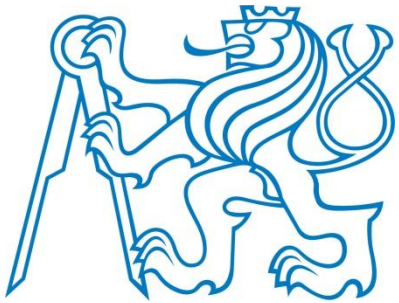
Experimental Data Analysis

in ©MATLAB

Lecture 12:

Dimension reduction using Principal Component Analysis, data interpretation

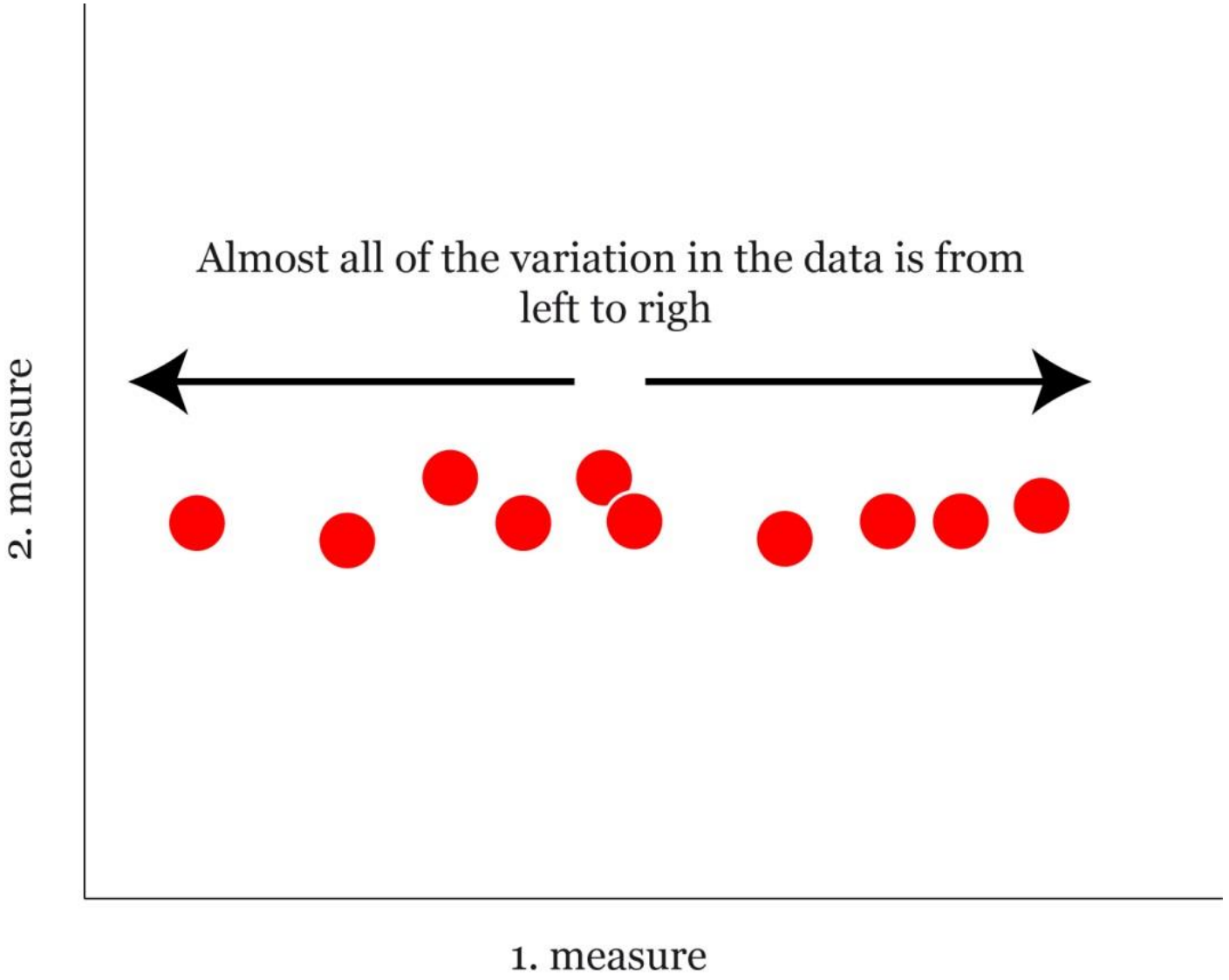
Jan Rusz
Czech Technical University in Prague



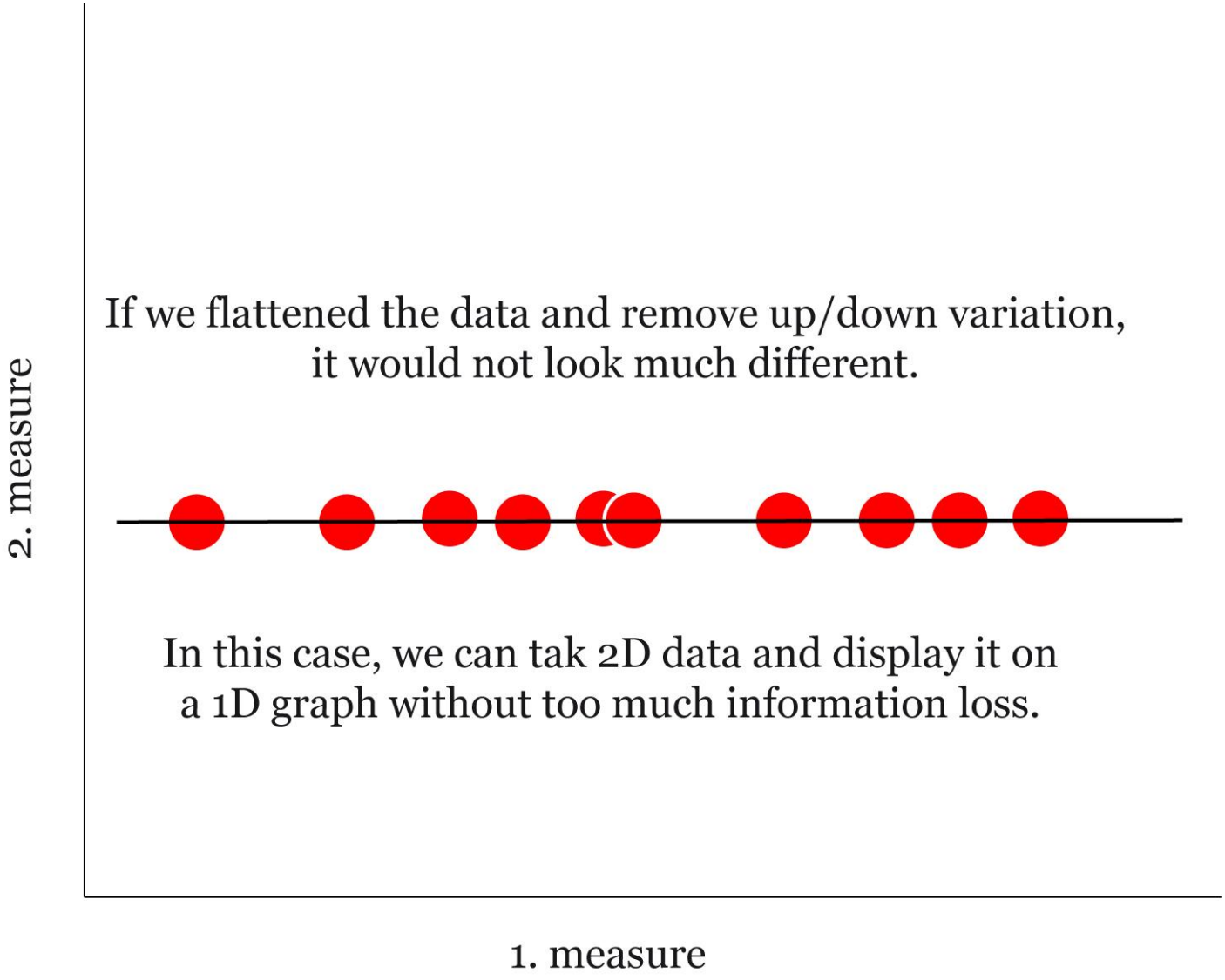
What if we have a big number of dimensions?

- 1 measure = 1D graph (number line)
- 2 measures = 2D graph (normal x/y graph)
- 3 measure = 3D graph (fancy graph with depth)
- 4 measure = 4D graph (you cannot draw it in 3D world)
- …
- 100 measures = 100D graph (etc.)

What if we have data that look like this?



What if we have data that look like this?



Example with movies in TV

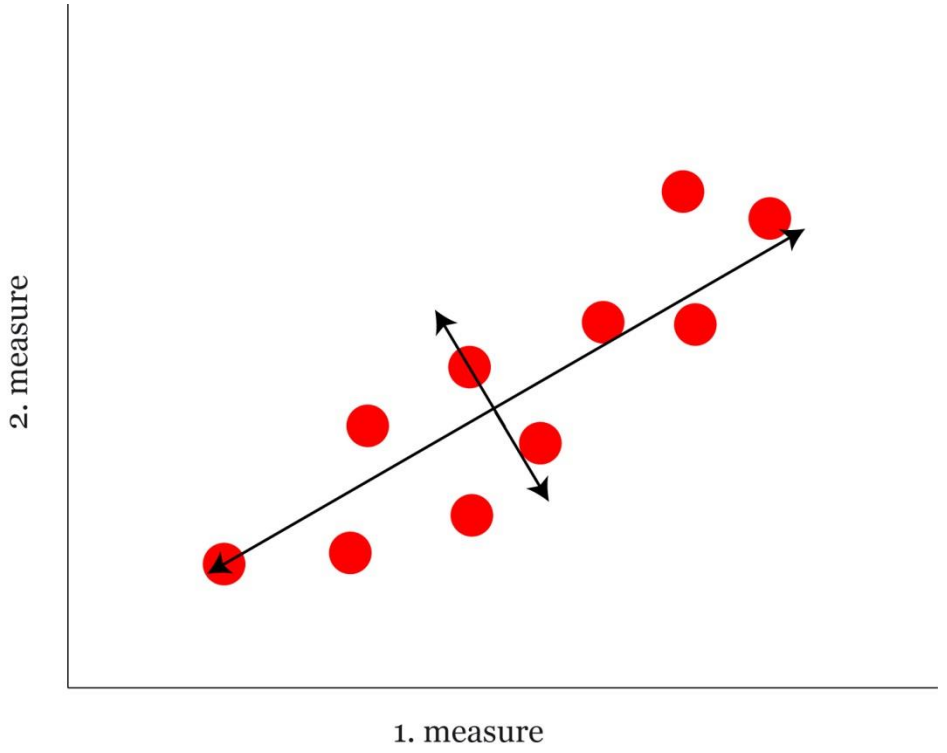
- TV and movies are almost always in 2D, even though subjects are 3D.
- This is OK. The 3rd dimension does not usually add much to the story.
- Things still look believable without it.
- People look like people, things look like things, even when they have no depth and are flat on screen.
- In other words, camera takes 3D information and flattens it to 2D without loss of too much information.

How to benefit from these information? Answer is Principal Component Analysis (PCA)

- Some dimensions (measures) are more important than others.
- PCA takes a dataset with lot of dimensions and flattens it to 2 or 3 (or more) dimensions, allowing us to look at it.
- PCA tries to find a meaningful way to flatten the data by focusing on the things that are different between individual measures.

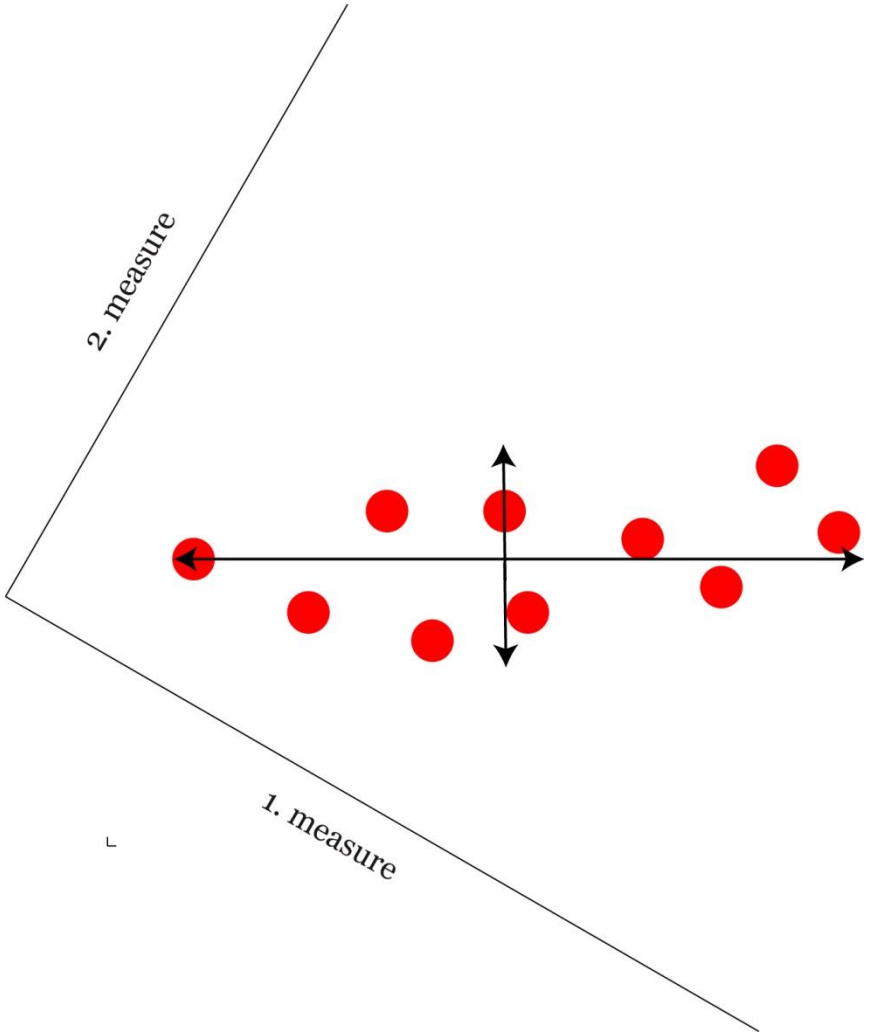
PCA example

- Dots are spread along a diagonal line.
- In other words, the maximum variation is between two endpoints of the first left-right line.
- Dots are also spread out little above and below the first line.
- In other words, the 2nd largest amount of variation is at the endpoints of the second up-down line.

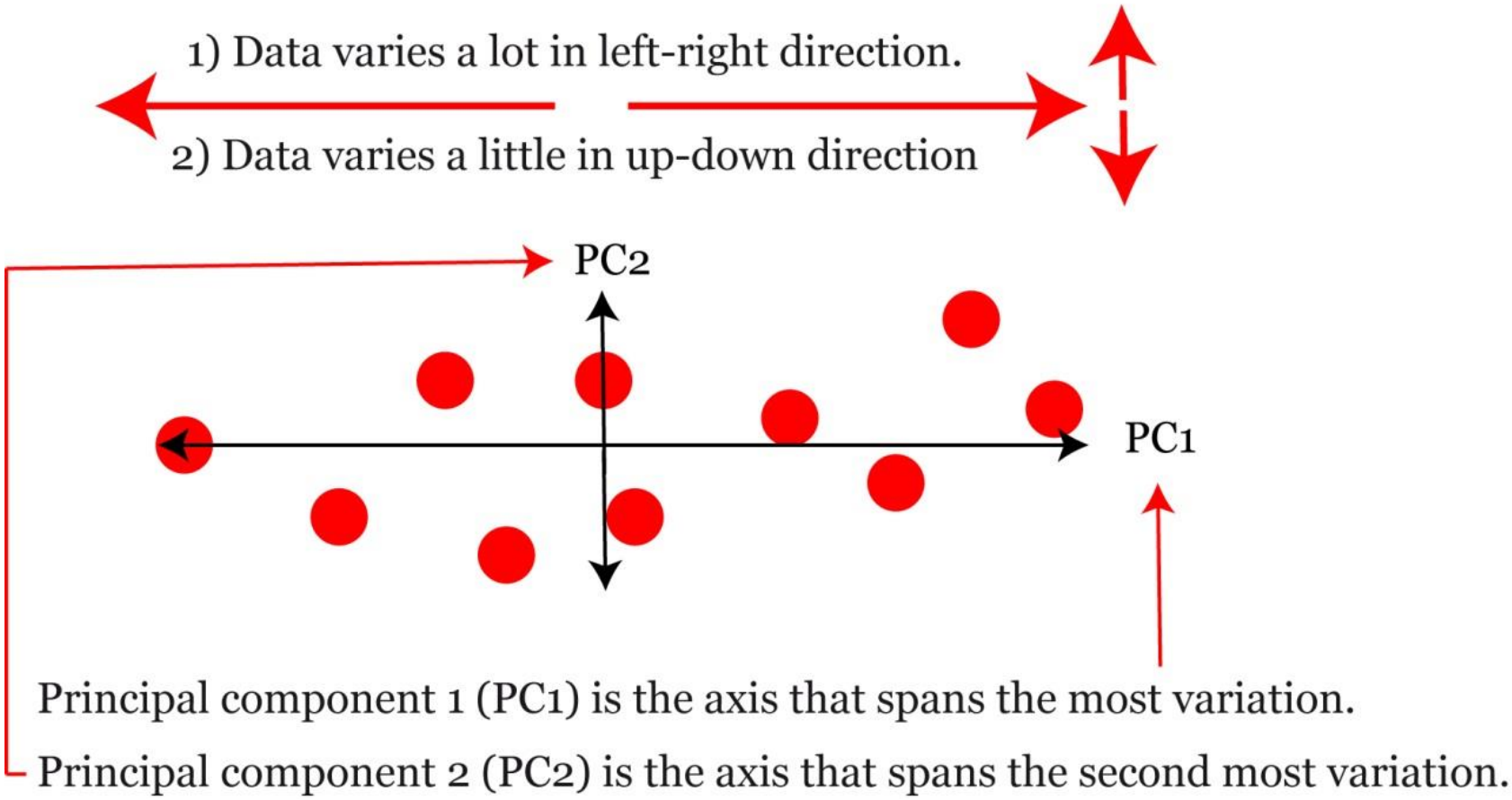


PCA example

- If we rotate the entire graph (i.e. perform the linear transformation using eigenvalues), the two lines now make new X and Y axes.
- This make the left-right and up/down variations easier to see.



PCA example


1) Data varies a lot in left-right direction.
2) Data varies a little in up-down direction
PC2
PC1
Principal component 1 (PC1) is the axis that spans the most variation.
Principal component 2 (PC2) is the axis that spans the second most variation.

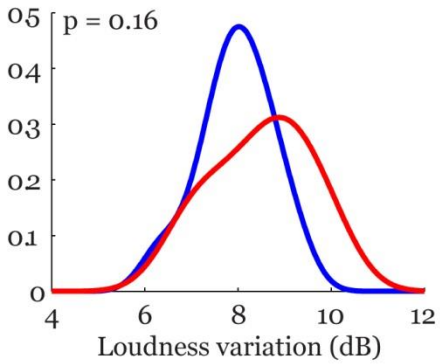
What if we had 4 measures?

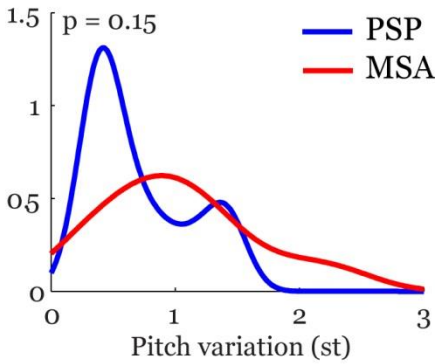
- PC1 would span the direction of the most variation.
- PC2 would span the direction of the 2nd most variation.
- PC3 would span the direction of the 3rd most variation.
- PC4 would span the direction of the 4th most variation.
- …
- If we had 100 measures, we would have 100 principal components.

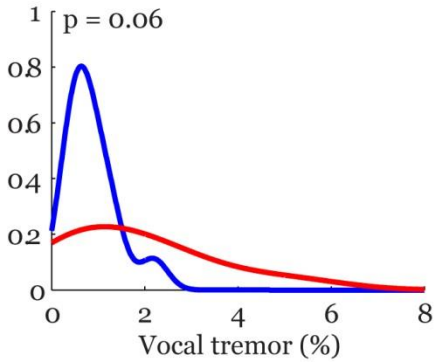
Research example

Matlab example 1

Progressive supranuclear palsy (PSP) and multiple system atrophy (MSA) represent atypical form of parkinsonism where the correct diagnosis is challenging. In addition, as both are rare diseases, it is difficult to obtain sufficient sample size for proper statistical power. We collected data from 12 PSP and 13 MSA subjects and measured three speech aspects of loudness variation, pitch variation and vocal tremor.







To correct Type I error

$p = 0.05/3 = 0.0167$

1. step: define basic parameters

A

25x3 double

	1	2	3
1	8.3059	1.4011	2.1900
2	7.1105	0.5787	1.3600
3	8.6655	0.3663	0.4300
4	7.3971	0.8993	1.1300
5	8.0218	0.7083	0.2900
6	7.8917	0.4120	0.7100
7	8.3066	0.3810	0.7000
8	6.3430	0.4250	0.5200
9	9.2462	0.2788	0.3600
10	8.7645	1.4253	1.1290
11	7.8460	0.4437	0.7875
12	7.6284	1.1517	0.7100
13	6.7261	1.9078	2.8900
14	8.4277	0.6254	2.2900
15	9.1019	1.3495	2.3700
16	9.8960	1.3701	1.5900
17	9.1974	0.1637	0.2200
18	10.0249	1.1726	5.3900
19	7.8946	2.3215	4.0500
20	8.6675	0.4982	0.3700
21	7.2036	0.2900	0.3400
22	8.8288	0.8556	0.4100
23	9.3600	0.8466	0.8000
24	7.9961	0.8184	1
25	7.1668	1.0825	1.8200

```
>> [n m] = size(A)
n =
    25
m =
     3

>> A_mean = mean(A)
A_mean =
    8.2407    0.8709    1.3543

>> A_std = std(A)
A_std =
    0.9545    0.5420    1.2707
```

2. step: normalize the data to z-score

B

25x3 double

	1	2	3
1	0.0683	0.9781	0.6577
2	-1.1842	-0.5391	0.0045
3	0.4450	-0.9310	-0.7274
4	-0.8839	0.0524	-0.1765
5	-0.2294	-0.3000	-0.8375
6	-0.3657	-0.8467	-0.5070
7	0.0690	-0.9039	-0.5149
8	-1.9883	-0.8227	-0.6565
9	1.0534	-1.0924	-0.7825
10	0.5487	1.0228	-0.1773
11	-0.4136	-0.7882	-0.4460
12	-0.6416	0.5180	-0.5070
13	-1.5869	1.9130	1.2086
14	0.1959	-0.4530	0.7364
15	0.9022	0.8829	0.7994
16	1.7342	0.9210	0.1855
17	1.0023	-1.3048	-0.8926
18	1.8693	0.5566	3.1760
19	-0.3627	2.6762	2.1215
20	0.4471	-0.6877	-0.7746
21	-1.0866	-1.0718	-0.7982
22	0.6161	-0.0283	-0.7431
23	1.1727	-0.0449	-0.4362
24	-0.2563	-0.0969	-0.2788
25	-1.1252	0.3903	0.3665

```
>> B = (A - repmat(A_mean,[n 1])) ./ repmat(A_std,[n 1])
% B = zscore(A)
```

3. step: calculate PCA

score

25x3 double

	1	2	3
1	1.1500	-0.1524	-0.2199
2	-0.5559	-1.0336	0.5619
3	-1.0855	0.6422	0.0570
4	-0.2287	-0.8732	-0.0212
5	-0.8353	-0.1563	-0.3492
6	-0.9999	-0.1669	0.2828
7	-0.9761	0.2696	0.2480
8	-1.3460	-1.7537	0.4169
9	-1.1394	1.2713	0.0332
10	0.6664	0.3061	-0.9170
11	-0.9241	-0.2267	0.2932
12	-0.1031	-0.7411	-0.6142
13	1.9217	-1.9749	-0.2131
14	0.2395	0.2921	0.8020
15	1.3166	0.6817	-0.1858
16	1.0405	1.4843	-0.7775
17	-1.3717	1.2690	0.1097
18	2.9253	1.6967	1.5669
19	3.2870	-0.9526	-0.2870
20	-0.9508	0.5898	-0.1444
21	-1.4751	-0.8192	0.3459
22	-0.4474	0.6070	-0.6033
23	-0.1537	1.1531	-0.4628
24	-0.3045	-0.2281	-0.0899
25	0.3500	-1.1841	0.1679

```
>> [PCA_coeff, score, latent] = pca(B);
```

- PCA_coeff = coefficients for the principal components
- score = representations of data in the principal component space
- latent = variance of the respective principal components

Hand calculation using **eigenvalues**:

```
>> [PCA_coeff D] = eig(cov(B));
>> latent = diag(D);
>> score = B*PCA_coeff;
```

PCA_coeff

3x3 double

	1	2	3
1	0.1583	0.9747	-0.1580
2	0.6892	-0.2237	-0.6892
3	0.7071	-2.4413e-...	0.7071

eigenvector of a linear transformation is a non-zero vector that only changes by an overall scale when that linear transformation is applied to it.

3. step: notes

- order of the principal components from *pca* is opposite of that from *eig(cov(B))*, i.e. *pca* orders the principal components so that the first one appears in column 1, whereas *eig(cov(B))* stores it in the last column.
- some of the coefficients from each method have the opposite sign as there is no "natural" orientation for principal components, so you can expect different software to produce different mixes of signs.

optional step: reverse information

A_new_minus_A

25x3 double

	1	2	3
1	0	2.2204e-...	8.8818e-...
2	0	0	-2.2204e-...
3	0	-1.6653e-...	-6.1062e-...
4	0	0	0
5	0	-2.2204e-...	-3.8858e-...
6	0	-1.6653e-...	-4.4409e-...
7	0	-2.2204e-...	-5.5511e-...
8	0	-1.6653e-...	-4.4409e-...
9	0	-1.6653e-...	-6.6613e-...
10	0	2.2204e-...	4.4409e-...
11	0	-1.1102e-...	-3.3307e-...
12	0	0	0
13	0	4.4409e-...	1.3323e-...
14	0	1.1102e-...	0
15	0	2.2204e-...	8.8818e-...
16	0	2.2204e-...	4.4409e-...
17	0	-2.7756e-...	-9.1593e-...
18	0	4.4409e-...	8.8818e-...
19	0	4.4409e-...	1.7764e-...
20	0	-2.7756e-...	-3.3307e-...
21	0	-1.6653e-...	-8.3267e-...
22	0	-1.1102e-...	-1.6653e-...
23	0	-1.1102e-...	-1.1102e-...
24	0	0	-2.2204e-...
25	0	0	2.2204e-...

```
% original scaled (z-score) data
>> B_new = (B * PCA_coeff) * PCA_coeff‘

% original non-scaled data
A_new = [(B_new .* repmat(A_std,[n 1])) + repmat(A_mean,[n 1])];

% alternative approach
A_new = [(score * PCA_coeff') .* repmat(A_std,[n 1]) + repmat(A_mean,[n 1])]
```

4. step: calculate cumulative variance

```
>> cumsum(var(score)) / sum(var(score))
ans =
   0.5731    0.9064    1.0000
```

- The first principal component contains 57% of the variance of the original data. Two first principal components contain 91% of the original data. A lossy data compression scheme which discarded the third principal component would compress 3 variables into 2, while losing only 9% of the variance.

```
>> r = corr(score,'type','pearson')
r =
    1.0000    0.0000   -0.0000
    0.0000    1.0000    0.0000
   -0.0000    0.0000    1.0000
```

- Principal components are completely uncorrelated

5. step: data reduction

reducedData

25x1 double

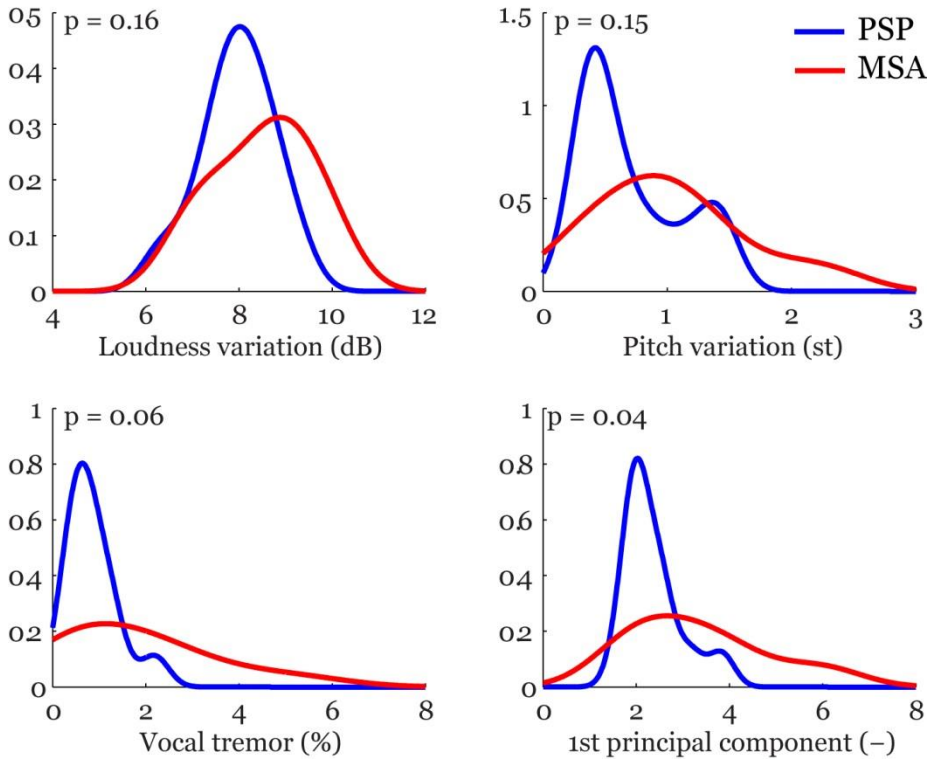
	1	2
1	3.8292	
2	2.4863	
3	1.9285	
4	2.5900	
5	1.9633	
6	2.0355	
7	2.0728	
8	1.6649	
9	1.9107	
10	3.1683	
11	2.1049	
12	2.5036	
13	4.4233	
14	3.3846	
15	4.0470	
16	3.6354	
17	1.7246	
18	6.2066	
19	5.7136	
20	1.9773	
21	1.5809	
22	2.2775	
23	2.6311	
24	2.5372	
25	3.1677	

```
% taking principal component with highest variance
>> reducedDimension = PCA_coeff(:,1)
reducedDimension =
    0.1583
    0.6892
    0.7071

% taking principal component compressing 3 measures
into 1
>> reducedData = A * reducedDimension
```

Research example

Progressive supranuclear palsy (PSP) and multiple system atrophy (MSA) represent atypical form of parkinsonism where the correct diagnosis is challenging. In addition, as both are rare diseases, it is difficult to obtain sufficient sample size for proper statistical power. We collected data from 12 PSP and 13 MSA subjects and measured three speech aspects of loudness variation, pitch variation and vocal tremor.



We do not need to correct Type I error
p should be < 0.05

How to “sell” the results of my project?

Before publishing your work, think about why you want to publish your work. Is it publishable?

- Have I done something new and interesting?
- Is there anything challenging in my work?
- Have I provided solution to some difficult problem?
- How novel is my approach?

- You have to known well state-of-the-art (perform quantitative literature review) in order not to perform redundant research/development, i.e. re-introduce the wheel.

Reviewers will likely have to answer following questions.

- Does the work contain sufficient new ideas, novelty or methods?
- **Is the work presented /written concisely and well organized?**
- **Are the methods/experiments presented in the way that they can be replicated again?**
- **Are the results presented adequately?**
- Is the discussion relevant, concise and well documented?
- **Are the conclusions supported by the data presented? Are the limitations documented and acceptable?**
- **Is the language acceptable (e.g., grammar)?**
- **Are figures and tables adequate and well designed? Are there some information duplicated?**
- Are all relevant references included and cited properly?

Pay attention to structure of the paper.

- **Introduction**: What did you/others previously do? Why did you do it?
- **Methods**: How did you do it?
- **Results**: What did you find?
- **Discussion**: What does it all mean?
- **Conclusion**: How the work advances the field from the present state of knowledge?

I. Introduction.

- What is the problem to be solved?
- Are there any existing solutions?
- Which is the best?
- What is its main limitation?
- What do you hope to achieve?

I. Tips for Introduction.

- Never use more words than necessary. Do not make this section into history lesson! Include only things necessary to introduce your own work. Long introduction put readers off.
- Do not use improper citations.
- Do not cite too many references irrelevant to the work.
- Introduction have to be organized from the global to particular point of view, guiding your readers to your objectives when writing the paper.
- Do not mix introduction with methods, discussion or conclusion. Always keep them separate to ensure that the paper flows logically from one section to the next.
- Motivation of the work should be clear after reading introduction.
- Hypotheses and objectives have to be clearly remarked at the end of the introduction.

II. Methods.

- Respond to the question of how the problem was studied.
- If you are proposing new method, you need to include detailed information allowing the others to reproduce the experiment.
- Do not repeat the details of already established methods.
- Methods provide critical knowledge; incomplete or incorrect methods represent major fail of the work.
- Always use standard system and nomenclature, for example international System of Units (SI).

II. Tips for Methods.

- Present methods in the logical order in which you did your research:
  - Description of the **materials and participants**.
  - Description of the **experiments** done (for example speech assessment).
  - Description of the **laboratory methods** (for example signal segmentation).
  - Description of the **statistical methods** used.
- Do not provide partial results as the part of methods.

III. Results.

- Provide results of your experiments, refers to the question “What have you found?”

III. Tips for Results.

- Ideally, present only representative results of your experiment. In other words, these results should be essential for discussion.
- Decide to present data in logical order that tells a clear story easy to understand. Generally, this will be in the same order as presented in methods section.
- You should provide clear results devoid of emotions and interpretations. You should never include references in this section. You are presenting **your results** here. Thus, you cannot refer to others, this is *discussion*.
- Do not list any methods in the results section.

IV. Discussion.

- You must explain your results to wider audience.
- Typically the most important section of your article.
- Here you get chance to sell your data.
- You have to make the discussion corresponding to the results but not reiterate the results.
- You need to **compare your results to previously observed results**.
- Do not ignore the work in disagreement with you findings, you should rather confront it and convince the readers that you are correct and better.

IV. Tips for Discussion.

- Avoid statements that go beyond what your results can support.
- Avoid unspecific expressions, quantitative descriptions are always preferred.
- Avoid sudden introduction of new terms and ideas; you have to introduce everything in the introduction.
- Speculations on possible interpretations are allowed, but these should be rooted in fact, rather than imagination.

To achieve good interpretation, think about:

- How do your results relate to original objectives outlined in introduction?
- Do the data support your hypothesis?
- Are your results consisted with what other investigators reported?
- Discuss weaknesses and discrepancies. Provide limitations. If your results were unexpected, try to explain why.
- Is there another way to interpret your results?
- What further research would be suitable to answer questions raised by your results?

IV. Conclusion.

- Shortly tells how your work advances the field from the present state of knowledge.
- It is typically separate section or last paragraph of discussion.

IV. Tips for Conclusion.

- Do not repeat the abstract or just list experimental results.
- Provide specific or global conclusion with respect to the objectives included in the introduction.
- Provide clear scientific justification for your work and indicate uses and extensions (if applicable). Suggest future experiments or point out those that are underway.

Figures and tables

- “Figure is worth of a thousand words.” Illustrations, including both figures and tables, are the most efficient way to present your results. Your data are driving force of the paper and thus your illustrations are critical.
- Whenever you choose table or figure, no illustration should replicate the information described elsewhere in the paper.
- Figures should preferably be freestanding, and figure and table have to be self-explanatory.

Tips for figures and tables

- Avoid too many curves in one figure.
- Select appropriate axis label size.
- Include clear symbols and data sets that are easy to distinguish.
- Preferably use color only when necessary.
- Lines joining data should be used only when you present time series or consecutive sample data.
- In tables, use same number of decimal numbers for each variable.

Abstract

- Summarize to prospective reader what you did and what the important findings in your research were.
- Together with the title, it represents advertisement of your paper.
- It gives key results but minimizes experimental details.
- Typically you have less than 250 words for abstract.

Tips for Abstract

- Use structured form, i.e. Introduction, Objectives, Methods, Results, Conclusion
- Two parts that are essential:
  - What has been done?
  - What are the main findings?

Tips for References

- You have to cite all the scientific publications on which your work is based. Do not over-inflate your paper with too many references. Avoid excessive self-citations and excessive citations of publications from the same region.
- Do not provide references to unpublished observations, articles that are not peer-reviewed, and minimize citations to articles not published in English. For example, Wiki and Google are not reliable source to cite.
- Make sure that all references are provided in the same format. You can use software such as *EndNote* to format your references.

Other tips

- Guideline introducing how to write a seminar paper including tips for developing a well-received paper.

https://www.wikihow.com/Write-a-Seminar-Paper