Factor and Component Analysis

esp. Principal Component Analysis (PCA)

Why Factor or Component Analysis?

- We study phenomena that can not be directly observed
  - ego, personality, intelligence in psychology
  - Underlying factors that govern the observed data

- We want to identify and operate with underlying latent factors rather than the observed data
  - E.g. topics in news articles
  - Transcription factors in genomics

- We want to discover and exploit hidden relationships
  - “beautiful car” and “gorgeous automobile” are closely related
  - So are “driver” and “automobile”
  - But does your search engine know this?
  - Reduces noise and error in results

Why Factor or Component Analysis?

- We have too many observations and dimensions
  - To reason about or obtain insights from
  - To visualize
  - Too much noise in the data
  - Need to “reduce” them to a smaller set of factors
  - Better representation of data without losing much information
  - Can build more effective data analyses on the reduced-dimensional space: classification, clustering, pattern recognition

- Combinations of observed variables may be more effective bases for insights, even if physical meaning is obscure

Factor or Component Analysis

- Discover a new set of factors/dimensions/axes against which to represent, describe or evaluate the data
  - For more effective reasoning, insights, or better visualization
  - Reduce noise in the data
  - Typically a smaller set of factors: dimension reduction
  - Better representation of data without losing much information
  - Can build more effective data analyses on the reduced-dimensional space: classification, clustering, pattern recognition

- Factors are combinations of observed variables
  - May be more effective bases for insights, even if physical meaning is obscure
  - Observed data are described in terms of these factors rather than in terms of original variables/dimensions

Basic Concept

- Areas of variance in data are where items can be best discriminated and key underlying phenomena observed
  - Areas of greatest “signal” in the data
- If two items or dimensions are highly correlated or dependent
  - They are likely to represent highly related phenomena
  - If they tell us about the same underlying variance in the data, combining them to form a single measure is reasonable
    - Parsimony
    - Reduction in Error
- So we want to combine related variables, and focus on uncorrelated or independent ones, especially those along which the observations have high variance
- We want a smaller set of variables that explains most of the variance in the original data, in more compact and insightful form

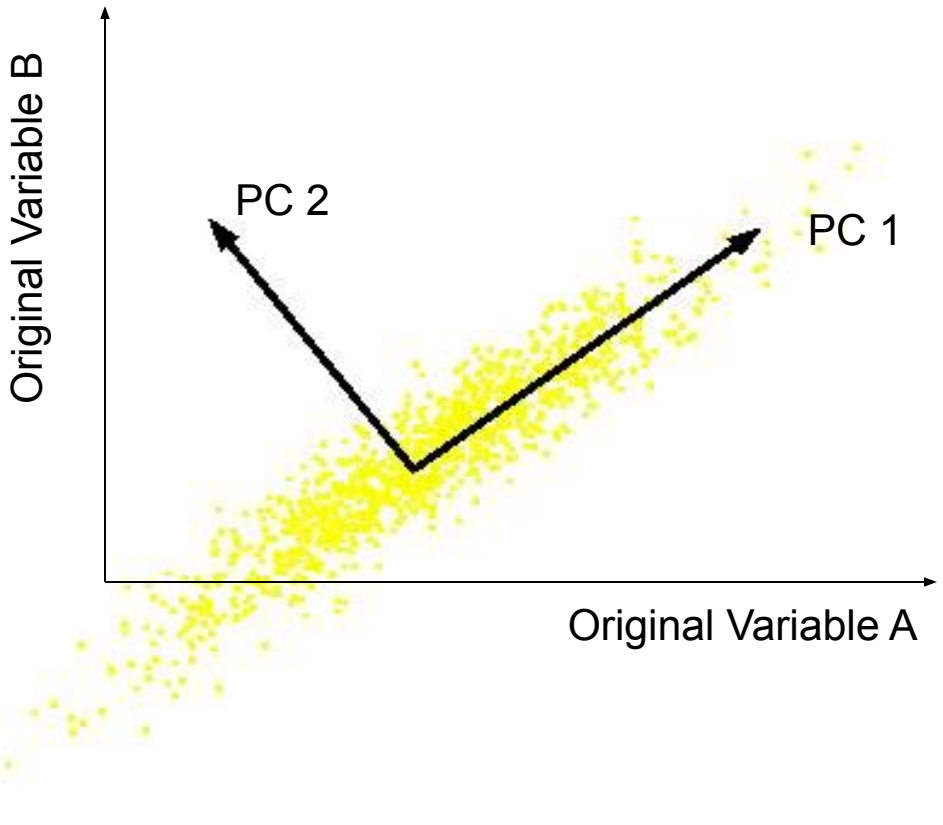
Basic Concept

- What if the dependences and correlations are not so strong or direct?
- And suppose you have 3 variables, or 4, or 5, or 10000?
- Look for the phenomena underlying the observed covariance/co-dependence in a set of variables
  - Once again, phenomena that are uncorrelated or independent, and especially those along which the data show high variance
- These phenomena are called “factors” or “principal components” or “independent components,” depending on the methods used
  - Factor analysis: based on variance/covariance/correlation
  - Independent Component Analysis: based on independence

Principal Component Analysis

- Most common form of factor analysis
- The new variables/dimensions
  - Are linear combinations of the original ones
  - Are uncorrelated with one another
    - Orthogonal in original dimension space
  - Capture as much of the original variance in the data as possible
  - Are called Principal Components

What are the new axes?



- Orthogonal directions of greatest variance in data
- Projections along PC1 discriminate the data most along any one axis

Principal Components

- First principal component is the direction of greatest variability (covariance) in the data
- Second is the next orthogonal (uncorrelated) direction of greatest variability
  - So first remove all the variability along the first component, and then find the next direction of greatest variability
- And so on …

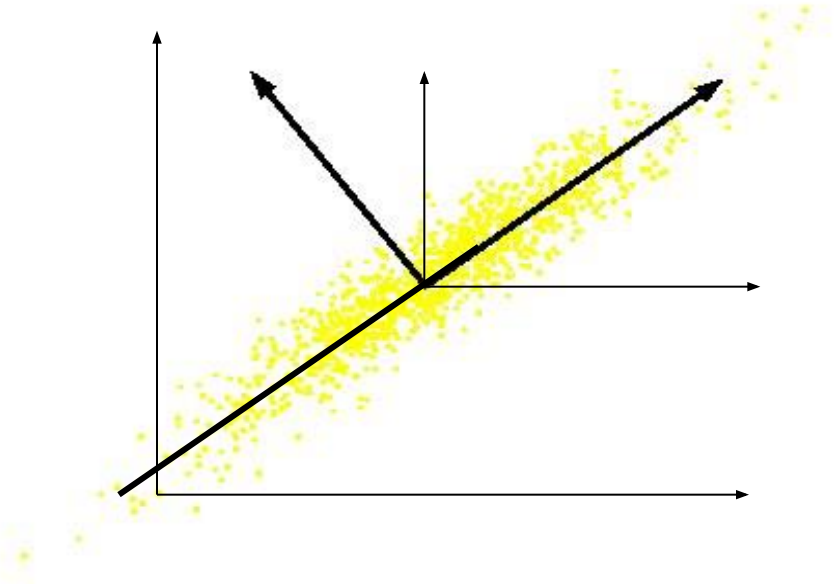
Principal Components Analysis (PCA)

- Principle
  - Linear projection method to reduce the number of parameters
  - Transfer a set of correlated variables into a new set of uncorrelated variables
  - Map the data into a space of lower dimensionality
  - Form of unsupervised learning

- Properties
  - It can be viewed as a rotation of the existing axes to new positions in the space defined by original variables
  - New axes are orthogonal and represent the directions with maximum variability

Computing the Components

- Data points are vectors in a multidimensional space
- Projection of vector **x** onto an axis (dimension) **u** is **u.x**
- Direction of greatest variability is that in which the average square of the projection is greatest
  - I.e. **u** such that $E((\mathbf{u}.\mathbf{x})^2)$ over all **x** is maximized
  - (we subtract the mean along each dimension, and center the original axis system at the centroid of all data points, for simplicity)
  - This direction of **u** is the direction of the first Principal Component



Computing the Components

- $E((\mathbf{u}.\mathbf{x})^2) = E((\mathbf{u}.\mathbf{x})(\mathbf{u}.\mathbf{x})^T) = E(\mathbf{u}.\mathbf{x}.\mathbf{x}^T.\mathbf{u}^T)$
- The matrix $\mathbf{C} = \mathbf{x}.\mathbf{x}^T$ contains the correlations (similarities) of the original axes based on how the data values project onto them
- So we are looking for w that maximizes $\mathbf{u}\mathbf{C}\mathbf{u}^T$, subject to **u** being unit-length
- It is maximized when w is the principal eigenvector of the matrix **C**, in which case
  - $\mathbf{u}\mathbf{C}\mathbf{u}^T = \mathbf{u}\lambda\mathbf{u}^T = \lambda$ if **u** is unit-length, where λ is the principal eigenvalue of the correlation matrix C
  - The eigenvalue denotes the amount of variability captured along that dimension

Why the Eigenvectors?

Maximise $\mathbf{u}^T\mathbf{x}\mathbf{x}^T\mathbf{u}$ s.t $\mathbf{u}^T\mathbf{u} = 1$

Construct Langrangian $\mathbf{u}^T\mathbf{x}\mathbf{x}^T\mathbf{u} - \lambda\mathbf{u}^T\mathbf{u}$

Vector of partial derivatives set to zero

$$\mathbf{x}\mathbf{x}^T\mathbf{u} - \lambda\mathbf{u} = (\mathbf{x}\mathbf{x}^T - \lambda\mathbf{I})\,\mathbf{u} = 0$$

As $\mathbf{u} \neq \mathbf{0}$ then $\mathbf{u}$ must be an eigenvector of $\mathbf{x}\mathbf{x}^T$ with eigenvalue λ

Singular Value Decomposition

The first root is called the prinicipal eigenvalue which has an associated orthonormal ($\mathbf{u}^T\mathbf{u} = 1$) *eigenvector* $\mathbf{u}$

Subsequent roots are ordered such that $\lambda_1 > \lambda_2 > \ldots > \lambda_M$ with rank($\mathbf{D}$) non-zero values.

Eigenvectors form an orthonormal basis i.e. $\mathbf{u}_i^T\mathbf{u}_j = \Delta_{ij}$

The eigenvalue decomposition of $\mathbf{x}\mathbf{x}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$

where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M]$ and $\mathbf{\Sigma} = \text{diag}[\lambda_1, \lambda_2, \ldots, \lambda_M]$

Similarly the eigenvalue decomposition of $\mathbf{x}^T\mathbf{x} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T$

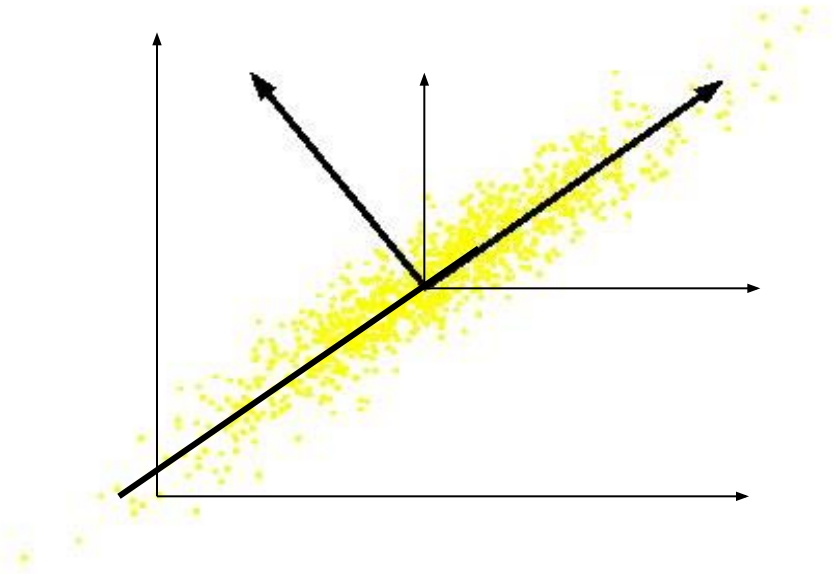
The SVD is closely related to the above $\mathbf{x} = \mathbf{U}\,\mathbf{\Sigma}^{1/2}\,\mathbf{V}^T$

The left eigenvectors **U,** right eigenvectors **V,**

singular values = square root of eigenvalues**.**

Computing the Components

- Similarly for the next axis, etc.
- So, the new axes are the eigenvectors of the matrix of correlations of the original variables, which captures the similarities of the original variables based on how data samples project to them



- Geometrically: centering followed by rotation
  - Linear transformation

PCs, Variance and Least-Squares

- The first PC retains the greatest amount of variation in the sample
- The k^{th} PC retains the k^{th} greatest fraction of the variation in the sample
- The k^{th} largest eigenvalue of the correlation matrix C is the variance in the sample along the k^{th} PC

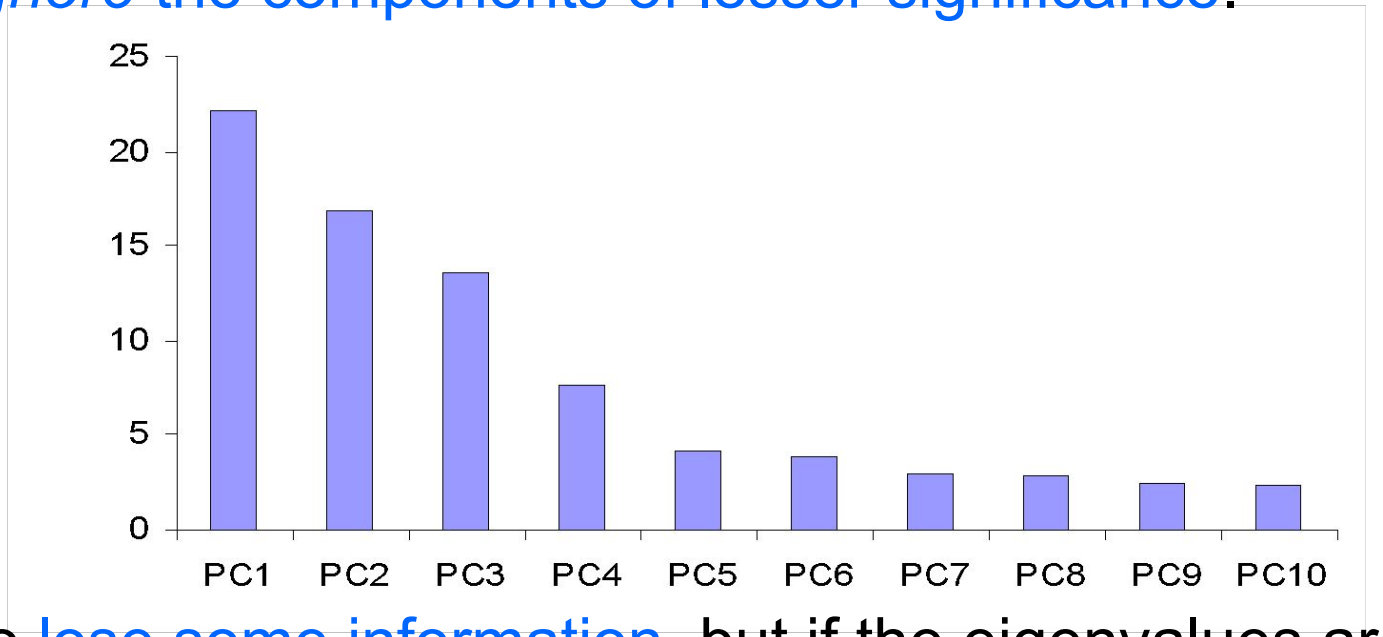
- The least-squares view: PCs are a series of linear least squares fits to a sample, each orthogonal to all previous ones

How Many PCs?

- For n original dimensions, correlation matrix is nxn, and has up to n eigenvectors. So n PCs.
- Where does dimensionality reduction come from?

Dimensionality Reduction

Can *ignore* the components of lesser significance.

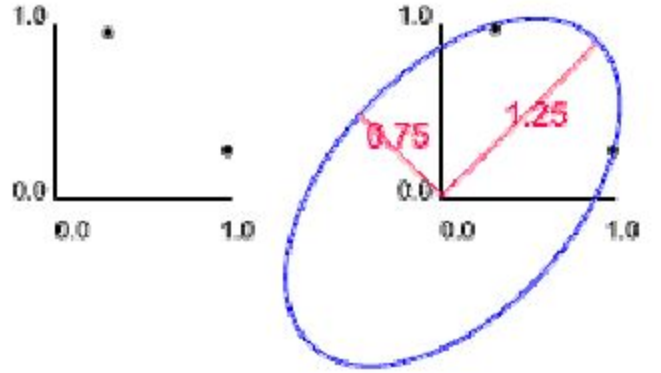


You do lose some information, but if the eigenvalues are small, you don't lose much

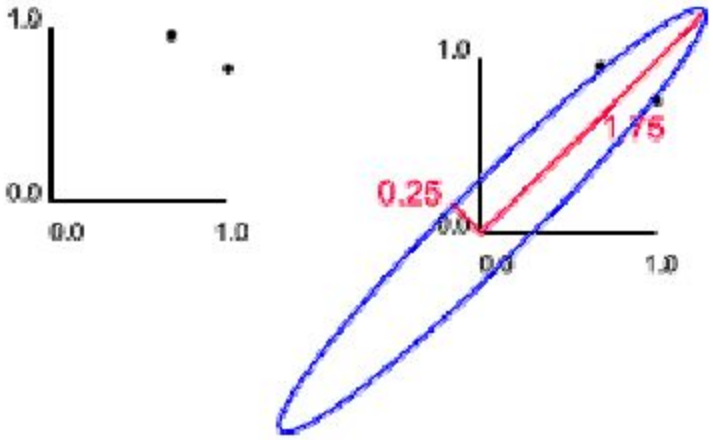
- n dimensions in original data
- calculate n eigenvectors and eigenvalues
- choose only the first p eigenvectors, based on their eigenvalues
- final data set has only p dimensions

Eigenvectors of a Correlation Matrix

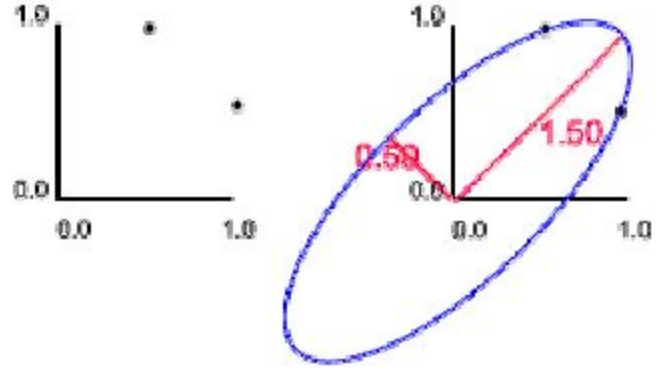
1.00 0.25
0.25 1.00



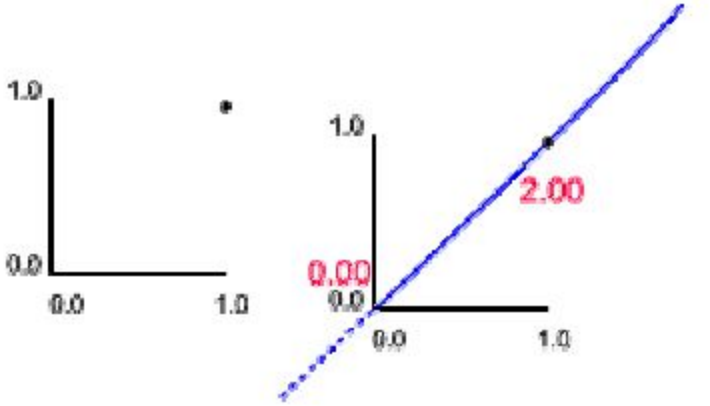
1.00 0.75
0.75 1.00



1.00 0.50
0.50 1.00



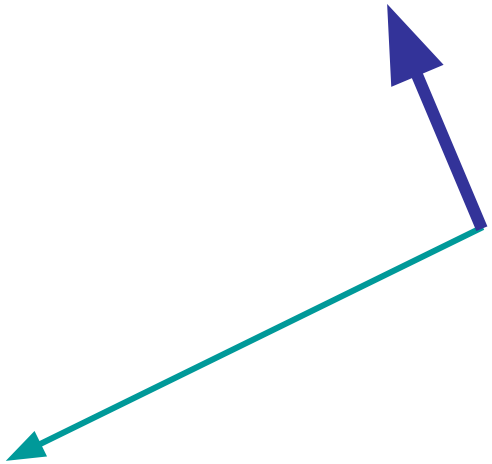
1.00 1.00
1.00 1.00



Limitations of PCA

Are the maximal variance dimensions the relevant dimensions for preservation?

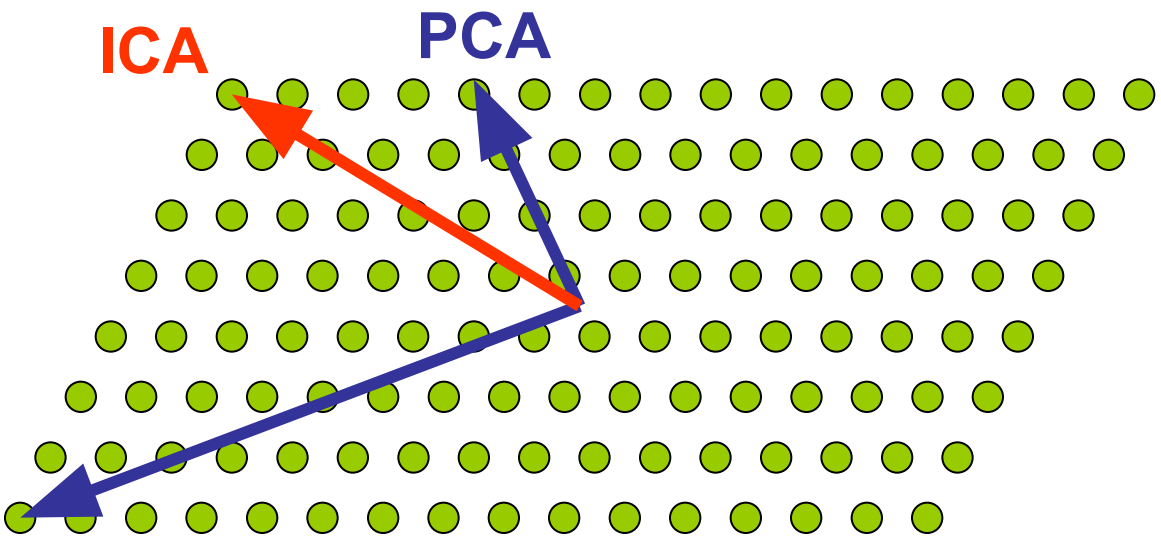
- **Relevant Component Analysis (RCA)**
- **Fisher Discriminant analysis (FDA)**



Limitations of PCA

Should the goal be finding independent rather than pair-wise uncorrelated dimensions

- **Independent Component Analysis (ICA)**



Limitations of PCA

- The reduction of dimensions for complex distributions may need non linear processing
- Curvilenear Component Analysis (CCA)
  - Non linear extension of PCA
  - Preserves the proximity between the points in the input space i.e. local topology of the distribution
  - Enables to unfold some varieties in the input data
  - Keep the local topology