BELK COLLEGE *of* BUSINESS
UNC CHARLOTTE

Predictive Modeling—Dimensionality Reduction

DSBA/MBAD 6211: Advanced Business Analytics

Spring 2024

High Dimensionality

- Real data usually have hundreds, thousands, or even millions of dimensions

- Examples:
  - Health status of patients: *100+ measured/recorded parameters from blood analysis, immune system status, genetic background, nutrition, alcohol- tobacco- drug-consumption, operations, treatments, diagnosed diseases, ...*
  - Seller reputation: *Seller rate, Percentage of seller good rate, Description score, Description votes, Service score, Service votes, Delivery score, Delivery votes, Placement on search results, Violation, Return conflict, Complaint*
  - Web documents, *where the dimensionality is the vocabulary of words*
  - Facebook graph, *where the dimensionality is the number of users*

The Curse of Dimensionality

- Problems with high-dimensional data
  - Number of samples required
  - Running time
  - Multicollinearity
  - Data becomes very sparse, some algorithms become meaningless (e.g., density-based clustering)
  - The complexity of some algorithms depends on the dimensionality, and they become infeasible

Dimensionality Reduction

- Usually, the data can be described with fewer dimensions, without losing much of the meaning of the data.
  - The data reside in a space of lower dimensionality
- Essentially, we assume that some of the data is noise, and we can approximate the useful part with a lower dimensionality space.
  - Dimensionality reduction does not just reduce the amount of data, it often brings out the useful part of the data

Why Reduce Dimensions

- Discover hidden patterns
- Remove redundant and noisy features
- Interpretation and visualization
- Easier storage and processing of the data

Approaches

1) Incorporating domain knowledge to remove or combine categories
2) Using data summaries to detect information overlap between variables
3) Using data conversion techniques such as converting categorical variables into fewer categories
4) Employing automated reduction techniques (e.g., PCA).

The Role of the Target Variable

- Supervised method
  - Use the target variable in the reduction method
- Unsupervised method
  - Ignore the target variable in the reduction method

Outcome

- Variable selection
  - Use one or a subset of the original variables as inputs into subsequent models
- Dimension reduction
  - Use combinations of the original variables as inputs into subsequent models

Comparison of Methods

Method	Target	Outputs
PCA	Not used	Constructed
Variable Clustering	Not used	Original or Constructed
LAR/LASSO	Used	Original
Linear discriminant analysis (LDA)	Used	Constructed

Dimensionality Reduction

- Purposes:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise
- Techniques
  - Principal Component Analysis (PCA)
  - Singular Value Decomposition (SVD)
  - Others: supervised and non-linear techniques

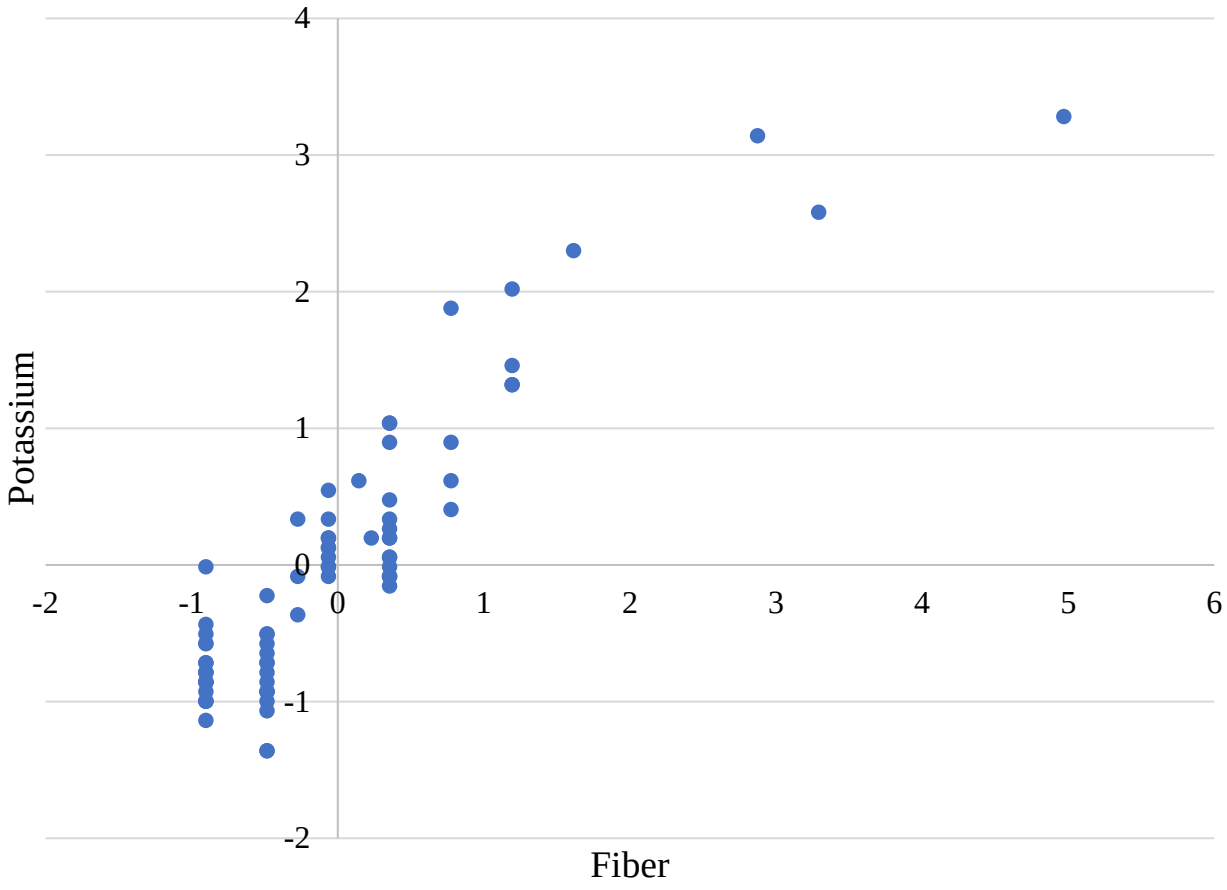
Principal Component Analysis (PCA)

- The idea:
  - Find linear combinations of variables that contains most, even if not all, of the information, so that these new variables can replace the original variables.

- PCA is one of the oldest and most popular dimensionality reduction procedure.

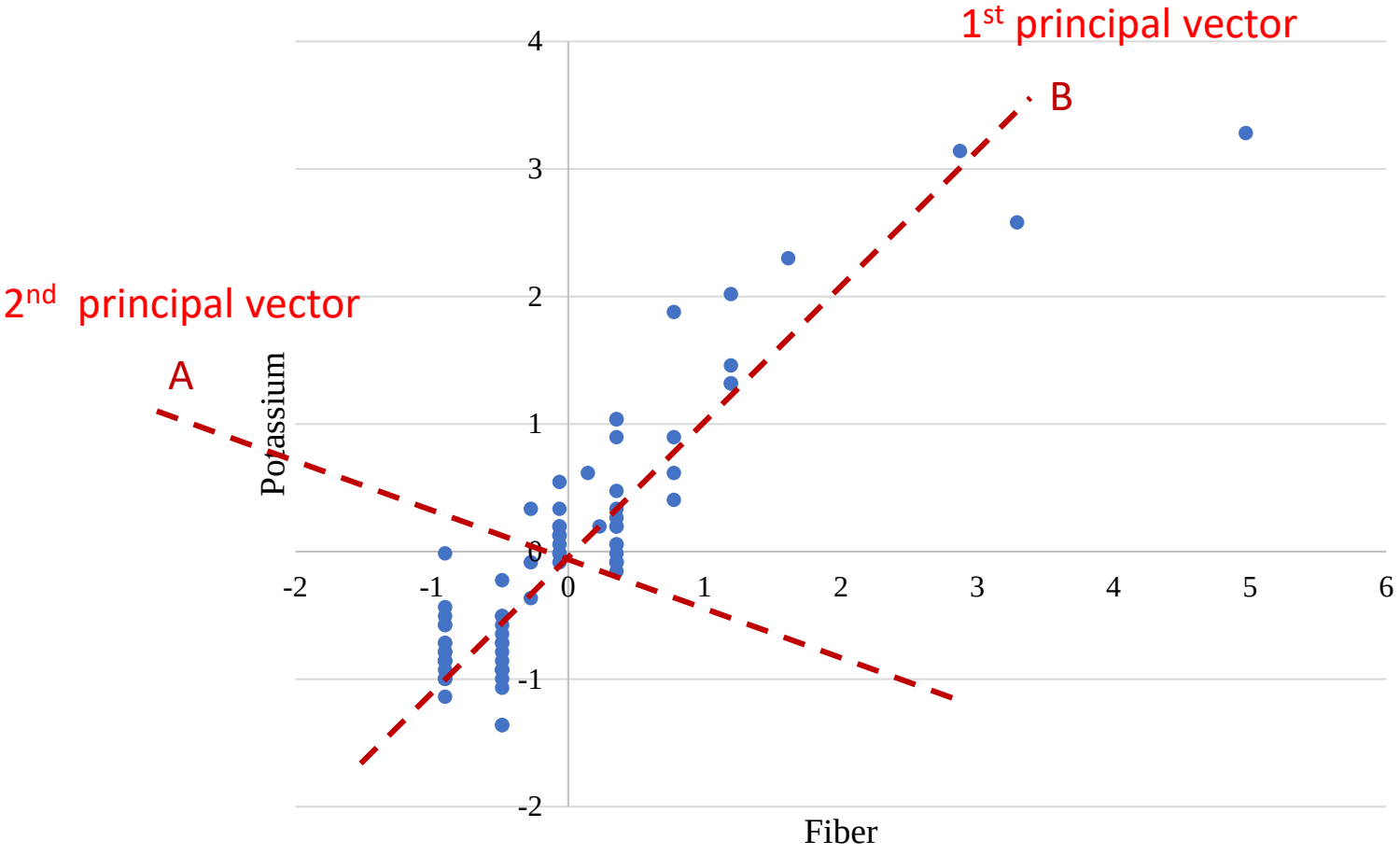
The First Principal Component

- Scaling the data



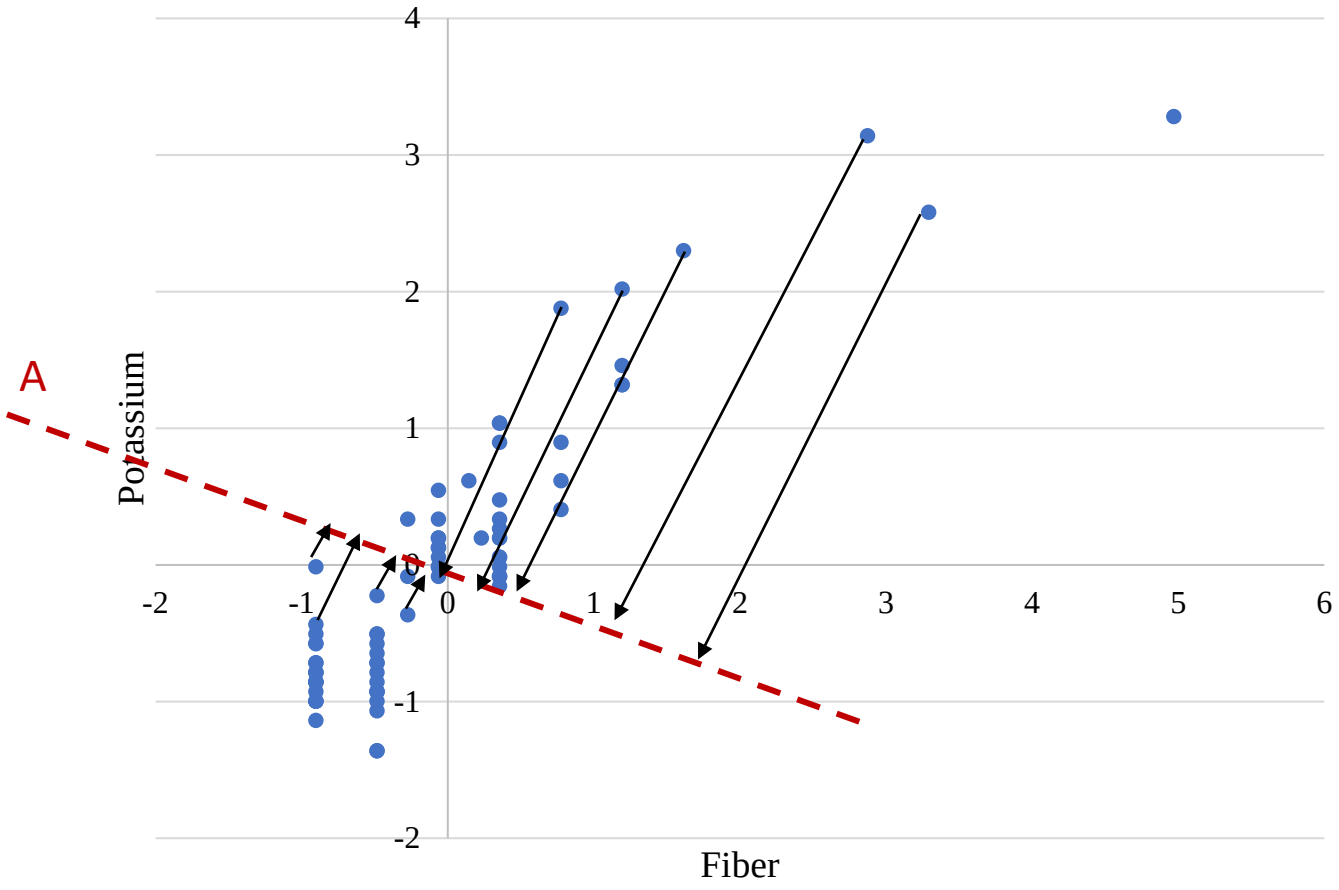
The First Principal Component

- Projection



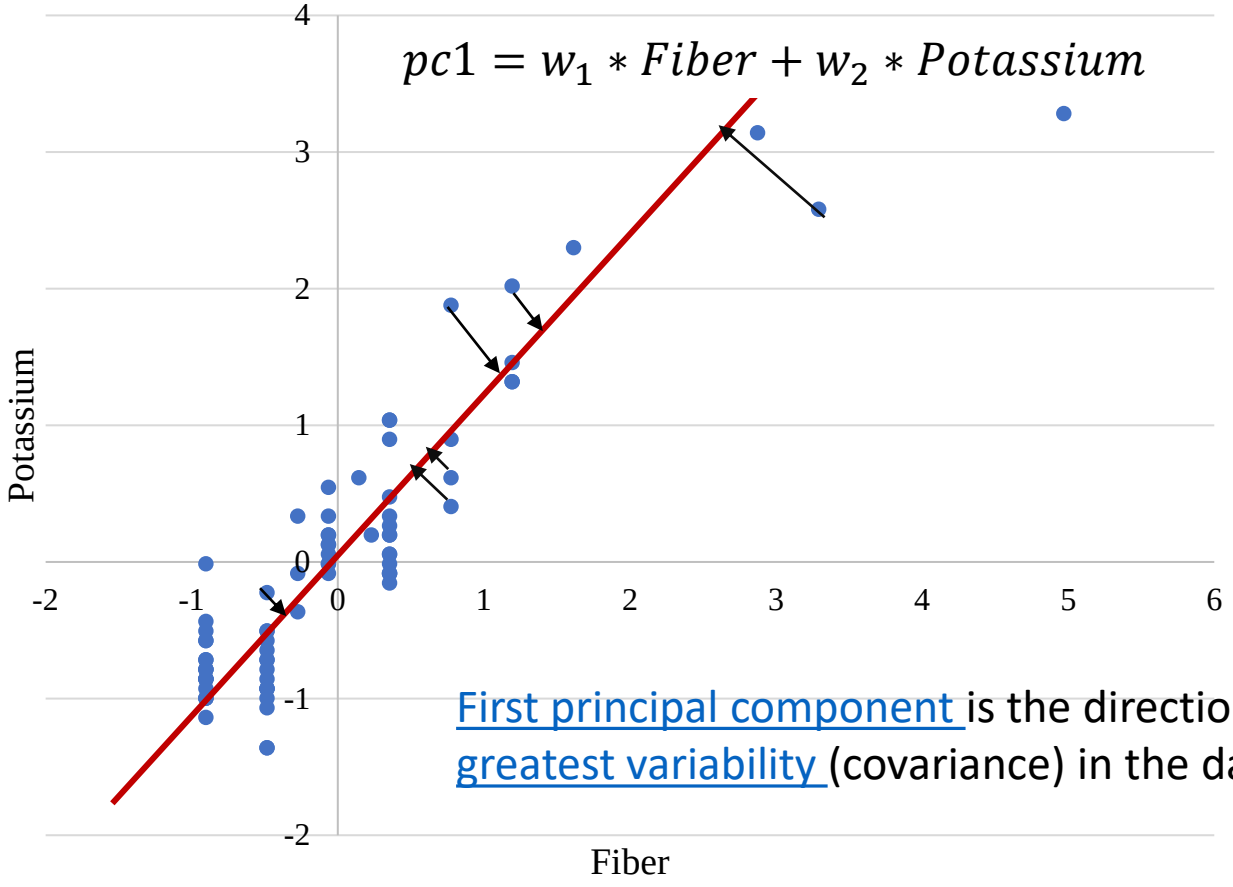
The First Principal Component

- Projection



The First Principal Component

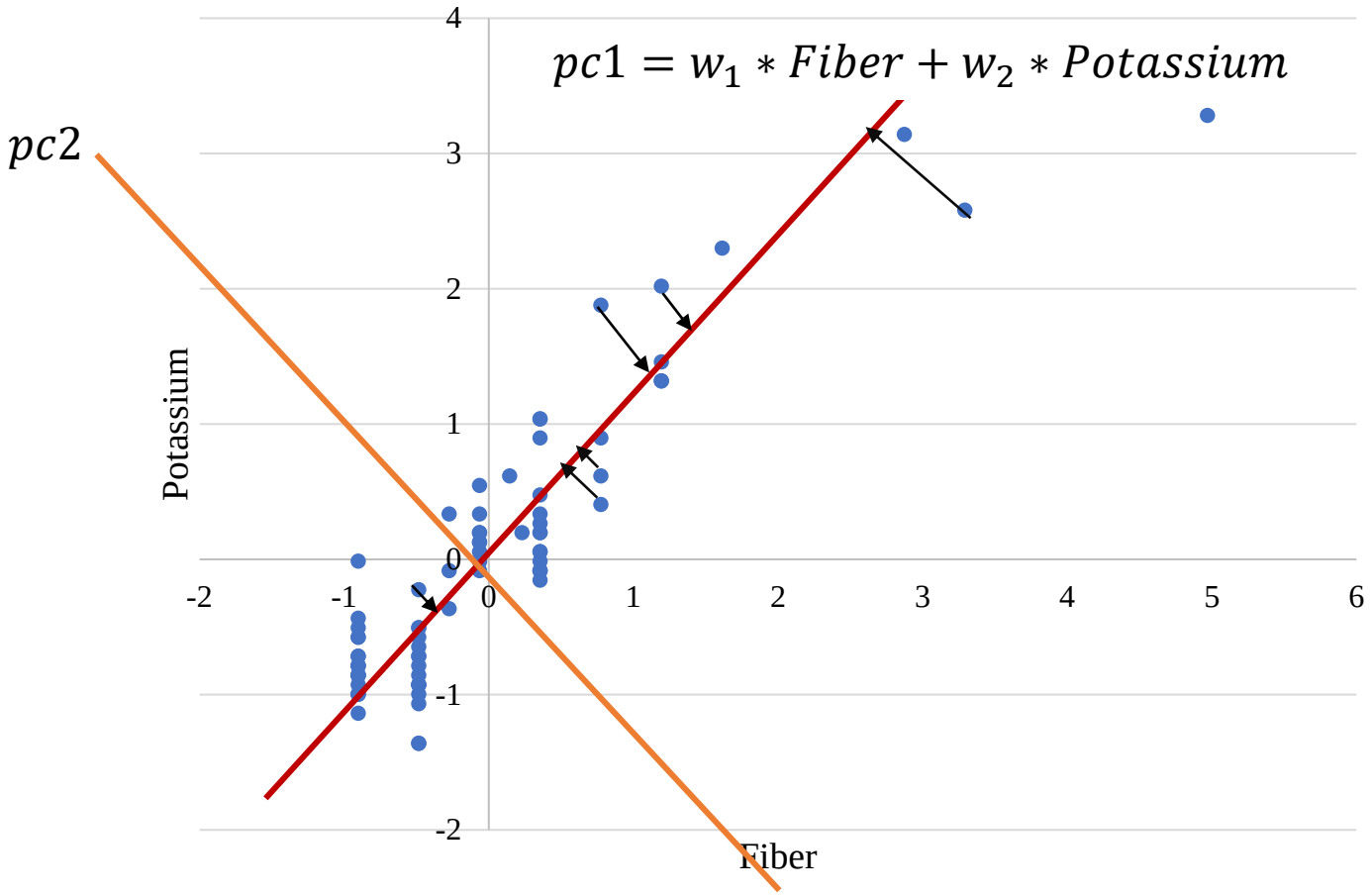
- Linear combination



First principal component is the direction of greatest variability (covariance) in the data

PCA approximating a high dimensional data set with a lower dimensional linear subspace

- The second principal component



UNC CHARLOTTE

PCA Learning Rule

- Objective Function:
  - Let $y_n = x_n^T w$ be the project of the data point on w. The variance of the entire projected data is given by: $\sum_{n=1}^{N}(y_n)^2 = \sum_{n=1}^{N}(x_n^T w)^2$. Thus, the first principal component is found by solving:

$$\text{Max} \sum_{n=1}^{N}(\boldsymbol{x}_n^T \boldsymbol{w})^2$$

$$\text{Subject to: } \boldsymbol{w}^T \boldsymbol{w} = 1$$

PCA

From N original variables: $x_1, x_2, \ldots, x_N$:

Produce k new variables: $y_1, y_2, \ldots, y_N$:

$$y_1 = w_{11}x_1 + w_{12}x_2 + \ldots + w_{1N}x_N$$
$$y_2 = w_{21}x_1 + w_{22}x_2 + \ldots + w_{2N}x_N$$
$$\ldots$$
$$y_N = w_{N1}x_1 + w_{N2}x_2 + \ldots + w_{NN}x_N$$

y_N's are *Principal Components*

such that:

y_N's are uncorrelated (orthogonal)

y_1 explains as much as possible of original variance in data set

y_2 explains as much as possible of remaining variance etc.

Eigenvector

$\{w_{11},w_{12},...,w_{1N}\}$ is 1st **Eigenvector** of correlation/covariance matrix, and **coefficients** of first principal component

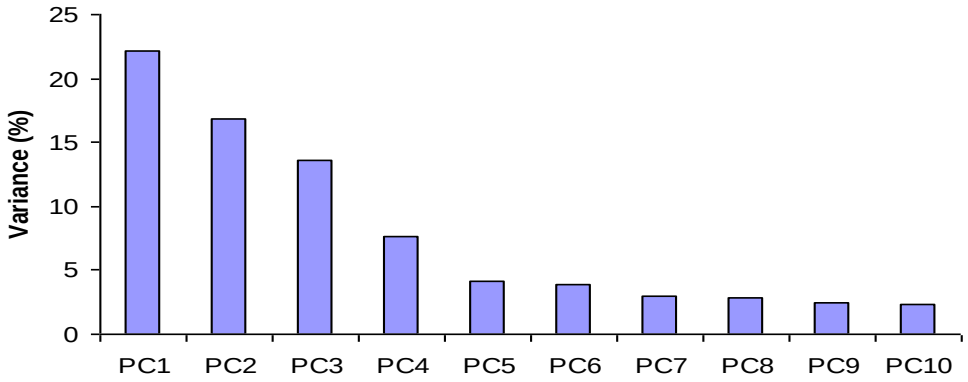
$\{w_{21},w_{22},...,w_{2N}\}$ is 2nd **Eigenvector** of correlation/covariance matrix, and **coefficients** of 2nd principal component

……

$\{w_{N1},w_{N2},...,w_{NN}\}$ is kth **Eigenvector** of correlation/covariance matrix, and **coefficients** of kth principal component

PCA for Dimension Reduction

- With *N* input variables, you can compute *N* principal components
- 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



PCA Results

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.21358154	1.44403082	0.7379	0.7379
2	0.76955072	0.75268297	0.2565	0.9944
3	0.01686775		0.0056	1.0000

Eigenvectors			
	Prin1	Prin2	Prin3
x1	0.650940	0.263685	0.711862
x2	0.645235	0.301851	-.701825
x3	-.399937	0.916164	0.026348

PCA example

Factor Analysis

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
item01 motivation	2.99	.918	71
item02 pleasure	3.58	.822	71
item03 competence	2.82	.915	71
item04 low motiv	2.21	.909	71
item05 low comp	1.61	.948	71
item06 low pleas	2.44	.996	71
item07 motivation	2.77	1.072	71
item08 low motiv	1.96	.917	71
item09 competence	3.32	.770	71
item10 low pleas	1.41	.748	71
item11 low comp	1.38	.763	71
item12 motivation	2.99	.837	71
item13 motivation	2.68	.807	71
item14 pleasure	2.86	.723	71

UNC CHARLOTTE

https://tandfbis.s3.amazonaws.com/rt-media/pdf/9781848729995/IBM_SPSS_5e_Chapter_4.pdf

PCA example

Eigenvalues refer to the variance accounted for, in terms of the number of "items' worth" of variance each explains. So, Factor 1 explains almost as much variance as in five items.

Percent of covariation among items accounted for by each factor before and after rotation.

Total Variance Explained

	Initial Eigenvalues			Rotation Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.888	34.916	34.916	3.017	21.549	21.549
2	2.000	14.284	49.200	2.327	16.621	38.171
3	1.613	11.519	60.719	1.784	12.746	50.917
4	1.134	8.097	68.816			
5	.904	6.459	75.275			
6	.716	5.113	80.388			
7	.577	4.125	84.513			
8	.461	3.293	87.806			
9	.400	2.857	90.664			
10	.379	2.710	93.374			
11	.298	2.126	95.500			
12	.258	1.846	97.346			
13	.217	1.551	98.897			
14	.154	1.103	100.000			

Extraction Method: Principal Axis Factoring.

Half of the variance is accounted for by the first three factors.

What are principal components

Rotated Factor Matrix[a]

	Factor		
	1	2	3
item05 low comp	-.897		
item03 competence	.780		
item01 motivation	.777		
item11 low comp	-.572		.355
item12 motivation		.721	
item13 motivation		.667	
item08 low motiv		-.619	
item04 low motiv		-.601	
item07 motivation	.412	.585	
item09 competence		.332	
item14 pleasure			-.797
item10 low pleas			.580
item02 pleasure	.487		-.535
item06 low pleas			.515

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

The items cluster into these three groups defined by the highest loading on each item.

PCA: Pros

- Constructed output variables are definitely uncorrelated.
- The selection order of the principal components is automatically determined.
- Often, a very small number of principal components must be kept in order to explain a lot of the variations in the data cloud.

PCA: Cons

- Difficult or impossible to interpret the constructed principal components.
- All original input variables still used, since they build the principal components.
- Misinterpretation of the coefficients of the linear combinations is common.