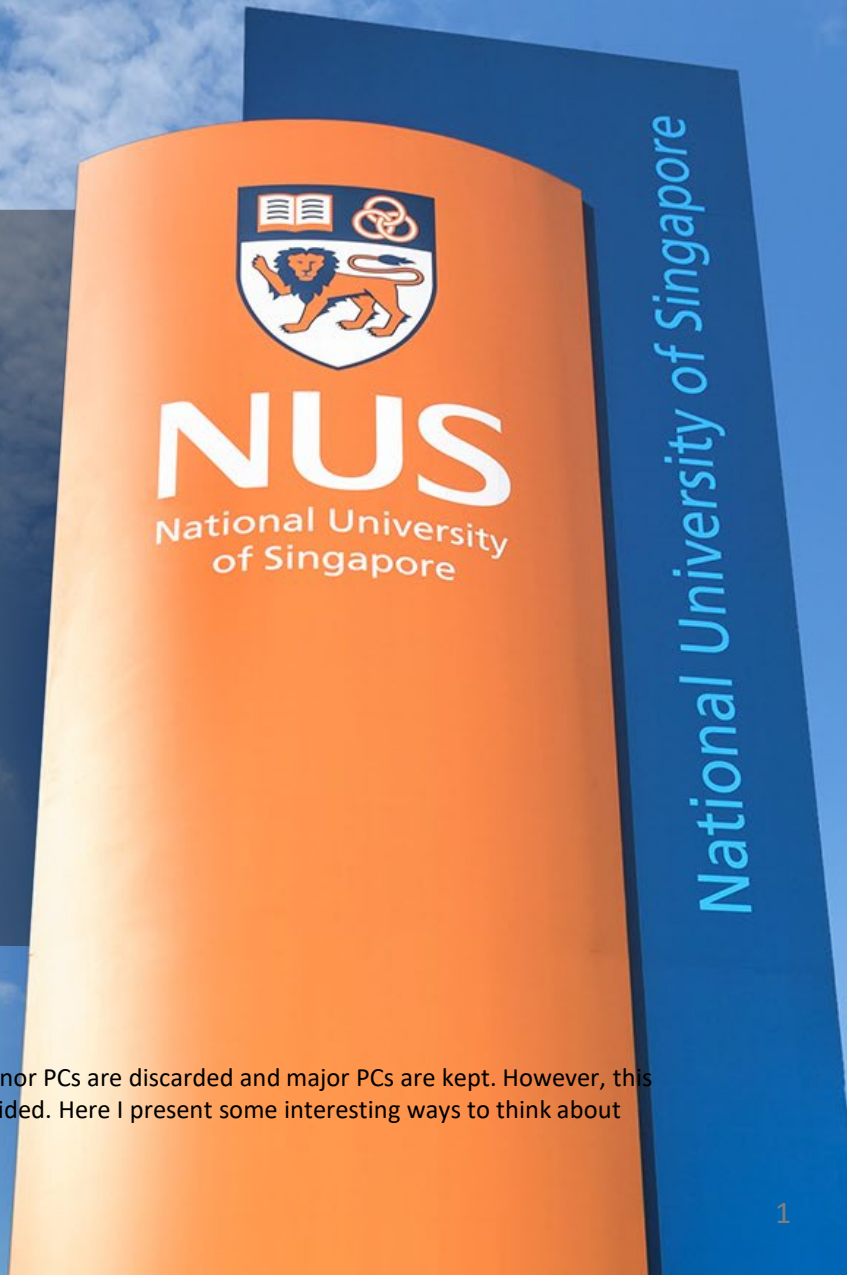
Art of principal component analysis

Wong Limsoon

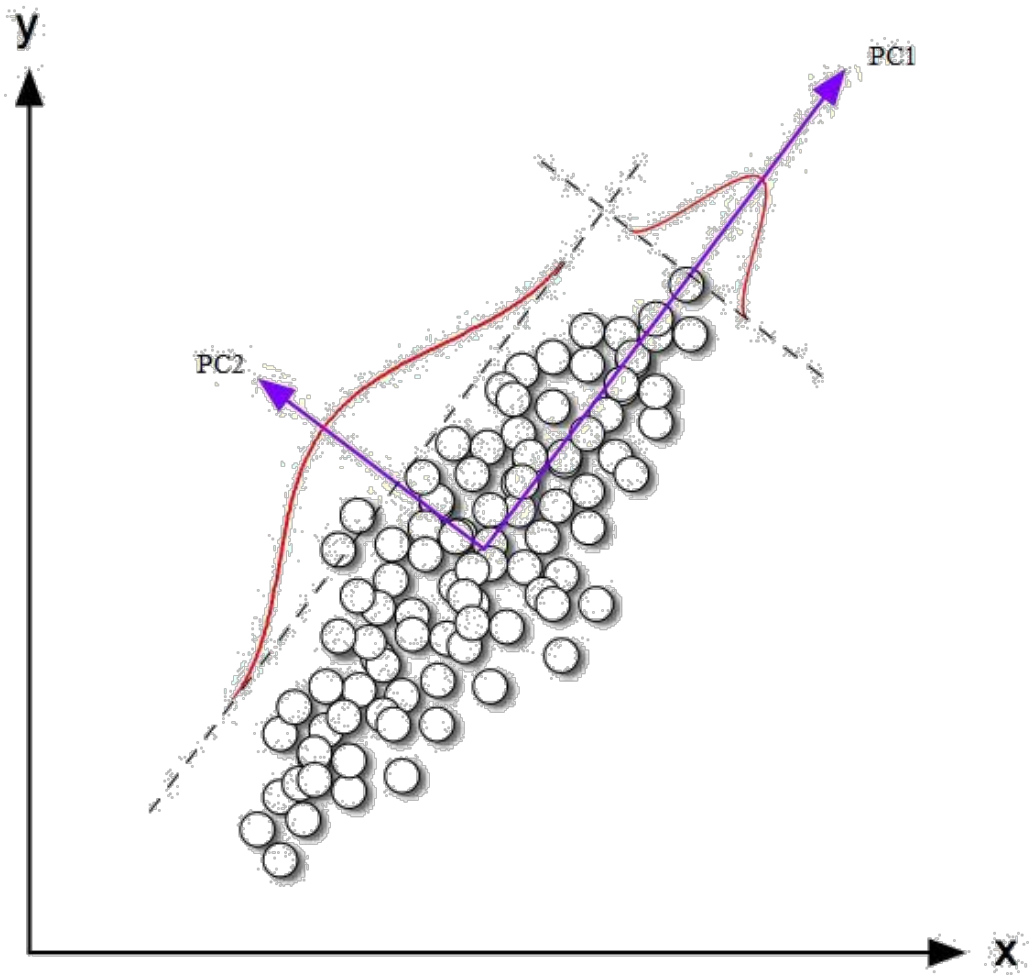


Outline: PCA is quite often thought of as a means for feature selection whereby minor PCs are discarded and major PCs are kept. However, this mechanical way of using PCA is often less than optimal and sometimes even misguided. Here I present some interesting ways to think about and use PCA.

© Copyright National University of Singapore. All Rights Reserved.

Principal component analysis (PCA)

PCA, intuitively



Credit: Alessandro Giuliani

PCA, mathematically

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

Inverting

$X_{inv} = Z V^T + \mu$

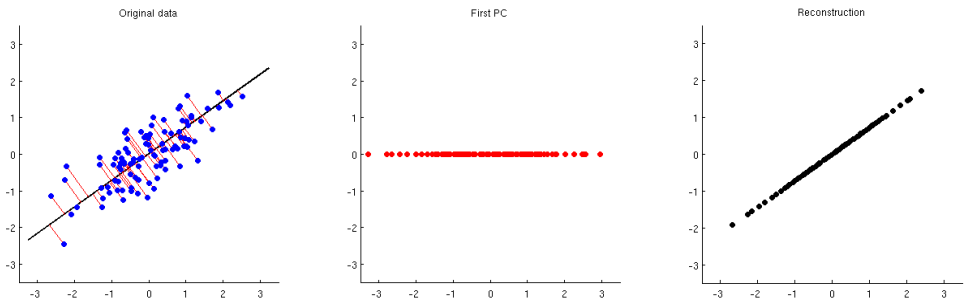


Image credit: https://stats.stackexchange.com/questions/229092/how-to-reverse-pca-and-reconstruct-original-variables-from-several-principal-com

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)

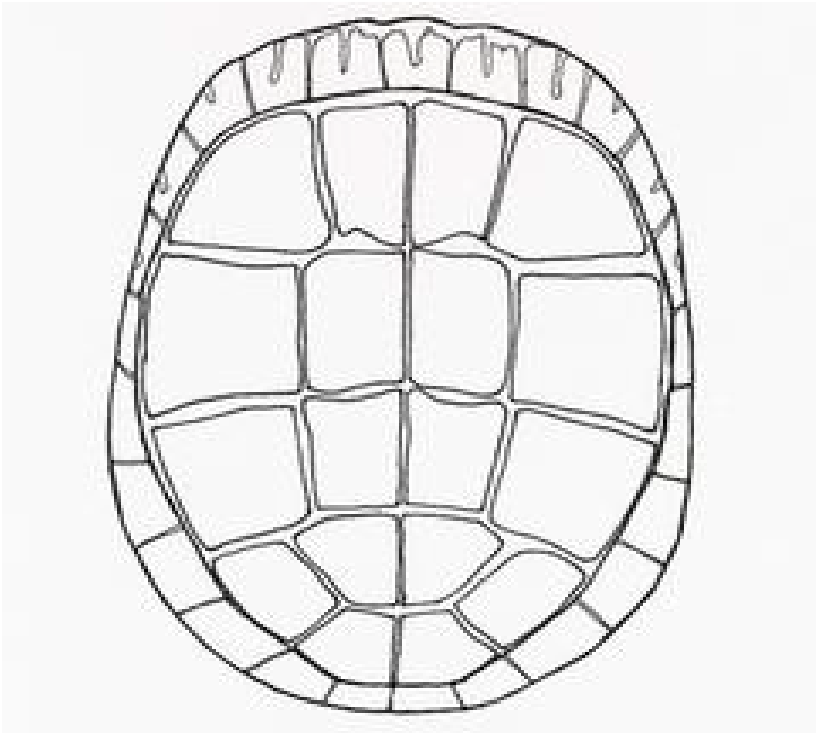
Credit: Alessandro Giuliani

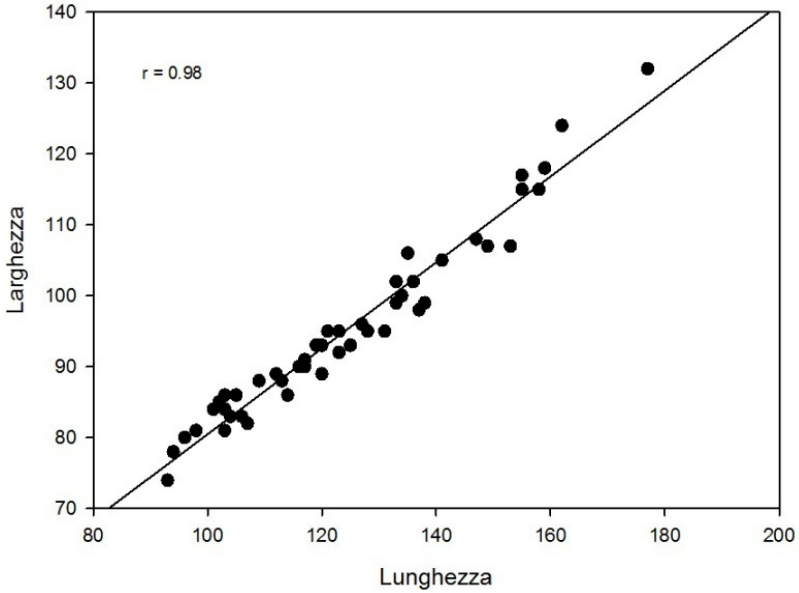
TABLE 1

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





Pearson Correlation Coefficients,

	length	width	height
length	1.00000	0.97831	0.96469
width	0.97831	1.00000	0.96057
height	0.96469	0.96057	1.00000

Width = 19,94 + 0,605*Length

Credit: Alessandro Giuliani

Credit: Alessandro Giuliani

Principal components

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

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

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

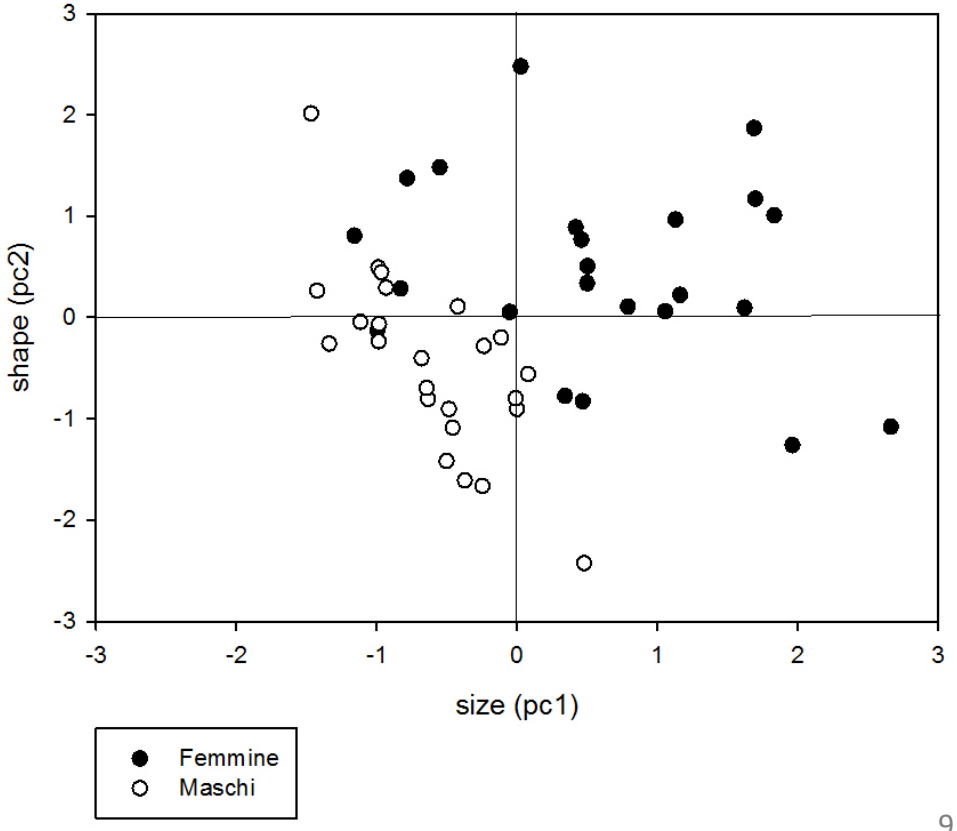
Presence of an overwhelming size component explaining system variance comes from the presence of a 'typical' common shape

Displacement along pc1 = size variation (all positive terms)

Displacement along pc2 = shape deformation (both positive and negative terms)

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

Exercise #1

Madrid and Warsaw are at almost the same distance to Latium cities

Are Madrid and Warsaw near each other?

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 Latium 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

Are PC2, PC3, … useless / non-informative?

PC2 & PC3 are …

So, you can tell Madrid is not near Warsaw

Intentionally left blank

Intentionally left blank

Intuitive points

PCA gives the axes that orthogonally account for variance in the data

PCs correspond to explanations / factors giving rise to the variance

Coefficient of a variable in a PC suggests how relevant that variable is for that PC

Surprising point

PCs accounting for a very small portion of the variance can also be informative, if you know how to find these

Exercise #2

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 - 99% variance & corresponds to distance

PC2, 3, 4, 5, … - <1% variance

Are PC2, 3, 4, 5, … useless? How do you know?

Variance is deconvoluted to real factors by PCA

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

Intentionally left blank

Pre-whitening to distinguish informative vs uninformative PCs

Intentionally left blank

PCA in biomarker selection

PCA in biomarker selection

When PCA is applied on gene expression data

PCs w/ large variance $\approx$ diff expressed pathways

Variables w/ large coefficient/loading in a PC $\approx$ key genes in the pathway associated w/ that PC

PCA can be a useful biomarker-selection approach

E.g., biomarkers $\approx$ genes w/ high loading

Loading of gene x = $\Sigma_j | \alpha_{xj} * \sigma_j^2 |$, where α_{xj} is coefficient of x in PC_j, and σ_j^2 is variance of PC_j

Acute lymphoblastic leukemia

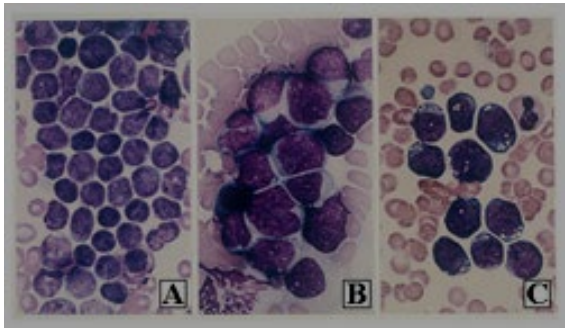
Major ALL subtypes

T-ALL, E2A-PBX,

TEL-AML, BCR-ABL,

MLL genom rearranged, Hyperdiploid>50

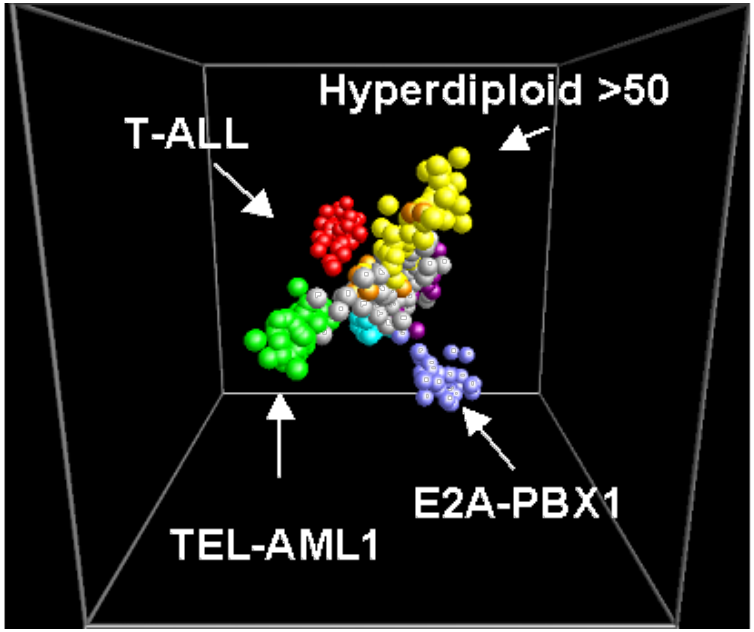
The subtypes look similar

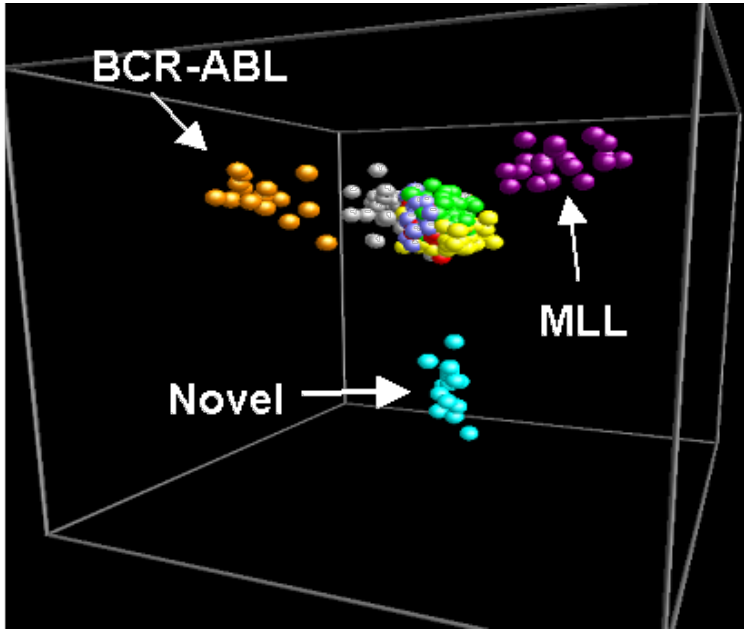


Diff subtypes respond differently to treatment

Can we diagnosis the subtypes based on gene expression profiling?

PCA in ALL subtype diagnosis





Identify genes with high variance

Perform PCA on them

Plot using PC1 to 3

Induction of hypothesis

PCs capture diff bio pathways

Values of PCs capture diff states of pathways

Hypothesis: If patient X has ALL subtype T, X's bio pathways are in state S_T

… and abduction during diagnosis

Observation: John's bio pathways are in state S_T

Abduction: John has ALL subtype T

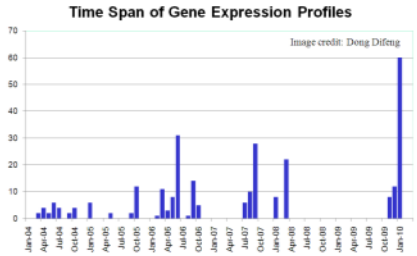
Batch effects

What are batch effects?

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

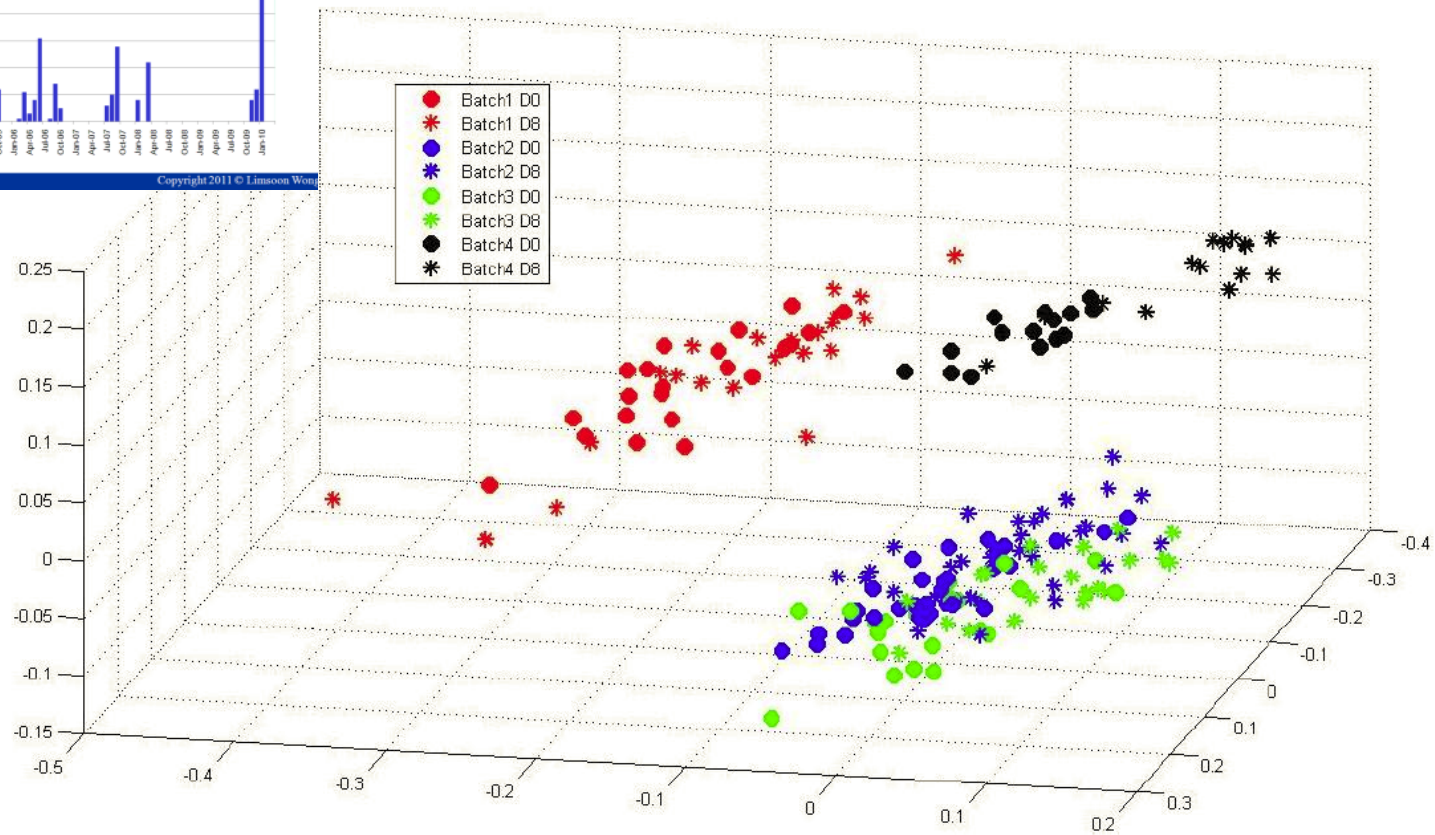
Batch effects is a big challenge faced in biological research, especially towards translational research and precision medicine

Sometimes, a gene expression study may involve batches of data collected over a long period of time…



Copyright 2011 © Limsoon Wong

PCA scatter plot



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

Image credit: Difeng Dong's PhD dissertation, 2011

Exercise #3

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

Paired boxplots of PCs

Intentionally left blank

Exercise #4

PCA can be used to detect presence of batch effects

Can you use PCA to remove batch effects?

If so, how?

PCA, mathematically

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

Inverting

$X_{inv} = Z V^T + \mu$

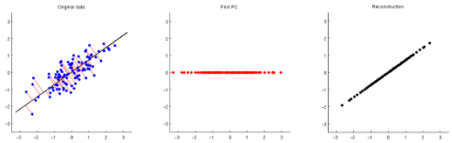


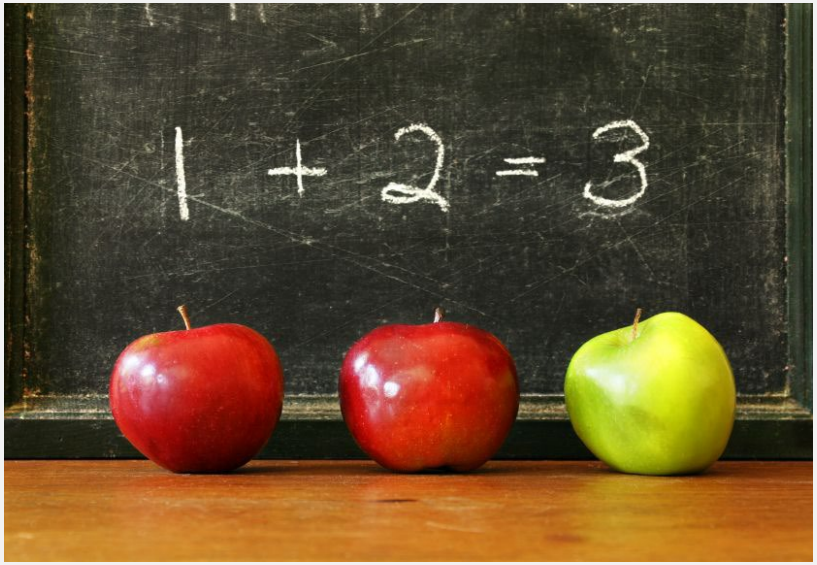
Image credit: https://stats.stackexchange.com/questions/229092/how-to-reverse-pca-and-reconstruct-original-variables-from-several-principal-com

Determine …

Intentionally left blank

Remove …

Intentionally left blank


1 + 2 = 3

What have we learned?

PCA is a useful paradigm

Low-loading PCs can be informative

“pre-whitening” can be used to identify informative PCs

PCA is not just a visualization tool; it can also be used for dealing with batch effects

High-loading PCs can be batch effects and need to be discarded

References

[PCA] Jolicoeur & Mosimann, *Growth*, 24:339-354, 1960

[PCA] Giuliani et al., *Physics Letters A*, 247:47-52, 1998

[Batch effects] Leek et al., *Nature Reviews Genetics*, 11:733-739, 2010

[Batch effects] Wang et al., *Molecular Biosystems*, 8:818-827, 2012

[FSNET] Goh & Wong, *BMC Genomics*, 18(Suppl 2):142, 2017