Principal Component Analysis

Instructor: Junghye Lee

Department of Industrial Engineering

junghyelee@unist.ac.kr

Issues on Independent Variables

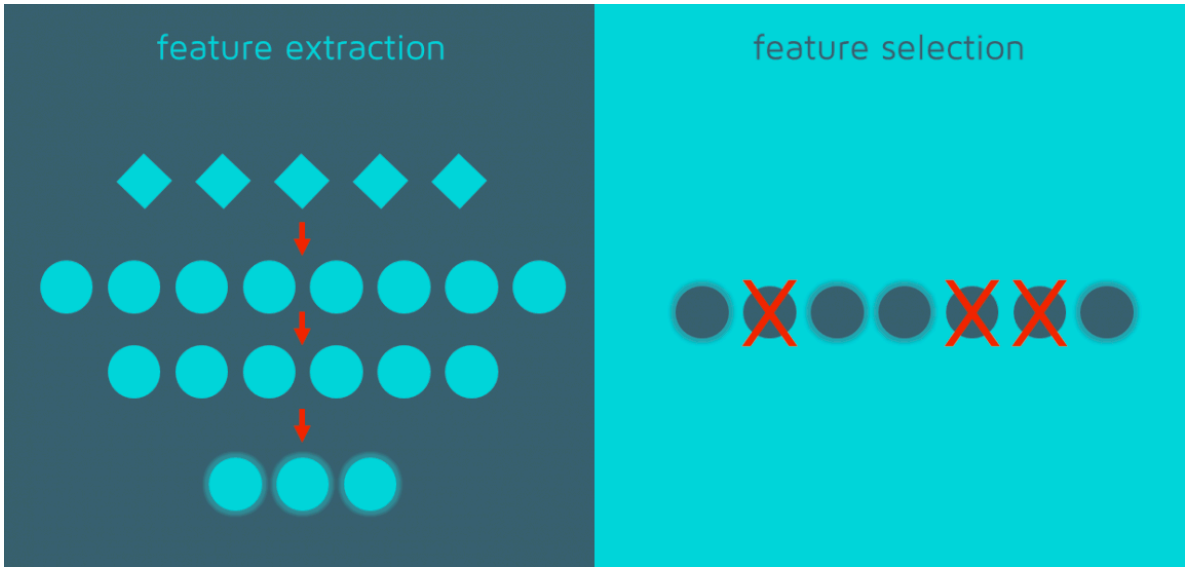
- In addition to the assumptions that we have discussed in the previous class, linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other.
  - e.g. BMI is correlated with Height and Weight

No.	Height	Weight	BMI	Gender	Diseased?
1	167	60	0.36	M	Yes
2	182	44	0.24	F	Yes
3	154	50	0.32	F	Yes
4	188	54	0.29	F	Yes
5	173	48	0.28	F	No
6	175	37	0.21	M	No

What can we do?

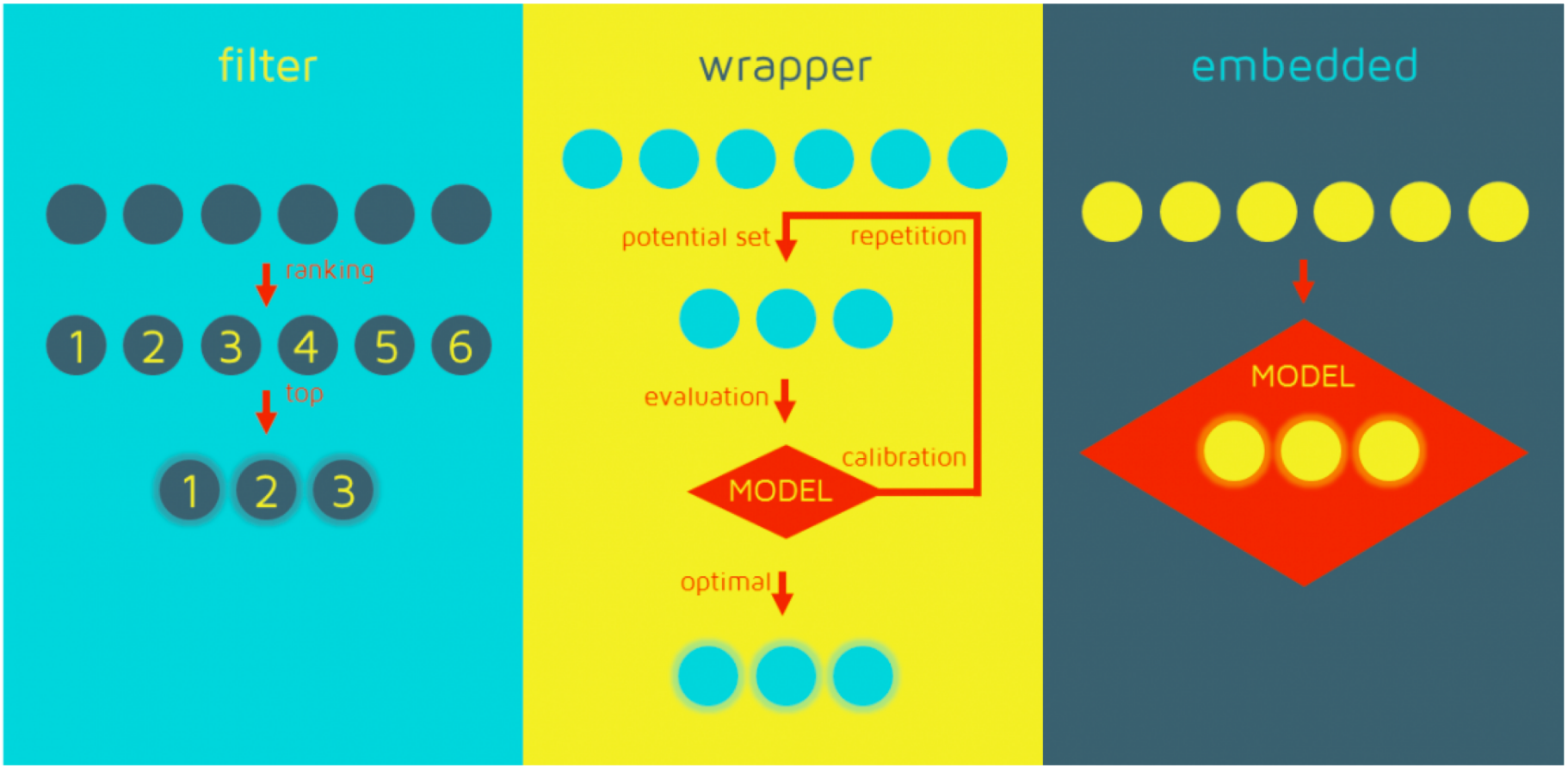
Dimensionality Reduction

- **Feature (variable) selection**
  - Given a set of *p* features, select a subset of features with the size *k (can keep the original meaning of the feature).*
- **Feature (variable) extraction**
  - Determine the appropriate subspace of dimensionality *k* from the original *p*-dimensional space, where $k \leq p$ (*has to lose the original meaning of the feature*).



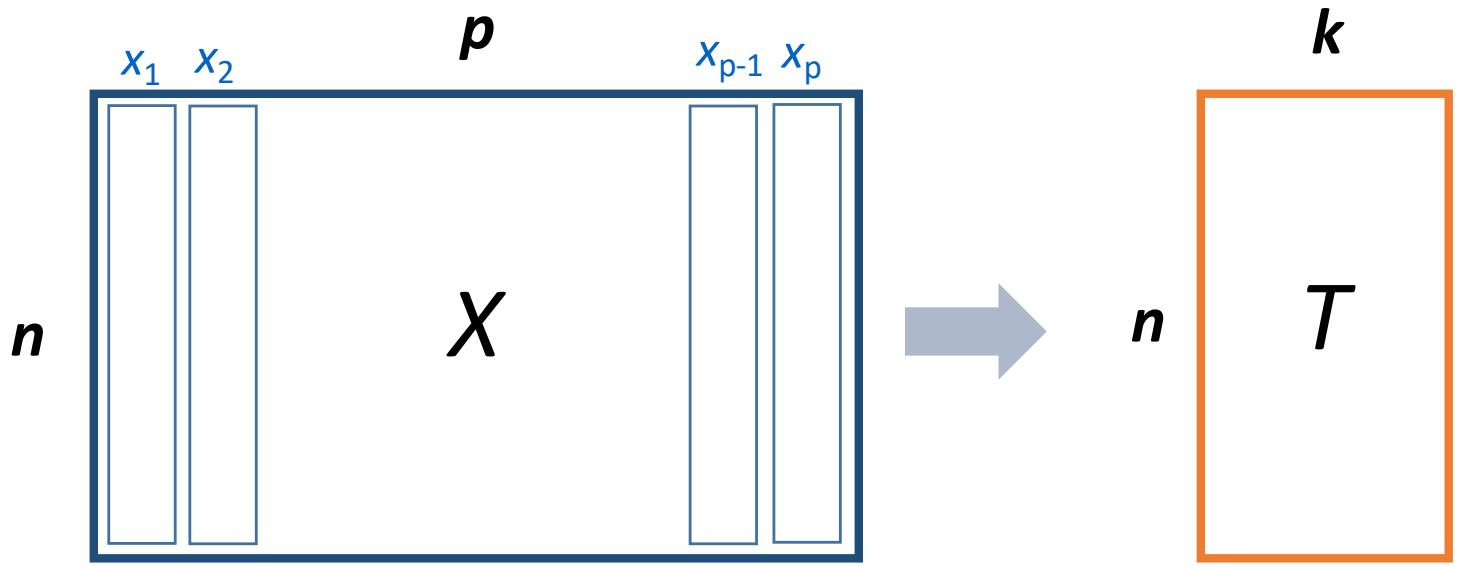
Feature Selection (Reference; optional)

- Three different categories
  - depending on how to stick the feature selection on the regressor/classifier



Feature Extraction

- **Summarization of data** with many (*p*) variables by a smaller set of (*k*) derived (synthetic, composite) variables.



- Balancing act between
  - Clarity of representation and ease of understanding
  - Oversimplification: loss of important or relevant information

Issues on Independent Variables

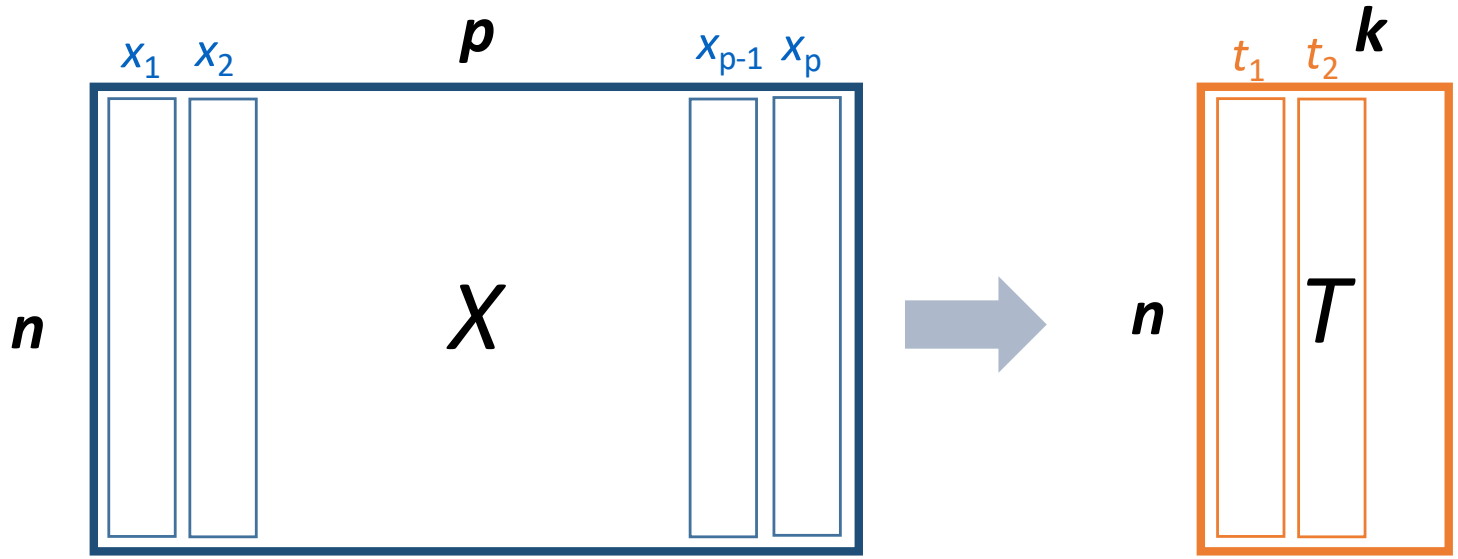
- **Independent variables can be correlated each other.** Then, linear regression works well?
  - To remove the collinearity (for the independent assumption)
- **What if we would like to derive latent variables?**
  - To make latent variables (for interpretability)
- **What if independent variables have noises?**
  - To reduce the noise (for the analysis accuracy)

Principal Component Analysis (PCA)

- **Principal Component Analysis (PCA) is the most well-known and widely used technique to resolve the multicollinearity issue.**
- PCA takes a data matrix of *n* objects by *p* variables, which may be correlated, and summarizes it by **uncorrelated axes** (principal components or principal axes) that are **linear combinations of the original *p* variables**.
- The first *k* components display **as much as possible of the variation** among objects.

PCA

- **Summarization of data** with many (*p*) variables by a smaller set of (*k*) derived (synthetic, composite) variables **to keep the the variance of the original data as maximum.**



- **Then how? By using the linear combination idea!**

PCA

From *k* original variables: $x_1, x_2, \ldots, x_p \in \mathbb{R}^n$ (given):

Produce k new variables: $t_1, t_2, \ldots, t_k \in \mathbb{R}^n$ (you should find):

$$t_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p$$

$$t_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p$$

$$\ldots$$

$$t_k = a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kp}x_p$$

Q1: How many linear combinations here?

Q2: What is the dimension of $\boldsymbol{X}$ and $\boldsymbol{T}$?

Q3: How can you represent these equations in a matrix format?

PCA

- **What is the purpose of this extraction?**
  - To remove the correlation among original variables
  - So, the final $t_1, t_2, \ldots, t_k$ should be independent to each other.

- **To make this happen together, then the parameters should be**

$\{a_{11}, a_{12}, \ldots, a_{1p}\}$ is 1st **Eigenvector** of correlation/covariance matrix, and **coefficients** of first principal component

$\{a_{21}, a_{22}, \ldots, a_{2p}\}$ is 2nd **Eigenvector** of correlation/covariance matrix, and **coefficients** of 2nd principal component

$\{a_{k1}, a_{k2}, \ldots, a_{kp}\}$ is *k*th **Eigenvector** of correlation/covariance matrix, and **coefficients** of *k*th principal component

Variance-Covariance Matrix

- Given two random variables X_j, X_j,

$$\mathrm{Var}(X_j) = \mathrm{Cov}(X_j, X_j) = \mathrm{E}\left[(X_j - \mathrm{E}[X_j])(X_j - \mathrm{E}[X_j])^T\right]$$

  - Degree of the spread of the variable

$$\mathrm{Cov}(X_j, X_l) = \mathrm{E}\left[(X_j - \mathrm{E}[X_j])(X_l - \mathrm{E}[X_l])^T\right]$$

  - Degree to which the variables are linearly correlated

- When there are *p* independent variables $\{X_j\}_{j=1}^{p}$,
  - Covariance variance matrix

$$K = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & \cdots & \mathrm{Var}(X_p) \end{bmatrix}$$

- We need estimators to estimate these from data.

Variance-Covariance Matrix Estimator

- Objects are represented as a cloud of *n* points in a multi-dimensional space with an axis for each of the *p* variables.

$$V_j = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{ij} - \bar{X}_j\right)^2$$

Variance of variable j

Centroid of the points of j

$$C_{jl} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{ij} - \bar{X}_j\right)\left(X_{il} - \bar{X}_l\right)$$

Covariance of variables j and l

Sum over all n objects

Value of variable j in object i

Mean of variable j

Value of variable l in object i

Mean of variable l

Variance-Covariance Matrix Estimate

- $\widehat{K} = S_X = \frac{1}{n-1}(\boldsymbol{X} - \overline{\boldsymbol{x}})^{\boldsymbol{T}}(\boldsymbol{X} - \overline{\boldsymbol{x}})$ including
  - $V_j = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{\boldsymbol{x}}_j\right)^2$
  - $C_{jl} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{\boldsymbol{x}}_j\right)(x_{il} - \overline{\boldsymbol{x}}_l)$

- PCA aims to derive uncorrelated features having the variance of the original data as maximum.
  - What is the variance of the original data?

Variance-Covariance Matrix Estimate

- $\widehat{K} = S_X = \frac{1}{n-1}(\boldsymbol{X} - \overline{\boldsymbol{x}})^{\boldsymbol{T}}(\boldsymbol{X} - \overline{\boldsymbol{x}})$ including
  - $V_j = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{\boldsymbol{x}}_j\right)^2$
  - $C_{jl} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{\boldsymbol{x}}_j\right)(x_{il} - \overline{\boldsymbol{x}}_l)$

- PCA aims to derive uncorrelated features having the variance of the original data as maximum.
  - What is the variance of the original data? $\mathrm{tr}(S)$
  - Is this fixed?

Variance-Covariance Matrix Estimate

- $\widehat{K} = S_X = \frac{1}{n-1}(\boldsymbol{X} - \bar{\boldsymbol{x}})^{\boldsymbol{T}}(\boldsymbol{X} - \bar{\boldsymbol{x}})$ including
  - $V_j = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{\boldsymbol{x}}_j\right)^2$
  - $C_{jl} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{\boldsymbol{x}}_j\right)(x_{il} - \bar{\boldsymbol{x}}_l)$

- PCA aims to derive uncorrelated features having the variance of the original data as maximum.
  - What is the variance of the original data? $\mathrm{tr}(S)$
  - Is this fixed?
  - $S_T = \frac{1}{n-1}(\boldsymbol{T} - \bar{\boldsymbol{t}})^{\boldsymbol{T}}(\boldsymbol{T} - \bar{\boldsymbol{t}})$
  - Then $\mathrm{tr}(S_T) = \mathrm{tr}(S_X)$ when $k = p$

Variance-Covariance Matrix Estimate

- $\widehat{K} = S_X = \frac{1}{n-1}(\boldsymbol{X} - \bar{\boldsymbol{x}})^{\boldsymbol{T}}(\boldsymbol{X} - \bar{\boldsymbol{x}})$ including
  - $V_j = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{\boldsymbol{x}}_j\right)^2$
  - $C_{jl} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{\boldsymbol{x}}_j\right)(x_{il} - \bar{\boldsymbol{x}}_l)$

- PCA aims to derive uncorrelated features having the variance of the original data as maximum.
  - What is the variance of the original data? $\mathrm{tr}(S)$
  - Is this fixed? Yes
  - $S_T = \frac{1}{n-1}(\boldsymbol{T} - \bar{\boldsymbol{t}})^{\boldsymbol{T}}(\boldsymbol{T} - \bar{\boldsymbol{t}})$
  - Then $\mathrm{tr}(S_T) = \mathrm{tr}(S_X)$ when $k = p$

Advanced

Maximize $\mathrm{tr}(S_T) = \mathrm{tr}\left(\frac{1}{n-1}(\boldsymbol{T} - \bar{\boldsymbol{t}})^{\boldsymbol{T}}(\boldsymbol{T} - \bar{\boldsymbol{t}})\right) = \mathrm{tr}(\frac{1}{n-1}(\boldsymbol{XA} - \bar{\boldsymbol{t}})^{\boldsymbol{T}}(\boldsymbol{XA} - \bar{\boldsymbol{t}}))$

s.t. $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I} \quad (k \leq p)$

Objective Function

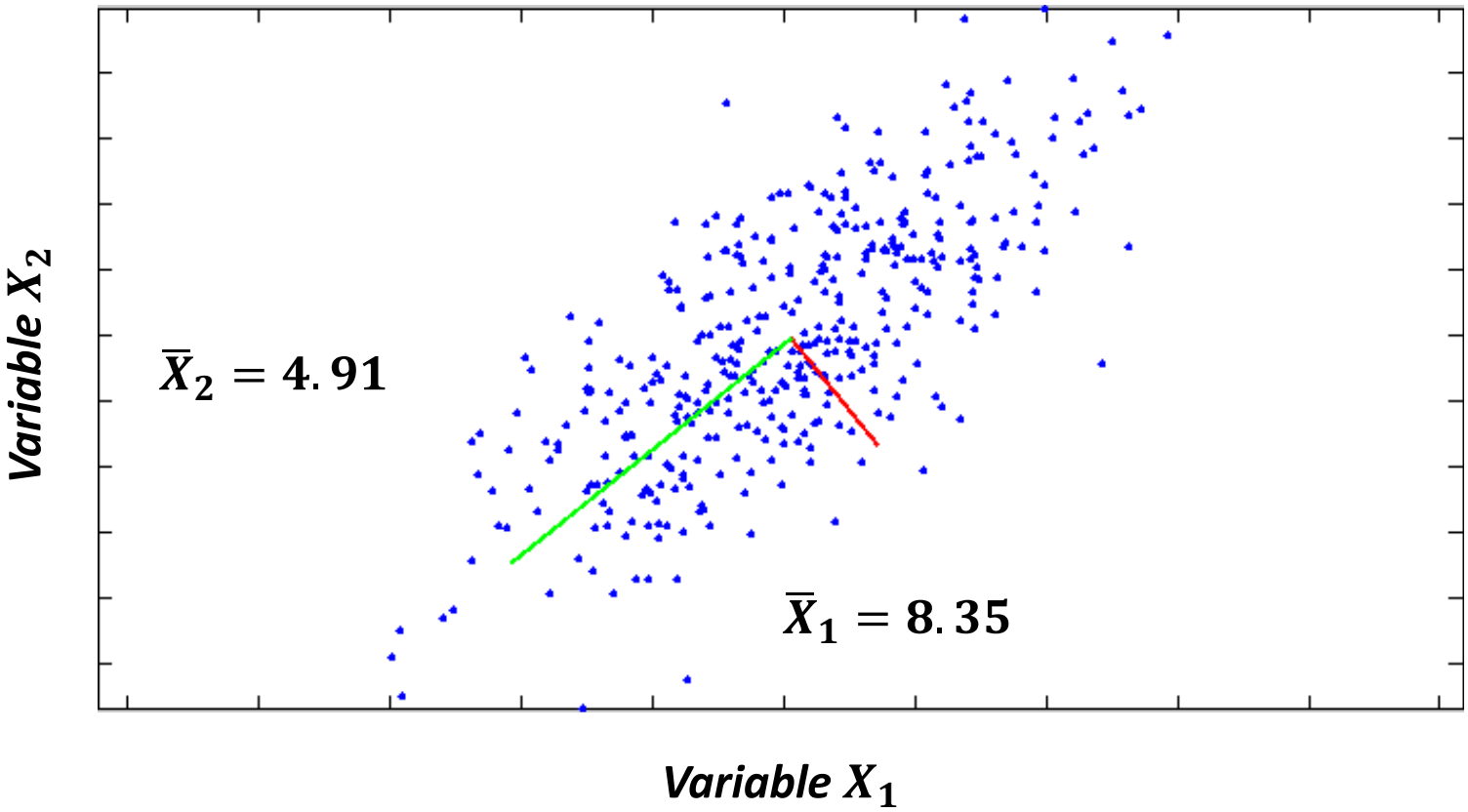
Geometric Rationale of PCA

- Objective of PCA is to **convert the *p*-dimensional space to new positions (principal axes)** that have the following properties:
  - Covariance among each pair of the principal axes is zero (the principal axes are uncorrelated).
  - Ordered such that principal axis 1 has the highest variance, axis 2 has the next highest variance, ... , and axis *p* has the lowest variance

- How come? Rigidly **rotate the axes!**

2D Example of PCA

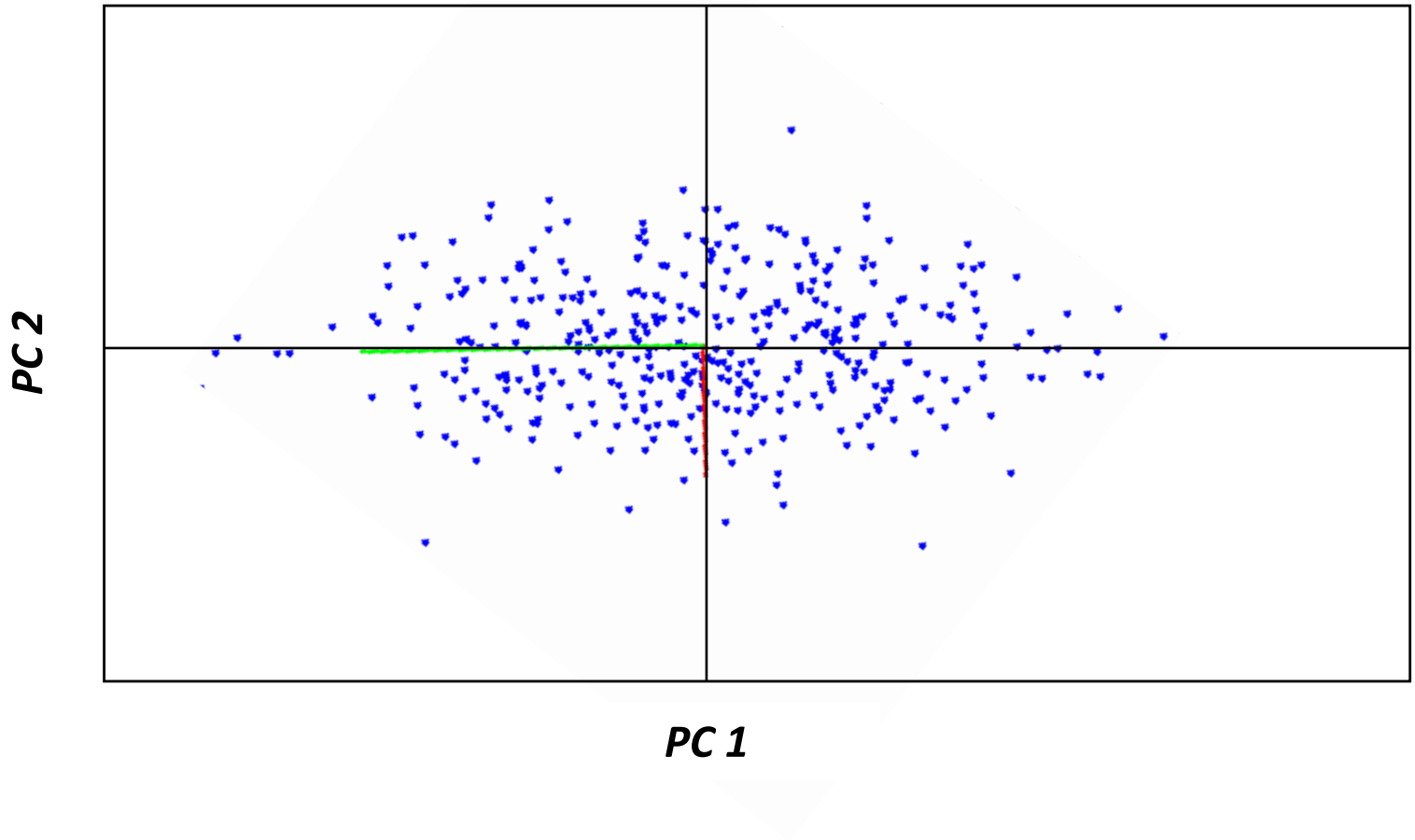
- Variables X_1 and X_2 have positive covariance and each has a similar variance.



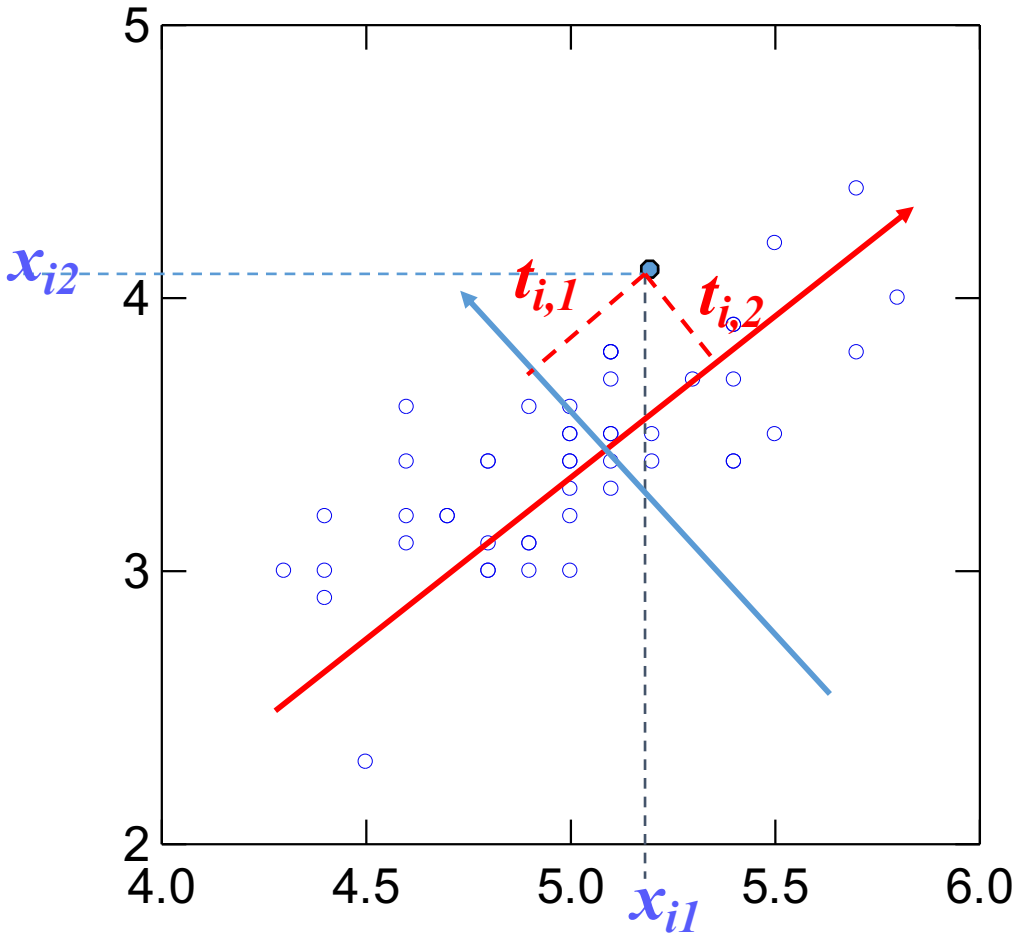
$V_1 = 6.67$ $V_2 = 4.94$ $C_{1,2} = 3.42$

Principal Components are Computed

- PC 1 has the highest possible variance ($V_1 = 9.88$)
- PC 2 has a variance of $V_2 = 2.03$
- PC 1 and PC 2 have zero covariance.



Comparison

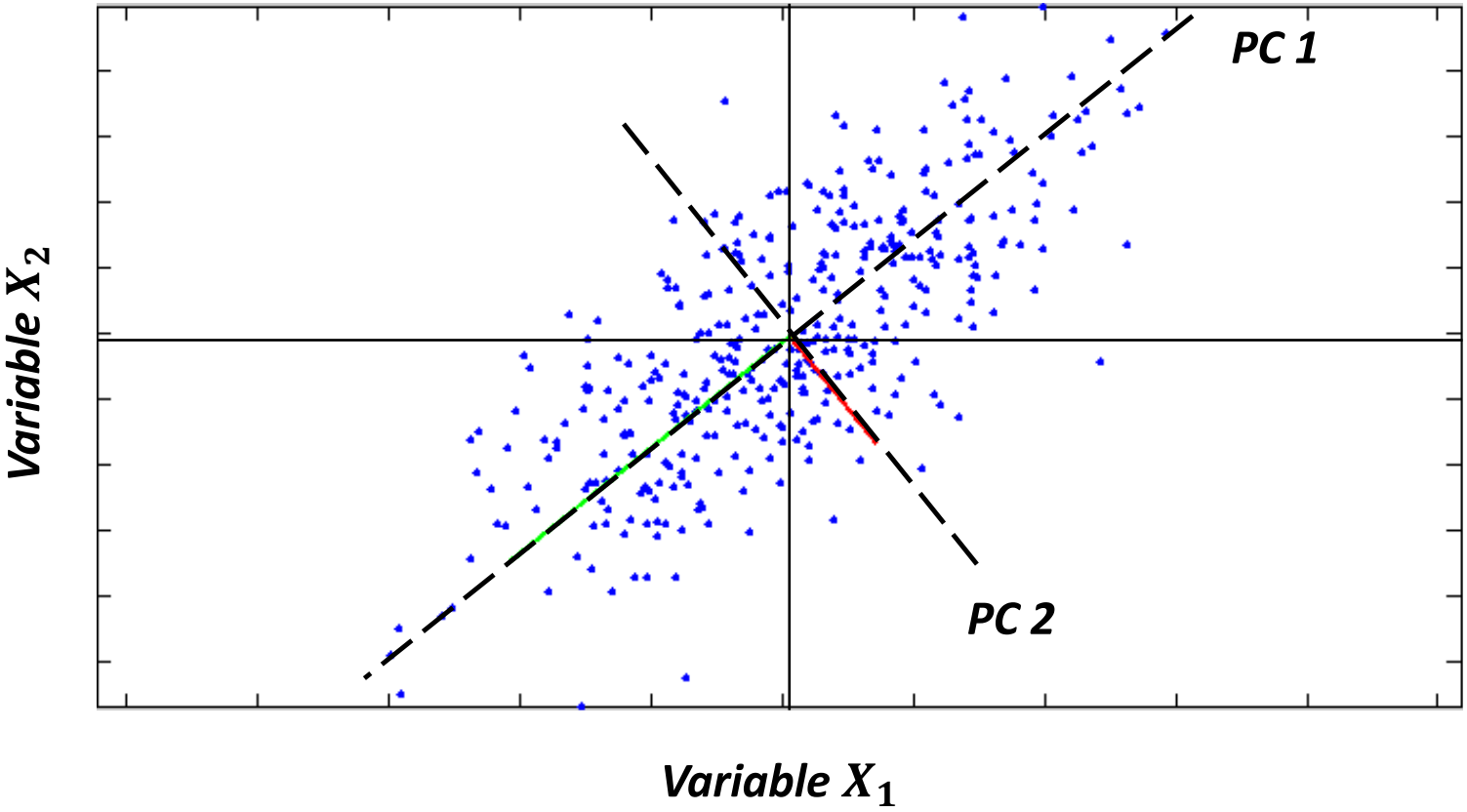


PCA

- Each factor is a linear combination of the original two variables.

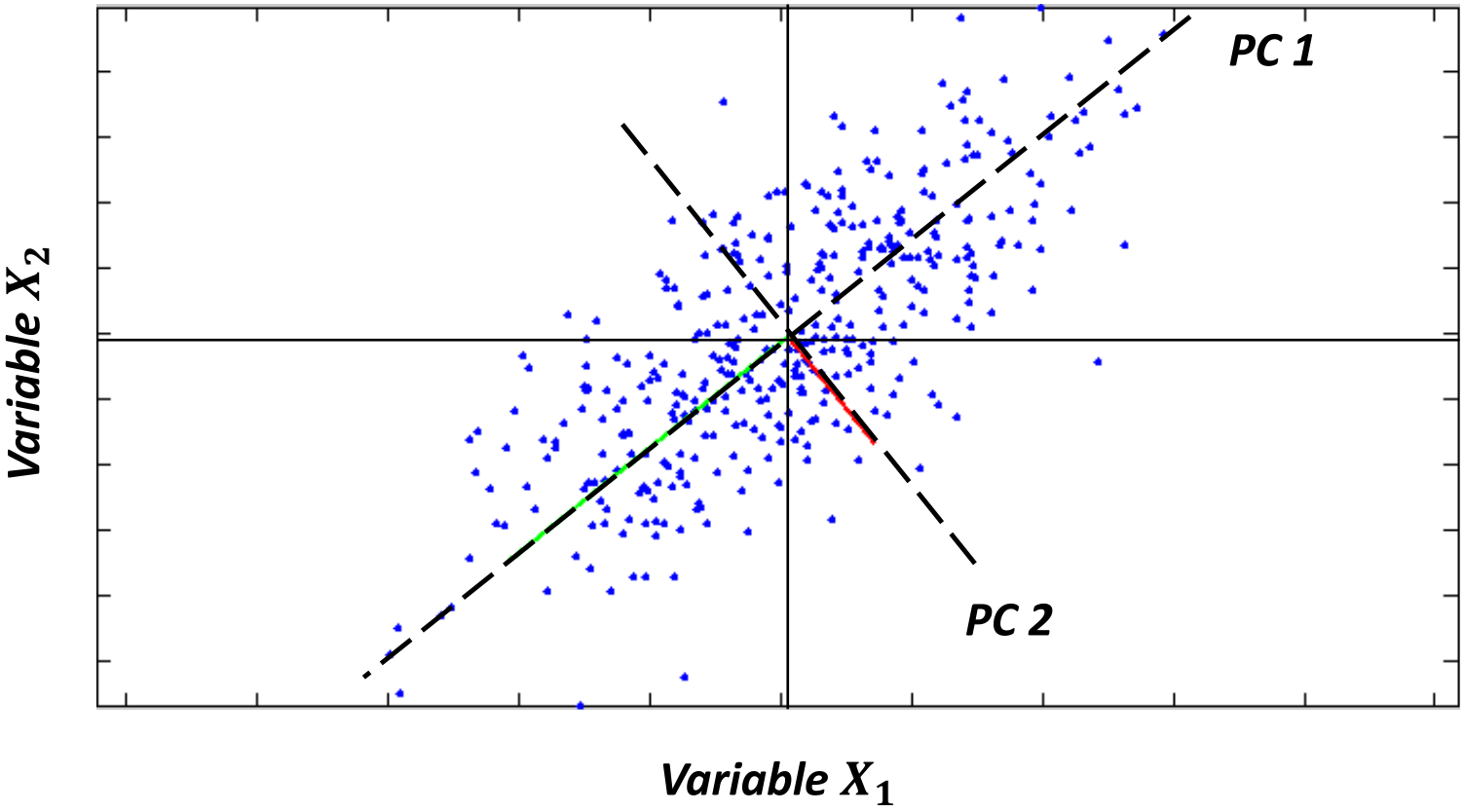
$$T_i = a_{i1}X_1 + a_{i2}X_2$$

  - a_{ij}'s are the coefficients for factor i multiplied by the measured value for variable j



PCA

- PC axes are a rigid rotation of the original variables.
- PC 1 is simultaneously the direction of **maximum variance and a least-squares "line of best fit"** (squared distances of points away from PC 1 are minimized).



Generalization to *p*-dimensions

- In practice nobody uses PCA with only 2 variables.
- The algebra for finding principal axes readily generalizes to *p* variables.
- PC 1 is the direction of **maximum variance** in the *p*-dimensional cloud of points.
- PC 2 is in the direction of the **next highest variance**, subject to the **constraint that it has zero covariance with PC 1**.
- PC 3 is in the direction of the **next highest variance**, subject to the **constraint that it has zero covariance with both PC 1 and PC 2** and so on... up to PC *p.*

Generalization to *p*-dimensions

- If we take the first *k* principal components, they define the *k*-dimensional "hyperplane of best fit" to the point cloud of the total variance of all *p* variables:
  - PCs 1 to *k* represent the maximum possible proportion of that variance that can be displayed in *k* dimensions.
  - i.e. the squared Euclidean distances among points calculated from their coordinates on PCs 1 to *k* are the best possible representation of their squared Euclidean distances in the full *p* dimensions.

$$T_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p$$

where $i = 1, \ldots, k$.

The Algebra of PCA

- First step is to calculate the cross-products matrix of variances and covariances among every pair of the *p* variables
- Square, symmetric matrix
- Diagonals are the variances, off-diagonals are the covariances.

	X_1	X_2
X_1	6.6707	3.4170
X_2	3.4170	6.2384

Variance-covariance Matrix

The Algebra of PCA

- In matrix notation, this is computed as

$$S = X^T X$$

  - where X is the *n* x *p* data matrix, with each variable.

	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$
$\boldsymbol{X_1}$	6.6707	3.4170
$\boldsymbol{X_2}$	3.4170	6.2384

Variance-covariance Matrix

The Algebra of PCA

- Sum of the diagonals of the variance-covariance matrix is called the trace.
- It represents the total variance in the data.
- It is the mean squared Euclidean distance between each object and the centroid in *p*-dimensional space.

	X_1	X_2
X_1	6.6707	3.4170
X_2	3.4170	6.2384

Trace = 12.9091

The Algebra of PCA

- Finding the principal axes involves **eigen analysis** of the cross-products matrix (S)
- The eigenvalues (latent roots) of S are solutions (λ) to the characteristic equation:

$$S\mathrm{v} = \lambda\mathrm{v}$$
$$|S - \lambda I| = 0$$

  - where $|\cdot|$: determinant of a matrix

The Algebra of PCA

- The eigenvalues, $\lambda_1, \lambda_2, \ldots \lambda_p$ are the **variances of the coordinates on each principal component axis**.
- The sum of all *p* eigenvalues equals the trace of *S* (the sum of the variances of the original variables).

	X_1	X_2
X_1	6.6707	3.4170
X_2	3.4170	6.2384

$$\lambda_1 = 9.8783$$
$$\lambda_2 = 3.0308$$

Trace = 12.9091

Note: $\boldsymbol{\lambda_1 + \lambda_2 = 12.9091}$

The Algebra of PCA

- Each eigenvector consists of *p* values which represent the "contribution" of each variable to the principal component axis
- Eigenvectors are uncorrelated (orthogonal)
  - Their inner product is zero.

Eigenvectors

	u_1	u_2
X_1	0.7291	-0.6844
X_2	0.6844	0.7291

0.7291*(-0.6844) + 0.6844*0.7291 = 0

* Eigenvectors of $\frac{1}{n-1}(\boldsymbol{X}-\bar{\boldsymbol{x}})^{\boldsymbol{T}}(\boldsymbol{X}-\bar{\boldsymbol{x}})$ = Eigenvectors of $(\boldsymbol{X}-\bar{\boldsymbol{x}})^{\boldsymbol{T}}(\boldsymbol{X}-\bar{\boldsymbol{x}})$
* Eigen decomposition provides you orthonormal eigenvectors.
* Of course, eigenvalues differ.

The Algebra of PCA

- Coordinates of each object i on the *k*th principal axis, known as the scores on PC *k*, are computed as

$$t_{ki} = a_{1k}x_{1i} + a_{2k}x_{2i} + \cdots + a_{pk}x_{pi}$$

  - where T is the *n* x *k* matrix of PC scores, X is the *n* x *p* centered data matrix and A is the *p* x *k* matrix of eigenvectors.

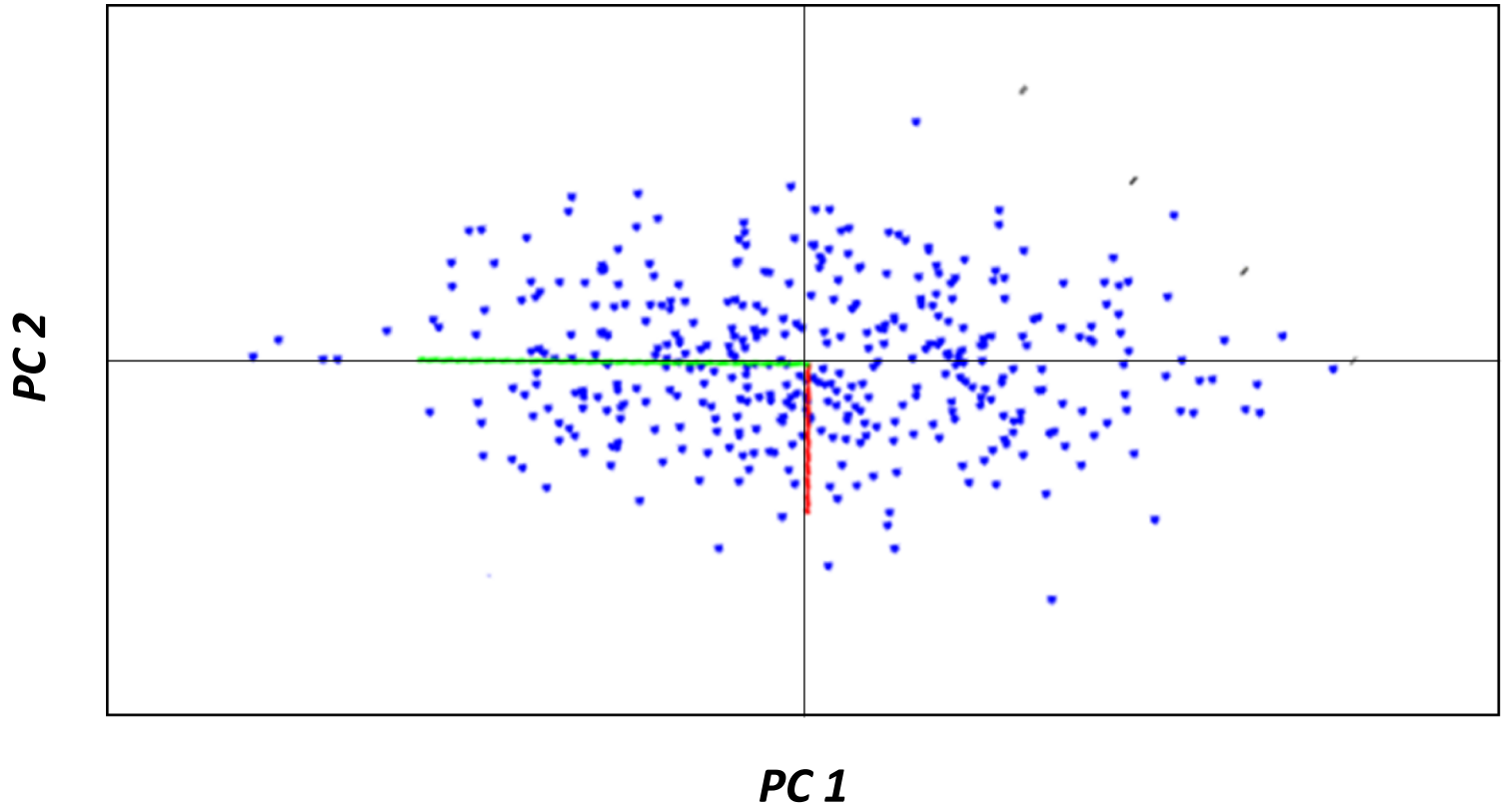
What about t_{kj} (for the object *j*)?

The Algebra of PCA

- Variance of the scores on each PC axis is equal to the corresponding eigenvalue for that axis.
- The *k*th eigenvalue represents **the variance displayed ("explained" or "extracted") by the *k*th PC axis**.
- The sum of the first *k* eigenvalues is the variance explained by the *k*-dimensional ordination.

Example

- λ_1=9.8783, λ_2=3.0308, Trace = 12.9091
- PC 1 displays (“explains”) 9.8783/12.9091 = 76.5% of the total variance.



The Algebra of PCA

- The cross-products matrix computed among the *p* principal axes has a simple form:
  - All off-diagonal values are zero (the principal axes are uncorrelated).
  - The diagonal values are the eigenvalues.

	PC_1	PC_2
PC_1	9.8783	0.0000
PC_2	0.0000	3.0308

Variance-covariance Matrix of the PC axes

A More Challenging Example

- Data from research on habitat definition in the endangered Baw Baw frog.
- 16 environmental and structural variables measured at each of 124 sites.
- Correlation matrix used because variables have different units.

Eigenvalues

Axis	Eigenvalue	% of Variance	Cumulative % of Variance
1	5.855	36.60	36.60
2	3.420	21.38	57.97
3	1.122	7.01	64.98
4	1.116	6.97	71.95
5	0.982	6.14	78.09
6	0.725	4.53	82.62
7	0.563	3.52	86.14
8	0.529	3.31	89.45
9	0.476	2.98	92.42
10	0.375	2.35	94.77

Interpreting Eigenvectors

- Each element of the eigenvectors represents the contribution of a given variable to a component.
- **(Pearson linear) Correlations** between variables and the principal axes are known as loadings.
- i.e., a_{pk}

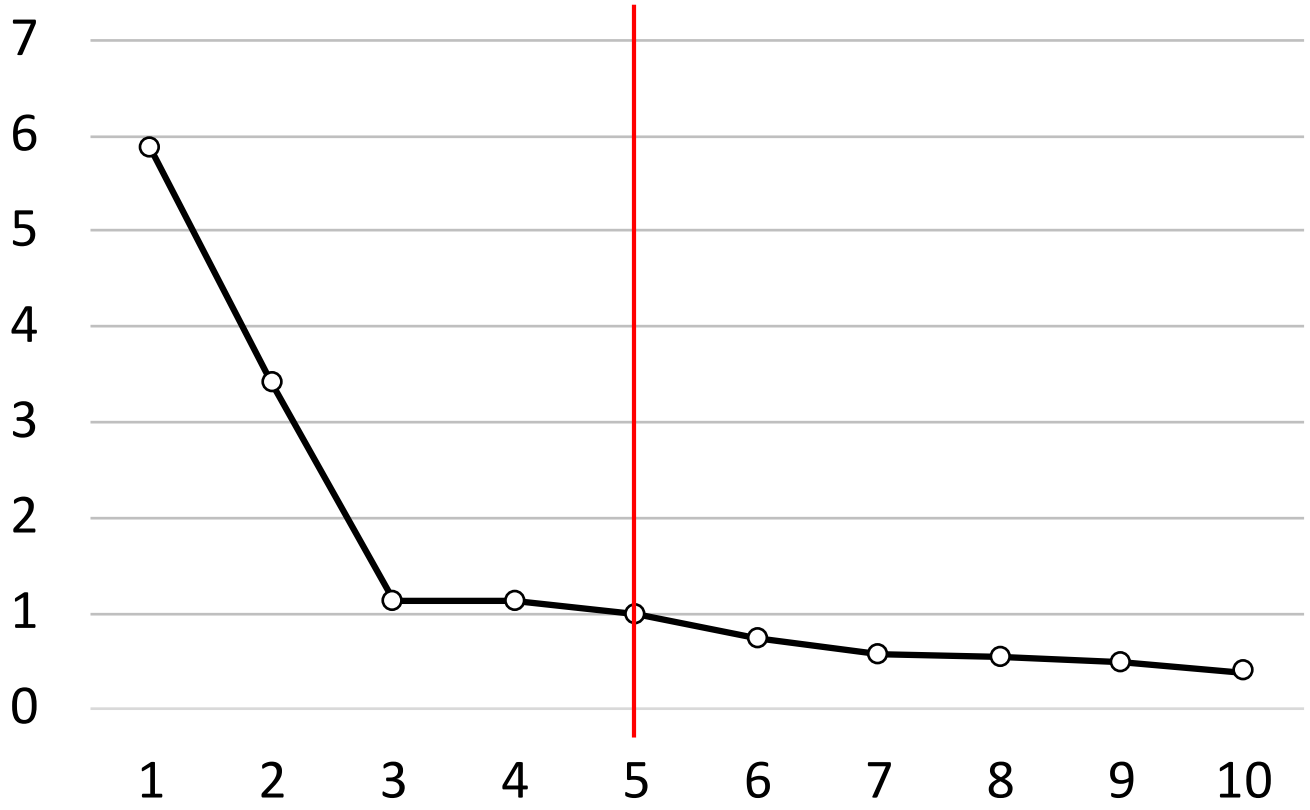
	PC1	PC2	PC3
Altitude	0.3842	0.0659	-0.1177
pH	-0.1159	0.1696	-0.5578
Cond	-0.2729	-0.1200	0.3636
TempSurf	0.0538	-0.2800	0.2621
Relief	-0.0765	0.3855	-0.1462
maxERht	0.0248	0.4879	0.2426
avERht	0.0599	0.4568	0.2497
%ER	0.0789	0.4223	0.2278
%VEG	0.3305	-0.2087	-0.0276
%LIT	-0.3053	0.1226	0.1145
%LOG	-0.3144	0.0402	-0.1067
%W	-0.0886	-0.0654	-0.1171
H1Moss	0.1364	-0.1262	0.4761
DistSWH	-0.3787	0.0101	0.0042
DistSW	-0.3494	-0.1283	0.1166
DistMF	0.3899	0.0586	-0.0175

How many axes are needed?

- Does the (*k+1*)th principal axis represent more variance than would be expected by chance?
- Several tests and rules have been proposed.
- A common “rule of thumb” when PCA is based on correlations is that axes with **eigenvalues > 1** are worth interpreting

Example

- Baw Baw Frog – PCA of 16 Habitat Variables



What are the assumptions of PCA?

- Assumes relationships among variables are **Linear**.
  - Cloud of points in *p*-dimensional space has linear dimensions that can be effectively summarized by the principal axes
- If the structure in the data is **Non-Linear** (the cloud of points twists and curves its way through *p*-dimensional space), the principal axes will not be an efficient and informative summary of the data.

When should PCA be used?

- PCA is useful for summarizing variables whose relationships are **approximately linear or at least monotonic**.
  - e.g. PCA of many soil properties might be used to extract a few components that summarize main dimensions of soil variation
- PCA is generally NOT useful for data having highly nonlinear relationships.

Summary

- Principal component (PC): Linear combination (variate) of the original variables. PC also represent **the underlying dimensions** (axes) that summarize or account for the original set of observed variables.
- PC loadings: Correlation between the original variables and the PC scores, and the key to understanding the underlying nature of a particular Principal component. Squared PC loadings indicate what percentage of the variance in an original variable is explained by a PC.
- PC score: Composite measure created **for each observation** on each PC extracted in the PCA.

Questions?