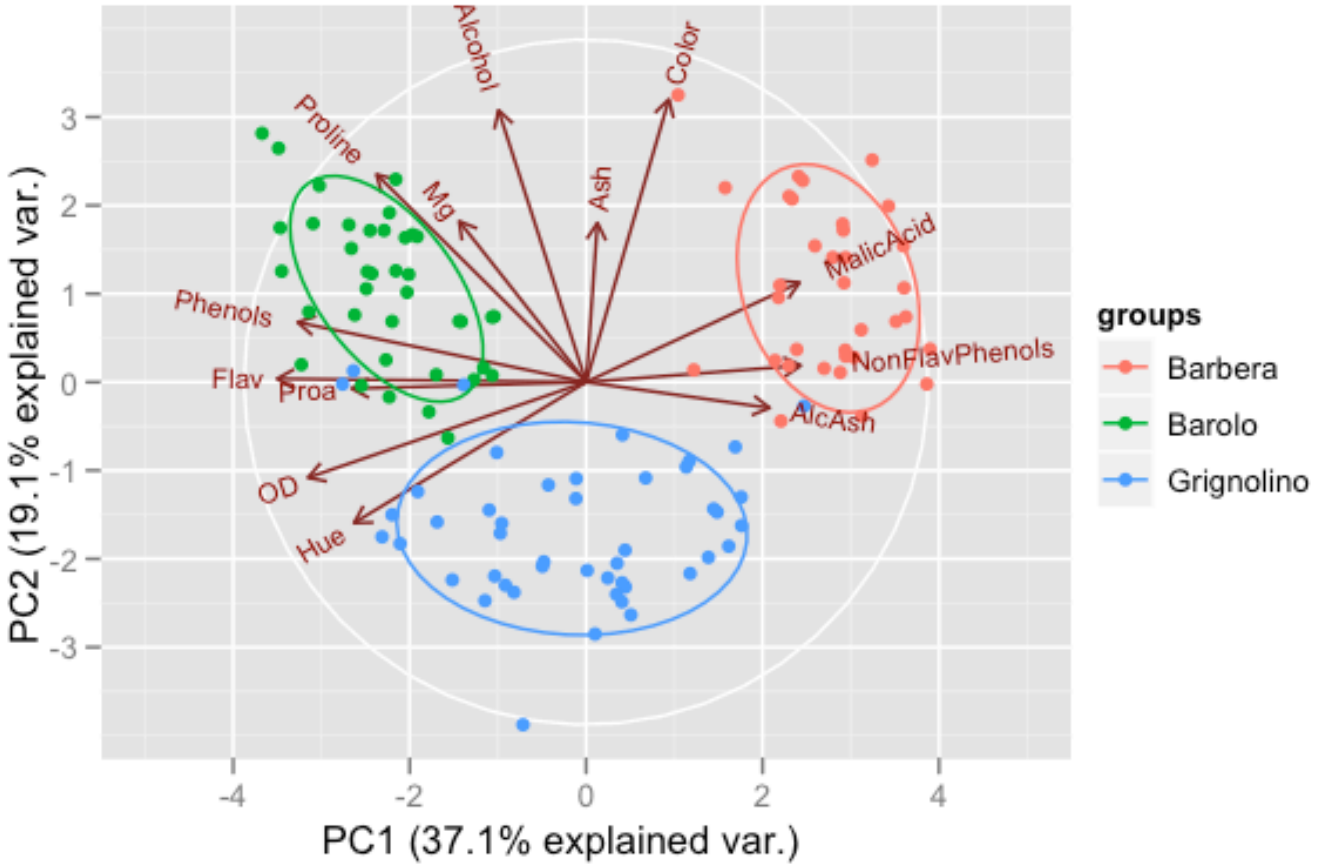
Note: this only holds true if $\mathbf{X}$ is centered.

Why PCA?

One can use PCA for two reasons:

- To explore the data;
- To preprocess the data.

PCA for data exploration



A biplot from https://stats.stackexchange.com/questions/7860

Total variance

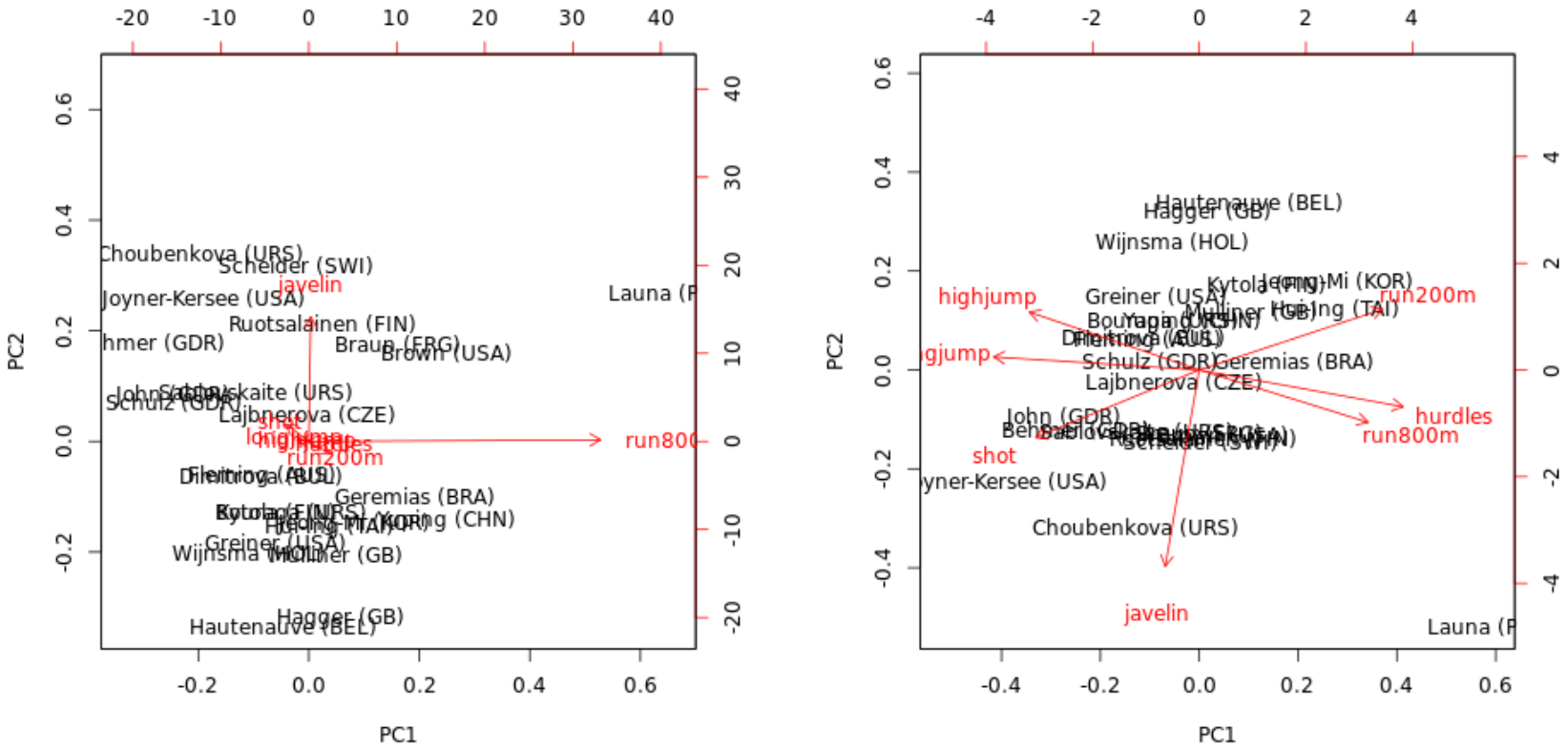
Total variance: $\sum_i \lambda_i = \sum_i \mathrm{Var}[\mathbf{x}_i]$ where λ_i are eigenvalues and $\mathbf{x}_i$ are data features. This is called the *trace* of the covariance matrix:

$$\mathrm{tr}(\mathbf{C}) = \mathrm{tr}(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top) = \mathrm{tr}(\mathbf{V}^\top\mathbf{V}\mathbf{\Lambda}) = \mathrm{tr}(\mathbf{\Lambda}).$$

Explained variance by PC i is defined as $\lambda_i / \mathrm{tr}(\mathbf{C})$.

PCA on correlation or covariance

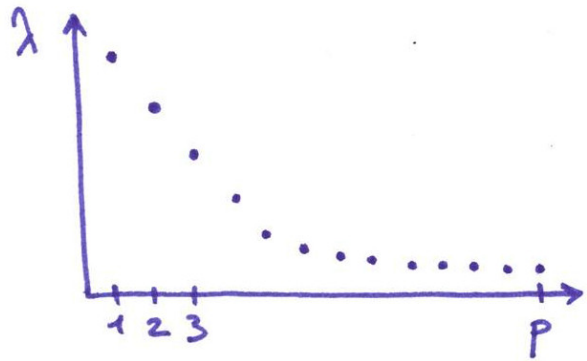
If features are on a different scale, it can make sense to standardize all of them (making $\mathbf{C}$ the correlation matrix):



https://stats.stackexchange.com/questions/53

The spectrum of the covariance matrix

The set of all the eigenvalues $\{\lambda_i\}$ is called the *spectrum*:

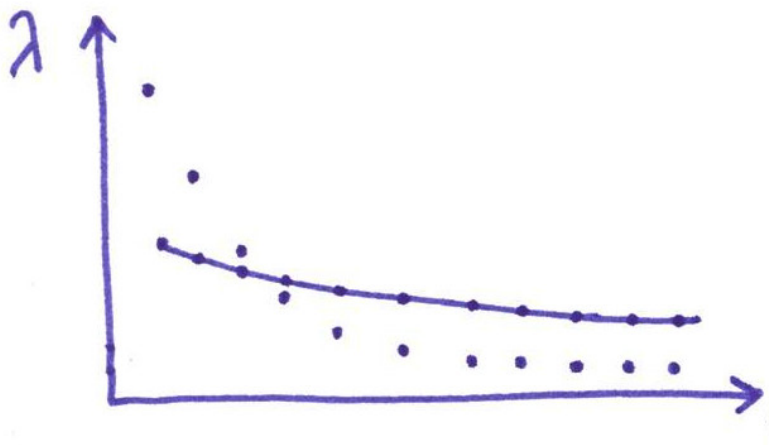


How to choose the number of PCs? There are many rules of thumb: look for an 'elbow'; capture 90% of the total variance; etc.

Better criteria: cross-validation and shuffling the features.

Shuffled spectrum

Shuffle every column of $\mathbf{X}$ independently:



PCA for preprocessing

PCA for preprocessing: reduce $\mathbf{X}$ to a small number k of PCs $\mathbf{XV}_k$ where $\mathbf{V}_k$ is a $p \times k$ matrix of unit-norm eigenvectors with the largest eigenvalues, then use $\mathbf{XV}_k$ for downstream processing.

Advantages: all correlations are zero; no small singular values / eigenvalues left; lower dimensionality; smaller size.

If you use all PCs, you simply rotate the data.

Principal component regression (PCR)

PCA followed by regression is called *principal component regression* (PCR). It is closely related to ridge regression.

Reminder (see Lecture 4):

$$\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{U}\mathbf{U}^\top\mathbf{y}$$

$$\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{ridge}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + n\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{U}\,\text{diag}\left\{\frac{s_i^2}{s_i^2 + n\lambda}\right\}\mathbf{U}^\top\mathbf{y}$$

PCR does hard thresholding of singular values:

$$\text{diag}\{\underbrace{1, 1, \ldots, 1}_{k}, 0, 0, \ldots 0\}.$$

The number of PCs k can serve as a regularization parameter, similar to the ridge penalty λ.

Probabilistic PCA (PPCA)

A different perspective on PCA. Consider a latent variable model:

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}_k, \mathbf{I}_k)$$
$$\mathbf{x} \mid \mathbf{z} \sim \mathcal{N}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

The mean and the covariance of the marginal distribution are:

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu},$$
$$\mathrm{Cov}[\mathbf{x}] = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}.$$

Goal: given a dataset $\mathbf{X}$, fit the model using maximum likelihood.

Solution: EM algorithm.

EM for PPCA

EM algorithm:

- E-step: given $\mathbf{W}, \boldsymbol{\mu}, \sigma^2$, find posterior distribution over $\mathbf{z}$ (it is Gaussian, so it is enough to compute $\mathbb{E}[\mathbf{z}]$ and $\mathrm{Cov}[\mathbf{z}]$).
- M-step: given $\mathbf{z}$, find $\mathbf{W}, \boldsymbol{\mu}, \sigma^2$ maximizing the likelihood.

It turns out that the maximum likelihood solution $\hat{\mathbf{W}}$ is given by $\mathbf{V}_k$ times a particular diagonal matrix. So PPCA is equivalent to PCA!

Factor analysis (FA)

Factor analysis:

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}_k, \mathbf{I}_k)$$
$$\mathbf{x} \mid \mathbf{z} \sim \mathcal{N}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Psi})$$

where $\boldsymbol{\Psi}$ is a diagonal matrix.

FA has been extremely popular in some social sciences. It is a probabilistic latent variable model slightly more general than PPCA.

FA does not have an analytic ML solution. But one can use EM to fit the model.