The data science of PCA: Myths, misuses, and missed signals

WONG Limsoon

Everyone knows principal component analysis (PCA), right?

PCA, mathematically

V is eigen vector of matrix A if A V = a V for some scalar a.
Unit eigen vector is an eigen vector of length 1.

X_{raw}	data matrix of n samples × p features
μ	mean vector of each feature
$X = X_{raw} - \mu$	the centered matrix
V	p × k unit eigenvectors of XT X; i.e. the PCA
Z = X V	the k PC projections

Inverting

X_{inv} = Z VT + μ

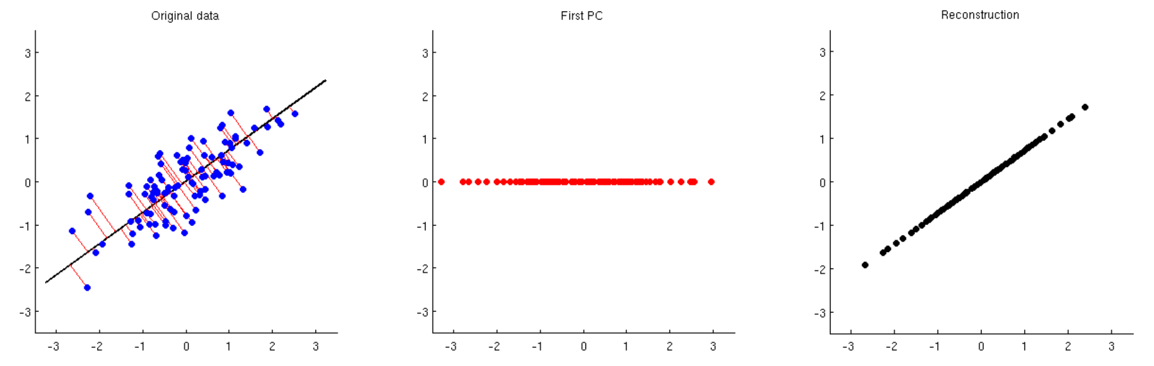
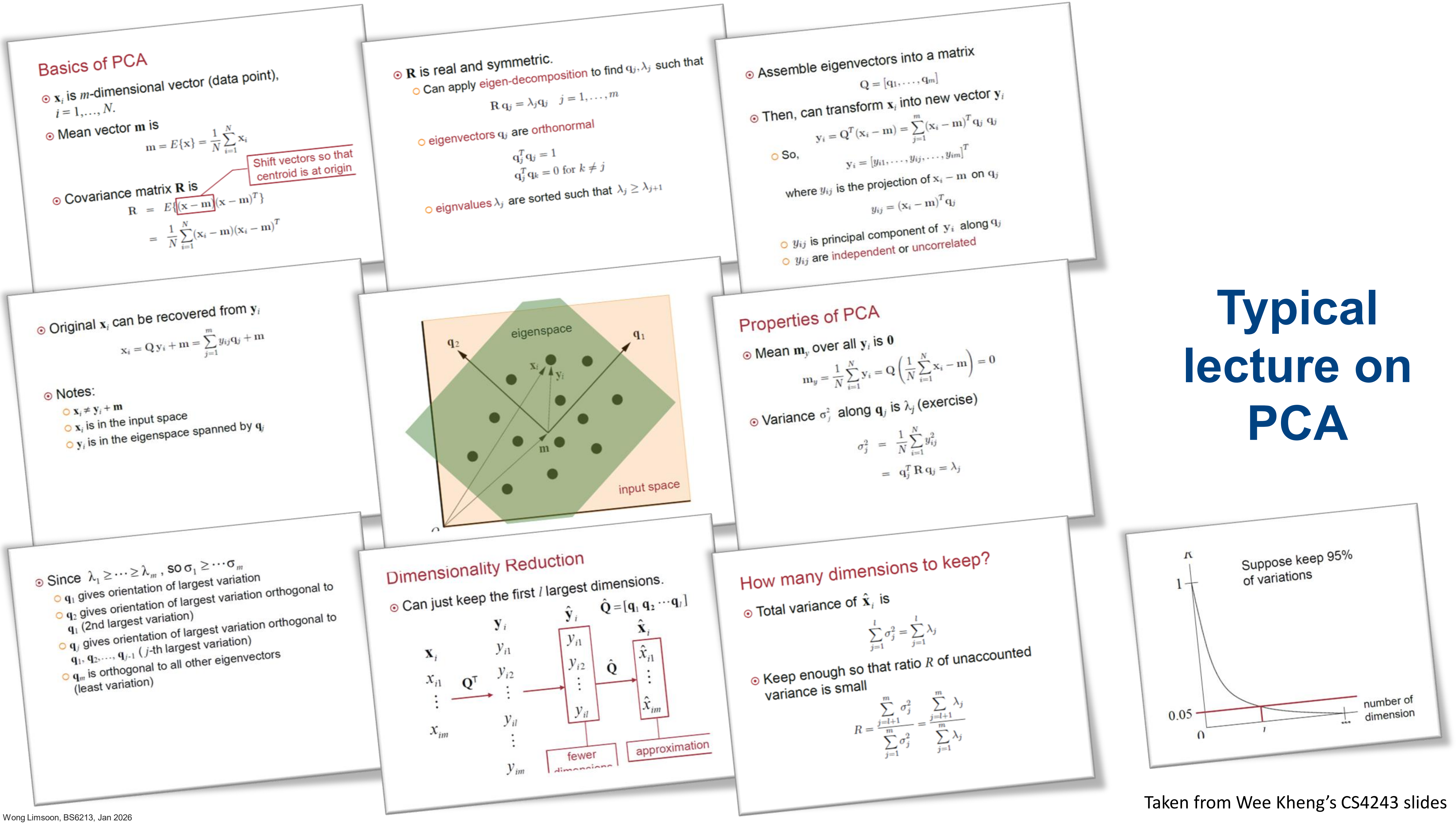


Image credit: https://stats.stackexchange.com/questions/229092/how-to-reverse-pca-and-reconstruct-original-variables-from-several-principal-com

Typical lecture on PCA



Taken from Wee Kheng's CS4243 slides

PCA, really

It deconvolutes variations in the data into (usually) meaningful directions

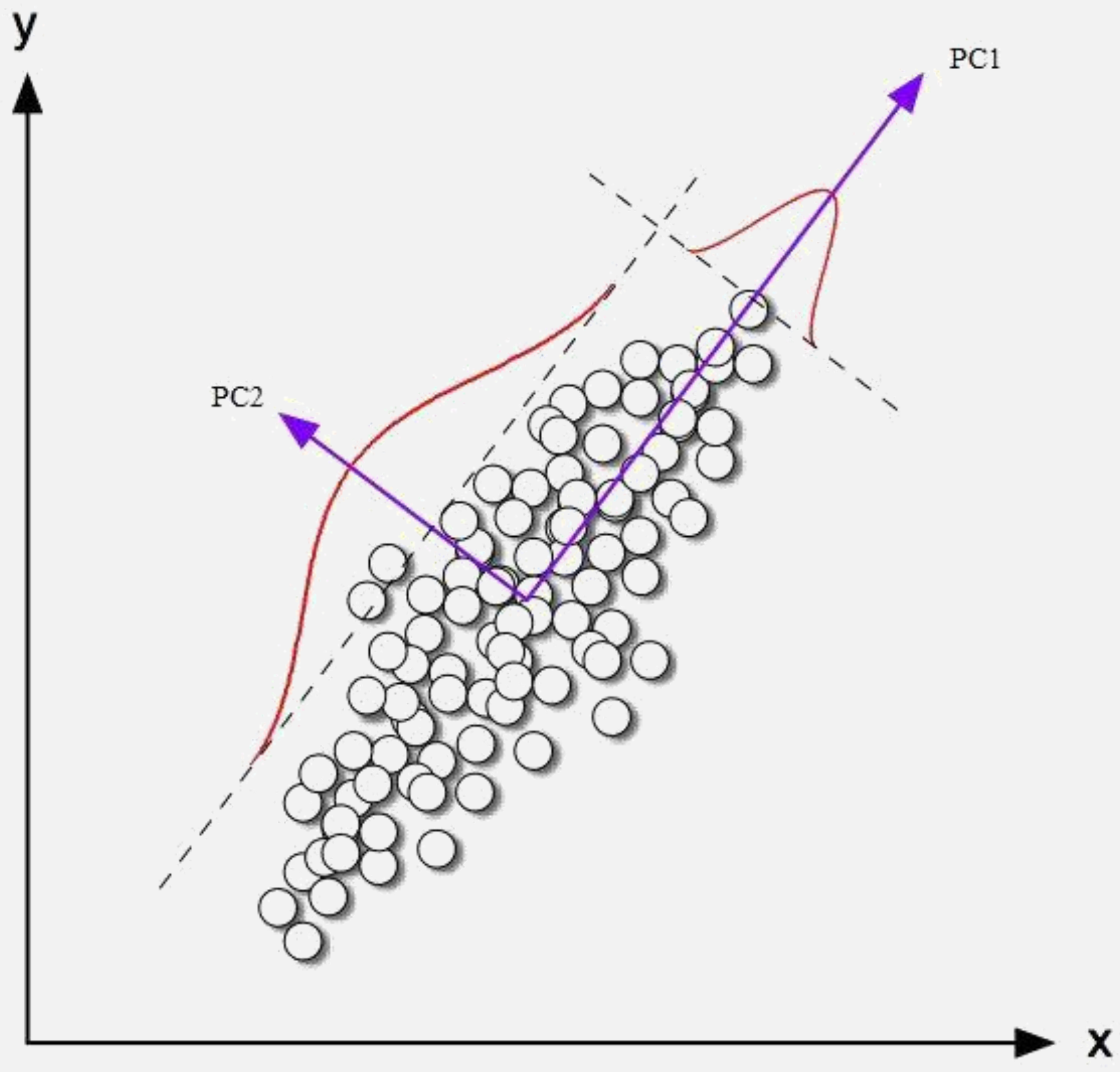


Image credit: Alessandro Giuliani

Growth, 1960, **24**, 339-354.

SIZE AND SHAPE VARIATION IN THE PAINTED TURTLE.[1] A PRINCIPAL COMPONENT ANALYSIS

PIERRE JOLICOEUR AND JAMES E. MOSIMANN[2]

Walker Museum, University of Chicago
and
Institut de Biologie, Université de Montréal

(Received for publication July 11, 1960)

TABLE 1
CARAPACE DIMENSIONS OF PAINTED TURTLES (*Chrysemys picta marginata*) IN MM.

24 Males			24 Females		
length	width	height	length	width	height
93	74	37	98	81	38
94	78	35	103	84	38
96	80	35	103	86	42
101	84	39	105	86	40
102	85	38	109	88	44
103	81	37	123	92	50
104	83	39	123	95	46
106	83	39	133	99	51
107	82	38	133	102	51
112	89	40	133	102	51
113	88	40	134	100	48
114	86	40	136	102	49
116	90	43	137	98	51
117	90	41	138	99	51
117	91	41	141	105	53
119	93	41	147	108	57
120	89	40	149	107	55
120	93	44	153	107	56
121	95	42	155	115	63
125	93	45	155	117	60
127	96	45	158	115	62
128	95	45	159	118	63
131	95	46	162	124	61
135	106	47	177	132	67

Credit: Alessandro Giuliani

Principal components

PC1= 33.78*Length +33.73*Width + 33.57*Height

PC2 = -1.57*Length – 2.33*Width + 3.93*Height

Variance of PC1

	PC1 (98%)	PC2 (1.4%)
Length	0,992	-0,067
Width	0,990	-0,100
Height	0,986	0,168

Loading / correlation of Length to PC2

Presence of an overwhelming size component explaining system variance comes from the presence of a 'typical' common shape

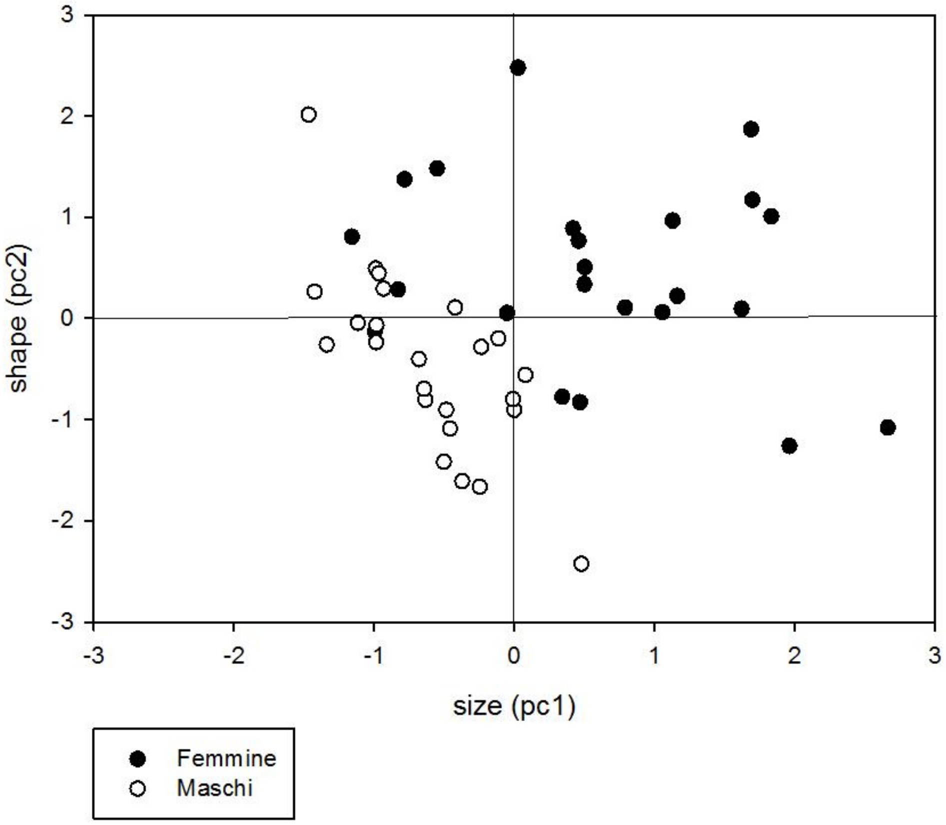
Displacement along pc1 = size variation (all positive terms)

Displacement along pc2 = shape deformation (both +ve and -ve terms)

Credit: Alessandro Giuliani

unit	sex	Length	Width	Height	PC1(size)	PC2(shape)
T25	F	98	81	38	-1,15774	0,80754832
T26	F	103	84	38	-0,99544	-0,1285916
T27	F	103	86	42	-0,7822	1,37433475
T28	F	105	86	40	-0,82922	0,28526912
T29	F	109	88	44	-0,55001	1,4815252
T30	F	123	92	50	0,027368	2,47830153
T31	F	123	95	46	-0,05281	0,05403839
T32	F	133	99	51	0,418589	0,88961967
T33	F	133	102	51	0,498425	0,33681756
T34	F	133	102	51	0,498425	0,33681756
T35	F	134	100	48	0,341684	-0,774911
T36	F	136	102	49	0,467898	-0,8289156
T37	F	137	98	51	0,457949	0,76721682
T38	F	138	99	51	0,501055	0,50628189
T39	F	141	105	53	0,790215	0,10640554
T40	F	147	108	57	1,129025	0,96505915
T41	F	149	107	55	1,055392	0,06026089
T42	F	153	107	56	1,161368	0,22145593
T43	F	155	115	63	1,687277	1,86903869
T44	F	158	115	62	1,696753	1,17117077
T45	F	159	118	63	1,833086	1,00956637
T46	F	162	124	61	1,962232	-1,261771
T47	F	177	132	67	2,662548	-1,0787317
T48	F	155	117	60	1,620491	0,09690818
T1	M	93	74	37	-1,46649	2,01289241
T2	M	94	78	35	-1,42356	0,26342486
T3	M	96	80	35	-1,33735	-0,258445
T4	M	101	84	39	-0,98842	0,49260881
T5	M	102	85	38	-0,98532	-0,2361914
T6	M	103	81	37	-1,11528	-0,0436547
T7	M	104	83	39	-0,96555	0,44687352
T8	M	106	83	39	-0,93257	0,29353841
T9	M	107	82	38	-0,98269	-0,066727
T10	M	112	89	40	-0,63393	-0,8042059
T11	M	113	88	40	-0,64405	-0,6966061
T12	M	114	86	40	-0,68078	-0,4047389
T13	M	116	90	43	-0,42133	0,10845233
T14	M	117	90	41	-0,48485	-0,9039457
T15	M	117	91	41	-0,45824	-1,0882131
T16	M	119	93	41	-0,37202	-1,610083
T17	M	120	89	40	-0,50198	-1,4175463
T18	M	120	93	44	-0,23552	-0,2831547
T19	M	121	95	42	-0,24581	-1,6640875
T20	M	125	93	45	-0,11305	-0,1986272
T21	M	127	96	45	-0,00023	-0,9047645
T22	M	128	95	45	-0,01035	-0,7971646
T23	M	131	95	46	0,079136	-0,559302
T24	M	135	106	47	0,477846	-2,4250481

Female turtles are larger and have more exaggerated height ☺



Credit: Alessandro Giuliani

A common “advice” on using PCA

“PCA is … used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible”

Wikipedia

This assumes variations in the first few PCs are more meaningful / useful than the other PCs. Is this a sound assumption?

Is there more to PCA?

Exercise

What kind of information is present in this table?

Is Madrid near Warsaw?

Distances of European cities (km) from the main cities of Latium

	Rome	Latina	Frosinone	Viterbo	Rieti
Amsterdam	430	447	449	415	409
Athens	347	321	331	346	364
Barcelona	283	305	293	292	271
Beograd	227	222	236	220	238
Berlin	393	400	409	374	373
Bern	227	249	247	220	205
Bonn	353	370	372	339	330
Bruselles	388	406	406	371	365
Bucharest	364	355	368	359	378
Budapest	268	261	274	246	259
Calais	418	448	446	418	405
Copenhagen	510	522	527	492	491
Dublin	622	645	641	615	600
Edinburgh	637	655	655	625	615
Frankfurt	318	333	336	302	295
Hamburg	435	448	453	417	414
Helsinki	727	729	739	706	713
Istanbul	452	430	443	443	464
Lisbon	615	637	622	624	604
London	474	494	493	464	456
Luxembourg	325	346	346	315	307
Madrid	449	470	458	460	440
Marseille	200	223	213	202	183
Moscow	782	773	785	759	774
Munich	230	245	250	216	213
Oslo	664	675	682	646	645
Paris	365	386	383	357	343
Prague	305	313	320	286	290
Sofia	294	273	286	280	301
Stockholm	653	658	668	632	636
Warsaw	435	433	444	413	421
Vienna	255	254	265	233	240
Zurich	227	246	246	214	205

PCA of distance matrix of European cities to Italian cities

Factor loadings and proportions of explained variance

Variables	Components				
	PC1	PC2	PC3	PC4	PC5
Rome	0.9997	0.0137	−0.0184	−0.0120	0.0001
Frosinone	0.9973	−0.0715	0.0132	0.0011	0.0029
Latina	0.9987	−0.0420	−0.0272	0.0058	−0.0024
Rieti	0.9909	0.0162	0.0393	−0.0009	−0.0023
Viterbo	0.9964	0.0837	−0.0070	0.0060	0.0017
Explained variance	0.9965	0.0029	0.000569	0.000043	0.000005

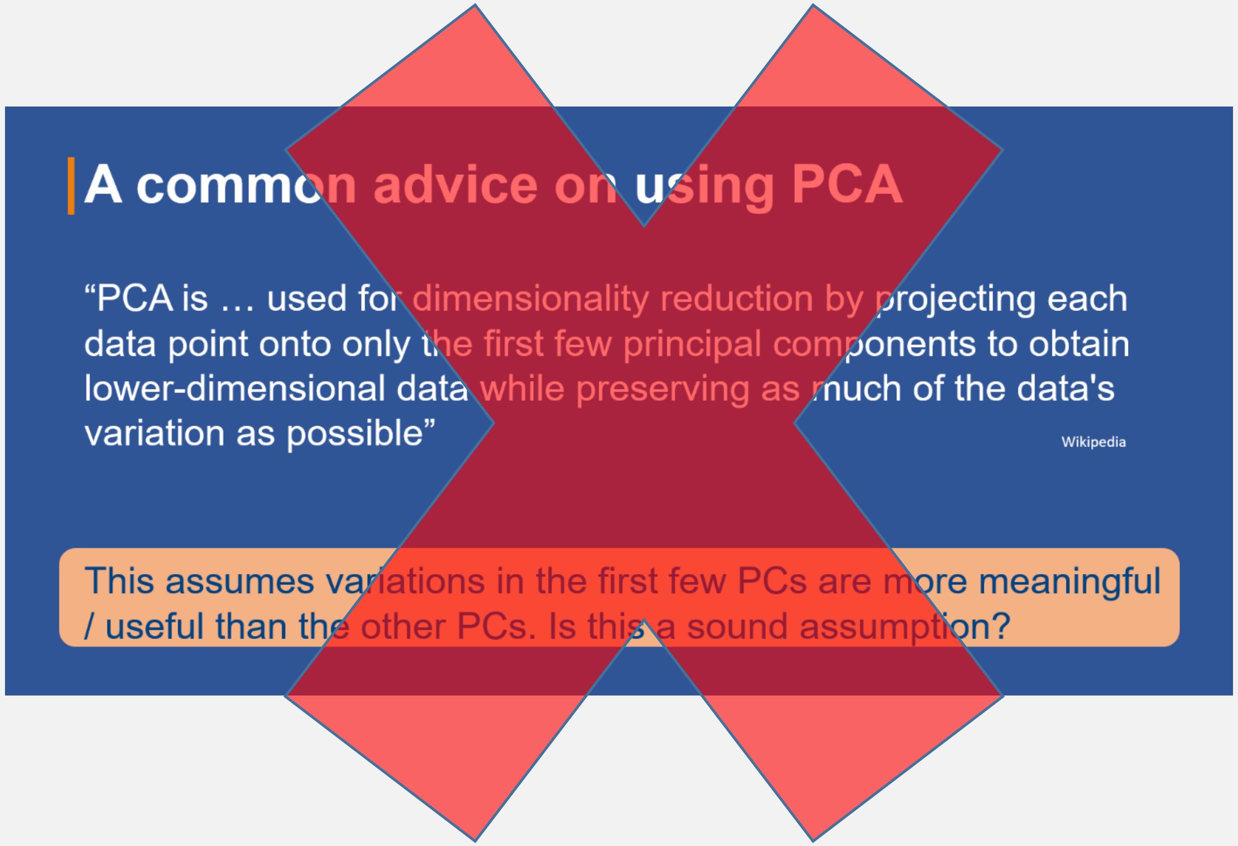
PC1 accounts for >99% of variance & correlates to distance of European cities to Latium cities

PC2, PC3, … account for < 1% of variance. What info do they correspond to? Noise?

PC2 & PC3 are …

What do you think?

PCs corresponding to < 1% of variation can be informative



Ready for a test?

Exercise

Factor loadings and proportions of explained variance

Variables	Components				
	PC1	PC2	PC3	PC4	PC5
Rome	0.9997	0.0137	−0.0184	−0.0120	0.0001
Frosinone	0.9973	−0.0715	0.0132	0.0011	0.0029
Latina	0.9987	−0.0420	−0.0272	0.0058	−0.0024
Rieti	0.9909	0.0162	0.0393	−0.0009	−0.0023
Viterbo	0.9964	0.0837	−0.0070	0.0060	0.0017
Explained variance	0.9965	0.0029	0.000569	0.000043	0.000005

PC1 ≈ distance of European cities to Latium

PC2 & PC3 ≈ angular orientation of European cities wrt Latium

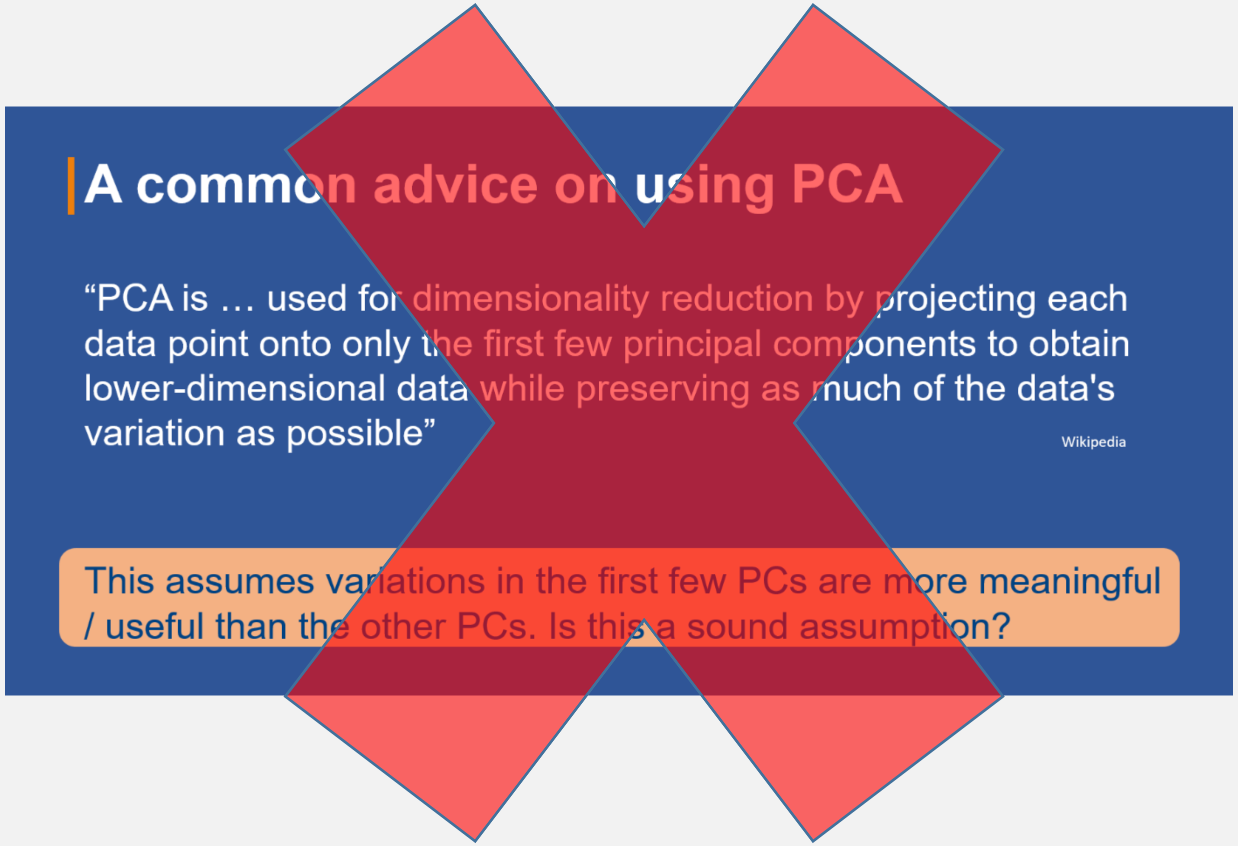
Do PC4 & 5 contain useful info? How would you know?

Variance is de-convoluted into real factors by PCA

PCs that don’t correspond to real factors are thus Gaussian-like residual noise

What do you think?

Top PCs can correspond to irrelevant or confounding information



Learning points

PCA deconvolutes data variations into meaningful direction

PCs corresponding to <1% of data variations can be meaningful

Top PCs can correspond to irrelevant or confounding info

You can tell which PCs are noise

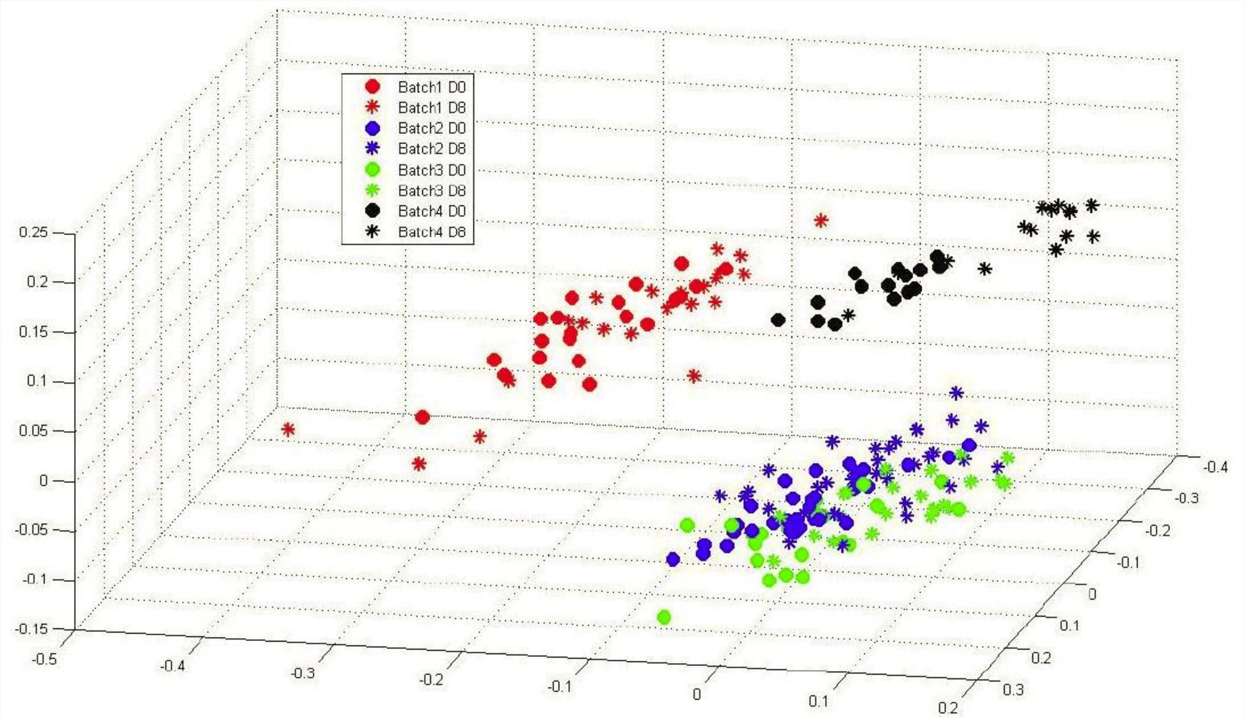
A little bit of logical thought goes a very long way

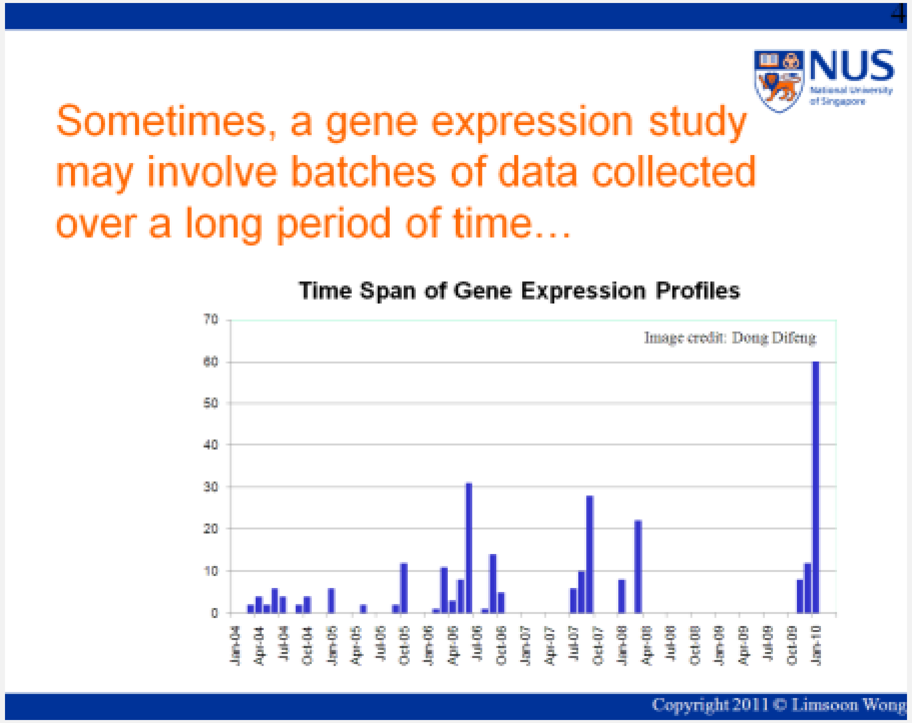
References

Giuliani et al., "On the constructive role of noise in spatial systems", *Physics Letters A*, 247:47-52, 1998

Batch effects & PCA

PCA scatter plot of patients' omics profiles





Samples from different batches are grouped together, regardless of subtypes and treatment response

Image credit: Difeng Dong's PhD dissertation, 2011

Batch effects

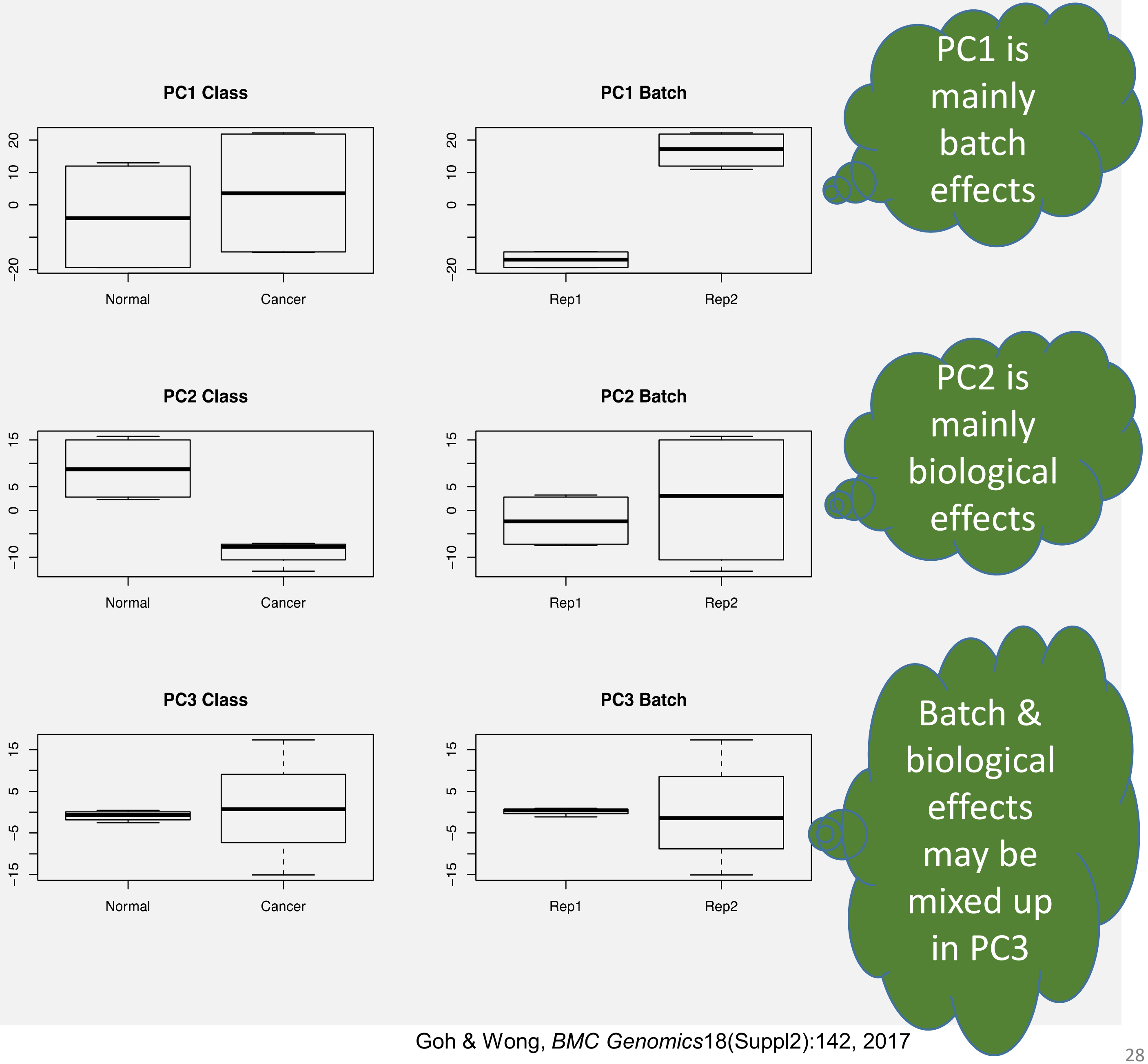
Batch effects are unwanted sources of variation caused by different processing date, handling personnel, reagent lots, equipment, etc.

Batch effects are a big challenge faced in biological research, especially towards translational research and precision medicine

How do you know which PCs are dominated by batch effects?

Paired boxplots of PCs

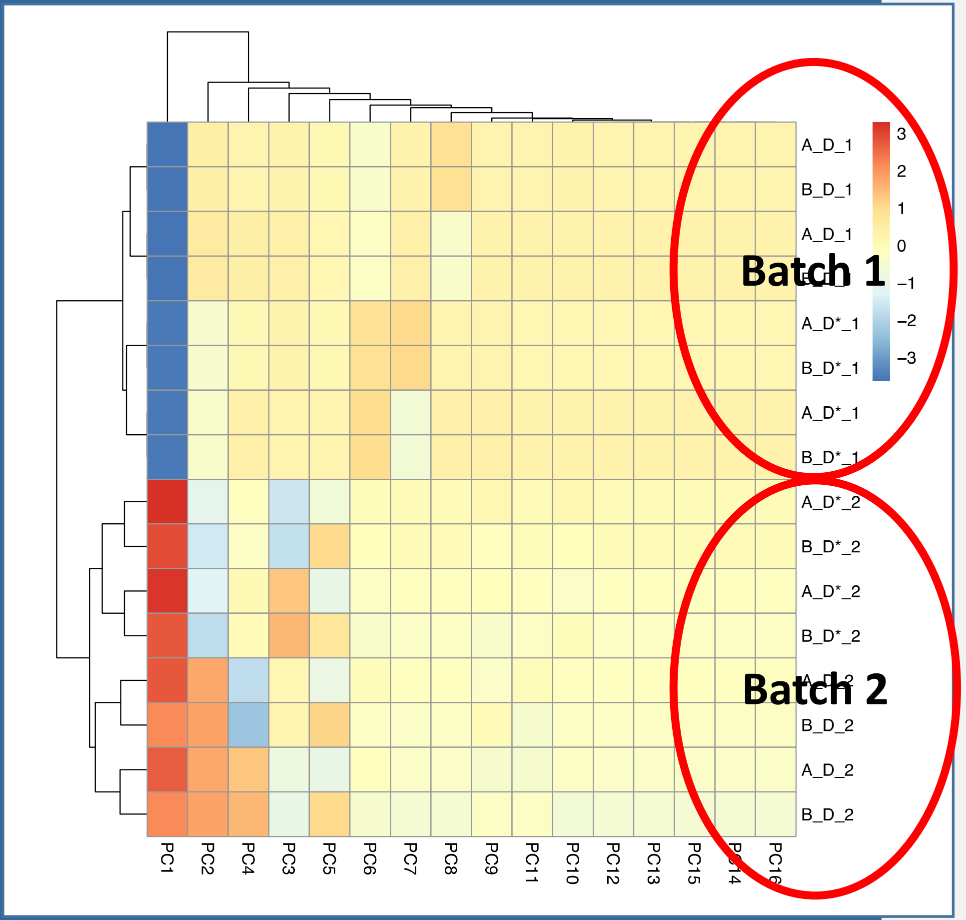
See which PC is enriched in batch effects by showing, side by side, distribution of values of each PC stratified by class and suspected batch variables



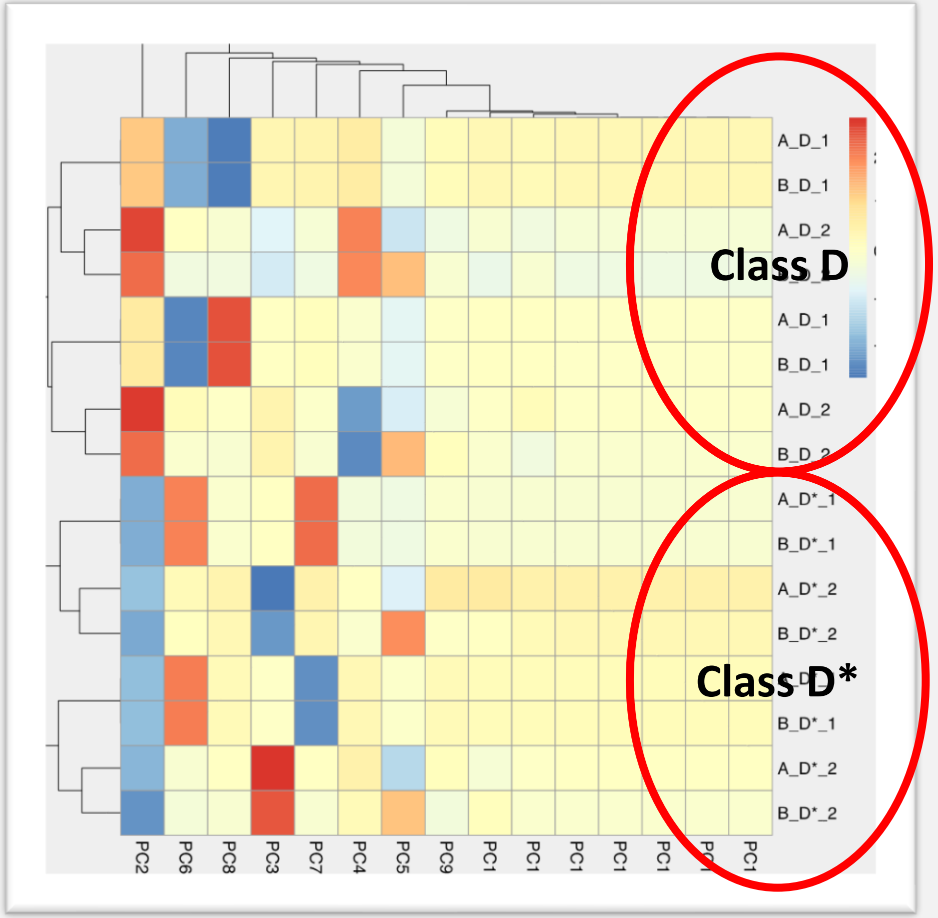
Goh & Wong, *BMC Genomics*18(Suppl2):142, 2017

Remove batch effects-laden PCs

(Notation: A/B_D/D*_1/2 refers to the dataset, class and batches respectively)



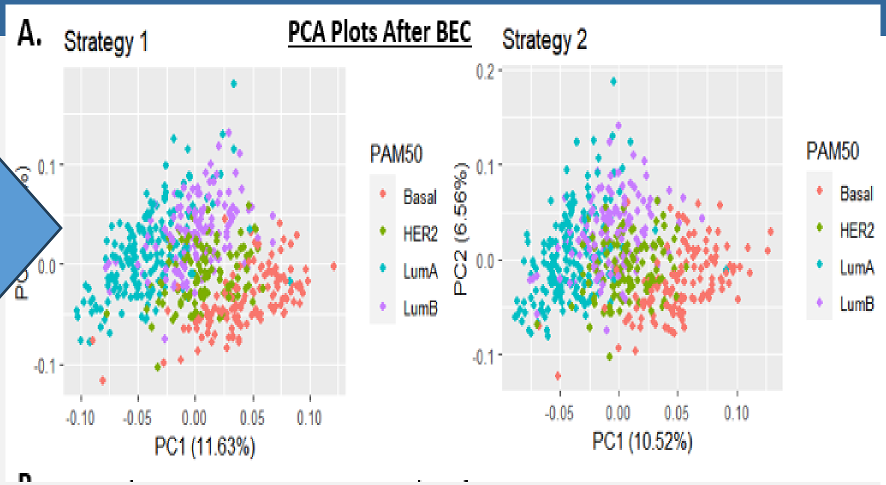
Batch effects dominate

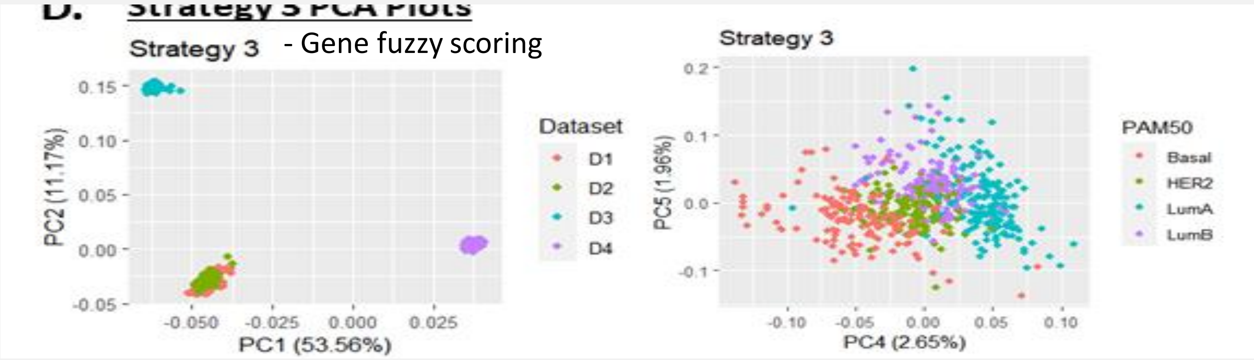


Class-effect discrimination recovered

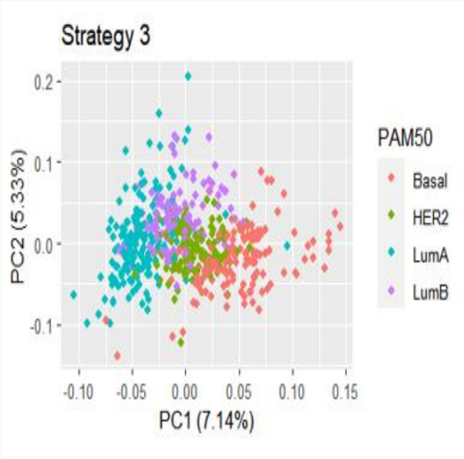
From Senuri’s PQE report

combat

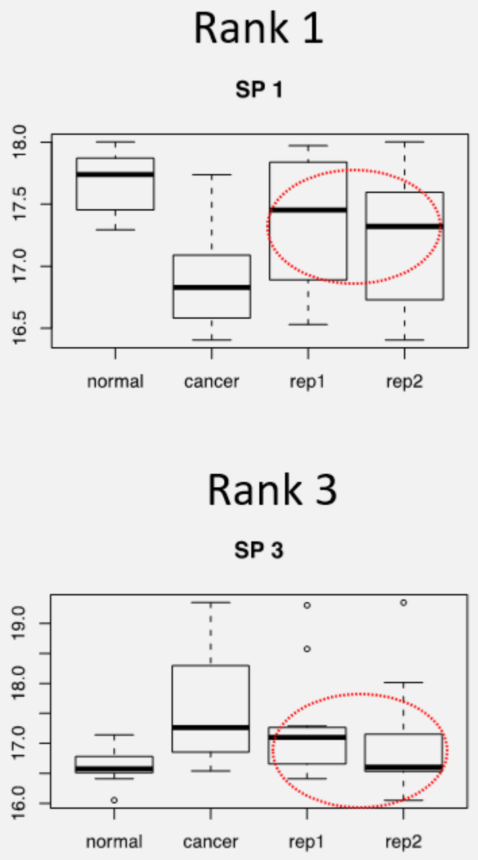




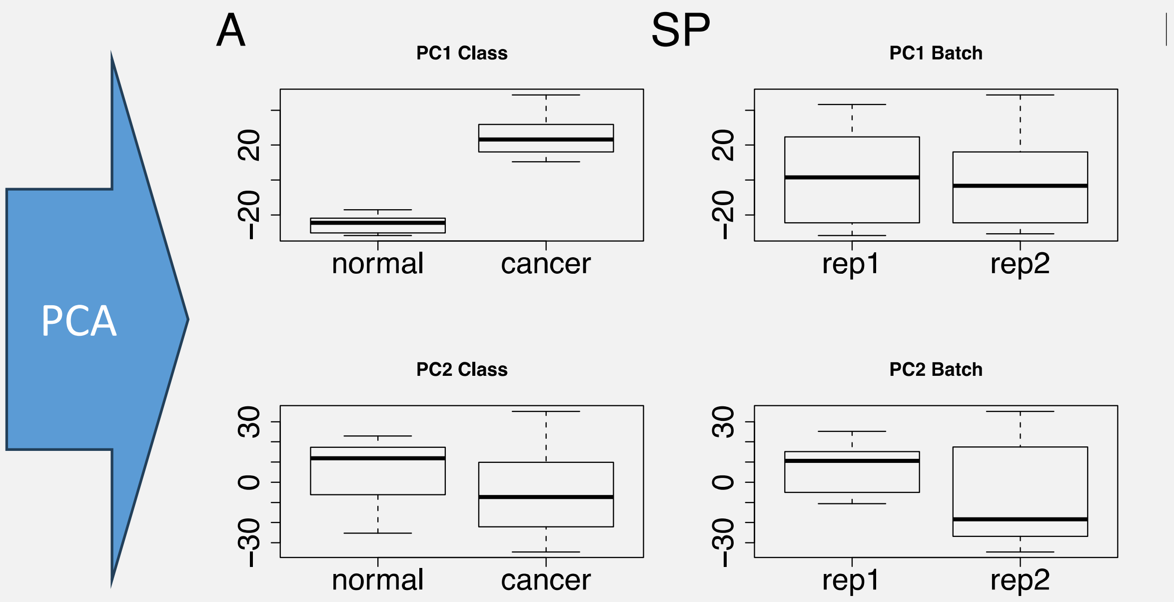
Drop
PC 1-3 &
invert



What if class and batch effects are confounded at the protein level?



Batch correction might inadvertently lose information on disease subpopulations (look like batch effects but are meaningful)



References

Goh & Wong, "Protein complex-based analysis is resistant to the obfuscating consequences of batch effects---a case study in clinical proteomics", *BMC Genomics*, 18(Suppl 2):142, 2017

Homework, due 24/1/2026

submit by email to dcswls@nus.edu.sg

Read this paper, focusing on Section 3.1, Segmentation by eigenbackground subtraction:

Oliver et al., "A Bayesian computer vision system for modeling human interactions", ***IEEE Transactions on Pattern Analysis and Machine Intelligence*****, 22(8):831-843, 2000**

Read the review report (attached below) produced by Gemini on this section.

Answer the following questions:

1/ Which parts of the Gemini report, if any, do you agree with?

2/ Which parts of the Gemini report, if any, do you disagree with? Why?

3/ Are there any major flaws that you think should be pointed out but are missed by Gemini?

Make 5-minutes presentation to class on 26/1/2026

Assessment of Section 3.1: Segmentation by Eigenbackground Subtraction

1. The Segmentation Method

Section 3.1 describes a background subtraction technique using Principal Component Analysis (PCA), forming an Eigenbackground model:

- Model Generation: The eigenspace is built from N sample images, computing the mean (μ_b) and covariance matrix (C_b). The top M eigenvectors (Φ_M) are retained to model the expected range of background appearances, which the authors claim includes "lighting variations over the day, weather variations, etc.".
- Foreground Detection: Moving objects are detected as outliers to this model because they "do not have a significant contribution" to the background eigenspace. Detection is achieved by calculating the Euclidean distance (Distance From Feature Space, DFFS) between the input image (I_i) and its reconstruction (B_i).
- Final Segmentation: The thresholded distance ($D_i > t$) produces a motion mask. This mask is then fed to a connected component algorithm to generate "blob descriptions" that characterize each person's shape.

2. Is the Idea Sound?

Yes, the Eigenbackground technique is a valid, historically significant approach. Mathematically, it correctly identifies moving objects by the high reconstruction error they produce when projected onto a background-trained eigenspace. The authors explicitly state they opted for this method due to its balance of "good results and less computational load" compared to alternatives like Mixture of Gaussians.

3. Flaws and Limitations

The method, while sound and efficient, suffers from significant practical and methodological constraints:

1. **Model Adaptation Ambiguity:** The paper claims it's "easy to **adaptively perform**" the subtraction but **does not describe an online update mechanism** for the PCA model itself. Without an explicit update loop, the static model cannot genuinely handle long-term changes or new stationary objects without full retraining.
2. **Critical Unjustified Assumptions:** The training process relies on the **strong, unstated assumption** that the N sample images contain *only* background. Any moving object present in the training set will be incorrectly learned as part of the background model.
3. **Lack of Principled Thresholding:** The paper **omits any discussion** of how the critical detection threshold (t) is chosen or tuned, which is a significant methodological omission.
4. **Sensitivity to Jitter:** As a **global** model, the eigenbackground is highly susceptible to **camera jitter or small misalignments**, causing spurious reconstruction errors across the entire frame.
5. **Context Misalignment:** While authors claim handling of "lighting... weather, etc.," the method's global nature and limited eigenvector capacity are **less suited to complex localized scene dynamics** than the paper's claims might suggest, particularly when applied outside of their static-background domain.