MCI5004: Molecular Biomarkers in Clinical Research

Principal Component Analysis in Biomarker Discovery

Wong Limsoon

NUS National University of Singapore

Plan

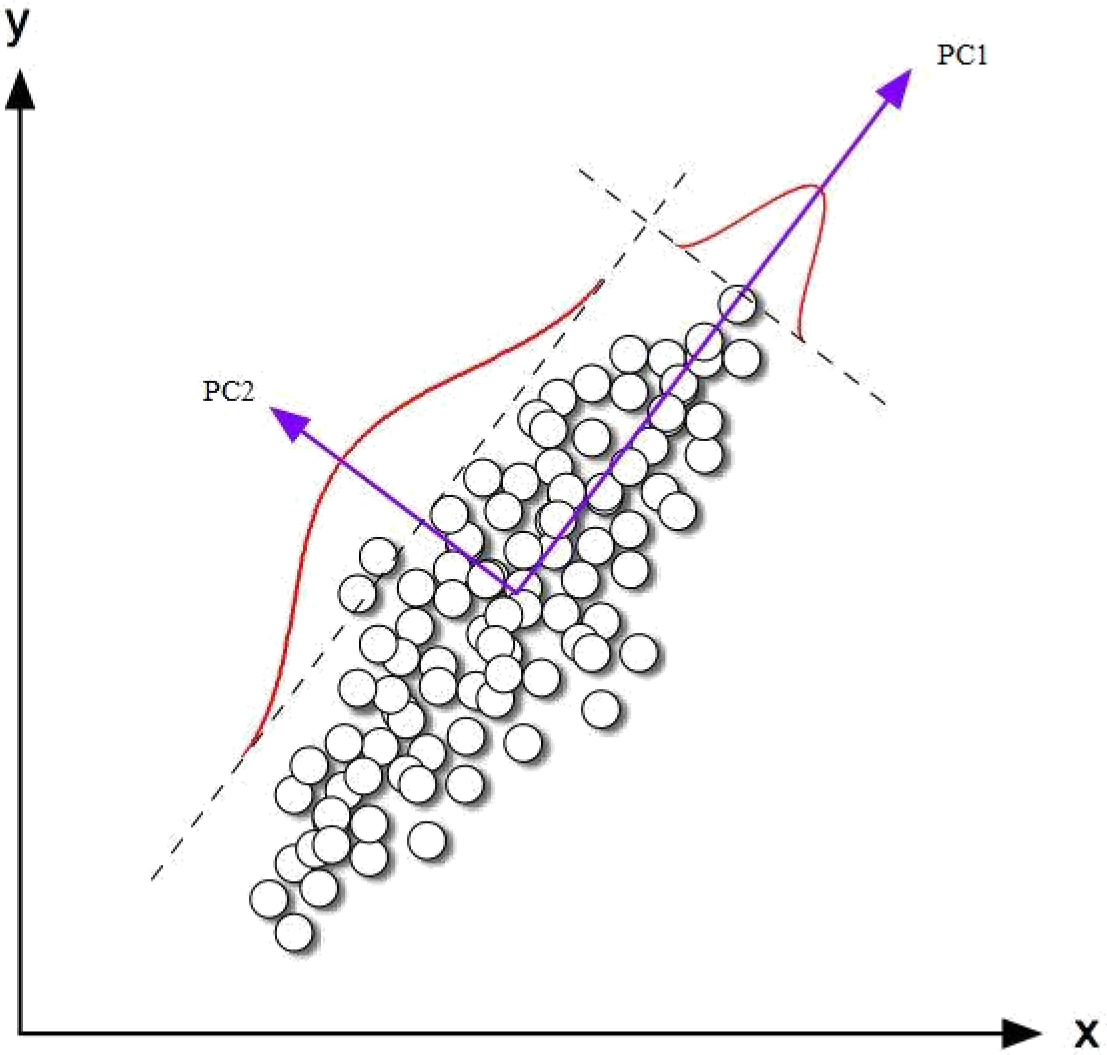
- **PCA**
- **PCA in biomarker selection**

- **Batch effects**
- **PCA for isolating batch effects**

- **PCA at the level of protein complexes / biological pathway subnetworks**

PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA, intuitively



Credit: Alessandro Giuliani

https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction

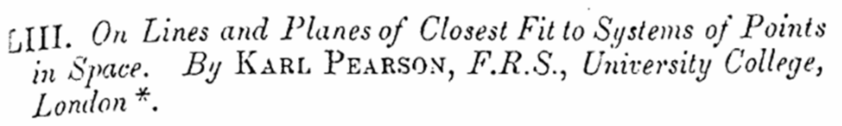
PCA, a la Pearson (1901)

)(98)(

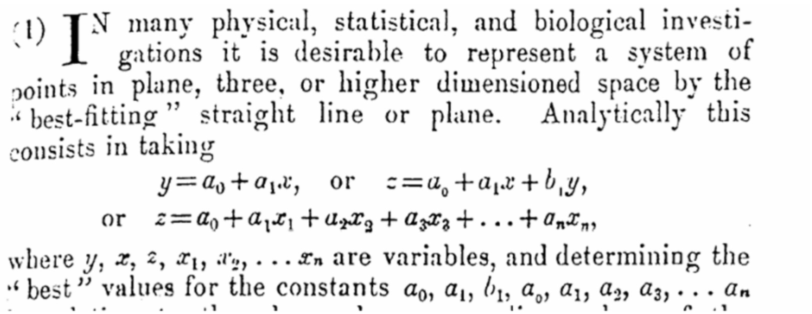
SULLE FUNZIONI BILINEARI

DI

E. BELTRAMI



LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, *F.R.S., University College, London* *.

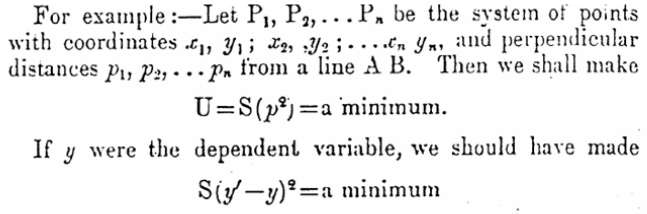


(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this consists in taking

$$y=a_0+a_1x, \quad \text{or} \quad z=a_0+a_1x+b_1y,$$

$$\text{or} \quad z=a_0+a_1x_1+a_2x_2+a_3x_3+\ldots+a_nx_n,$$

where y, x, z, x_1, x_2, . . . x_n are variables, and determining the "best" values for the constants a_0, a_1, b_1, a_0, a_1, a_2, a_3, . . . a_n

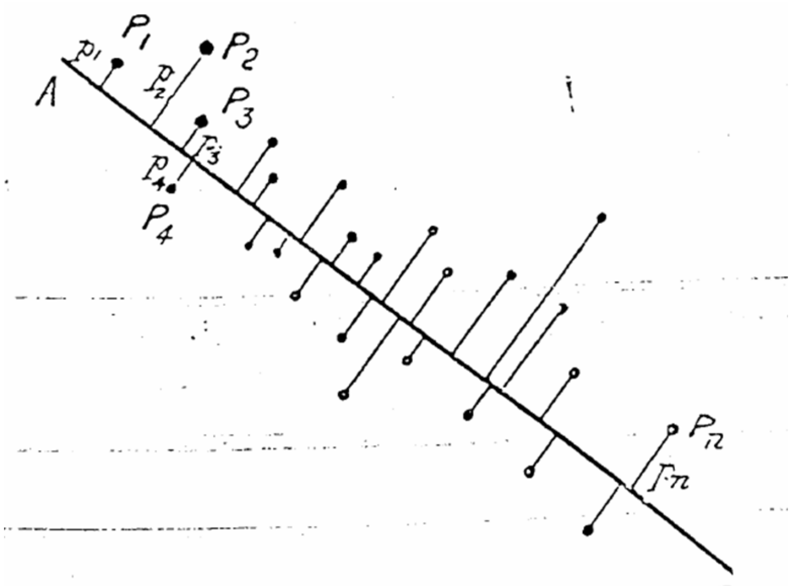


For example:—Let $P_1, P_2, \ldots P_n$ be the system of points with coordinates x_1, y_1; x_2, y_2; . . . x_n y_n, and perpendicular distances $p_1, p_2, \ldots p_n$ from a line A B. Then we shall make

$$U=S(p^2)=\text{a minimum.}$$

If y were the dependent variable, we should have made

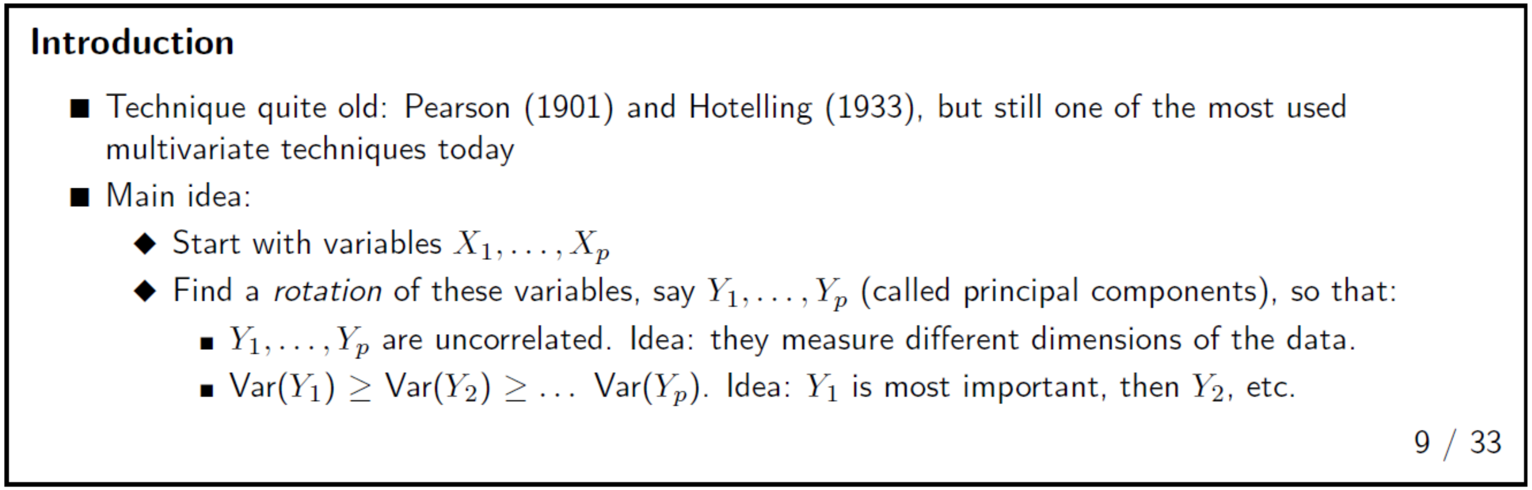
$$S(y'-y)^2=\text{a minimum}$$



Credit: Alessandro Giuliani

Credit: Marloes Maathuis

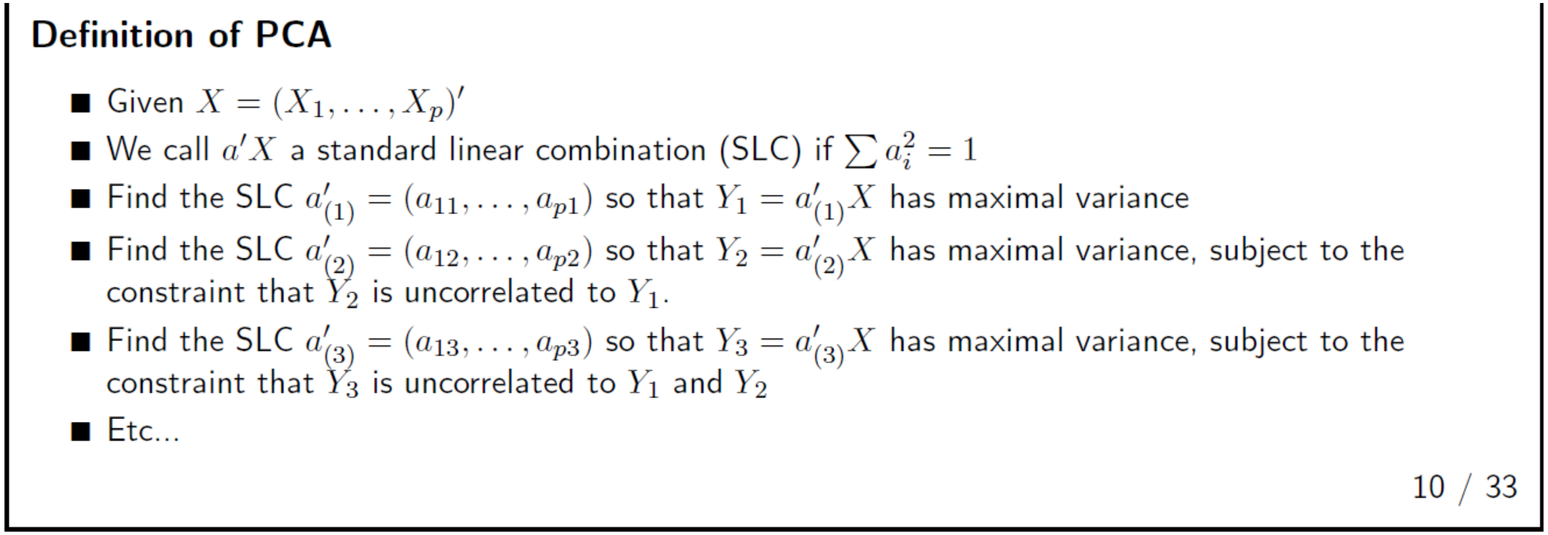
PCA, in modern English ☺



Introduction

- Technique quite old: Pearson (1901) and Hotelling (1933), but still one of the most used multivariate techniques today
- Main idea:
  - Start with variables $X_1, \ldots, X_p$
  - Find a *rotation* of these variables, say $Y_1, \ldots, Y_p$ (called principal components), so that:
    - $Y_1, \ldots, Y_p$ are uncorrelated. Idea: they measure different dimensions of the data.
    - $\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq \ldots \text{Var}(Y_p)$. Idea: Y_1 is most important, then Y_2, etc.

9 / 33



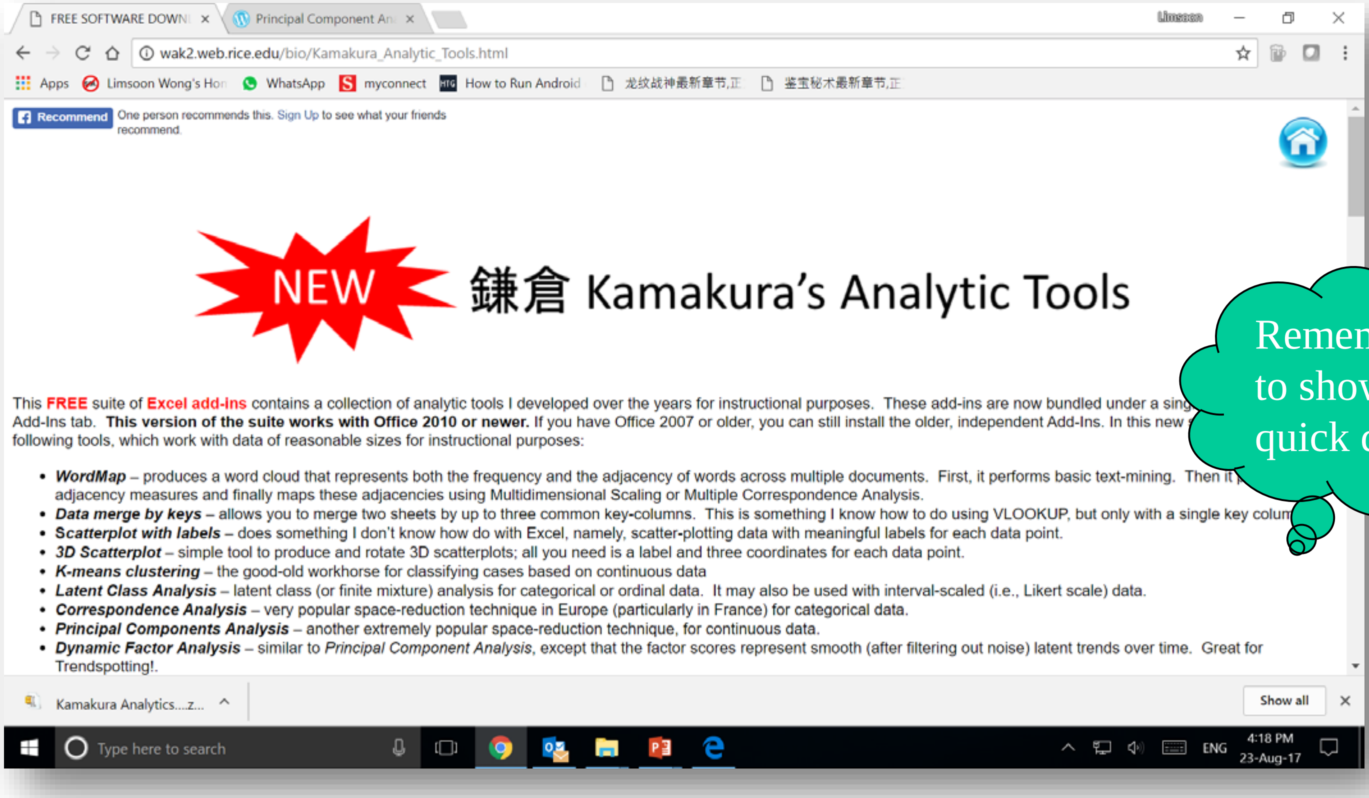
Definition of PCA

- Given $X = (X_1, \ldots, X_p)'$
- We call $a'X$ a standard linear combination (SLC) if $\sum a_i^2 = 1$
- Find the SLC $a'_{(1)} = (a_{11}, \ldots, a_{p1})$ so that $Y_1 = a'_{(1)}X$ has maximal variance
- Find the SLC $a'_{(2)} = (a_{12}, \ldots, a_{p2})$ so that $Y_2 = a'_{(2)}X$ has maximal variance, subject to the constraint that Y_2 is uncorrelated to Y_1.
- Find the SLC $a'_{(3)} = (a_{13}, \ldots, a_{p3})$ so that $Y_3 = a'_{(3)}X$ has maximal variance, subject to the constraint that Y_3 is uncorrelated to Y_1 and Y_2
- Etc...

10 / 33

Nice free Excel add-on

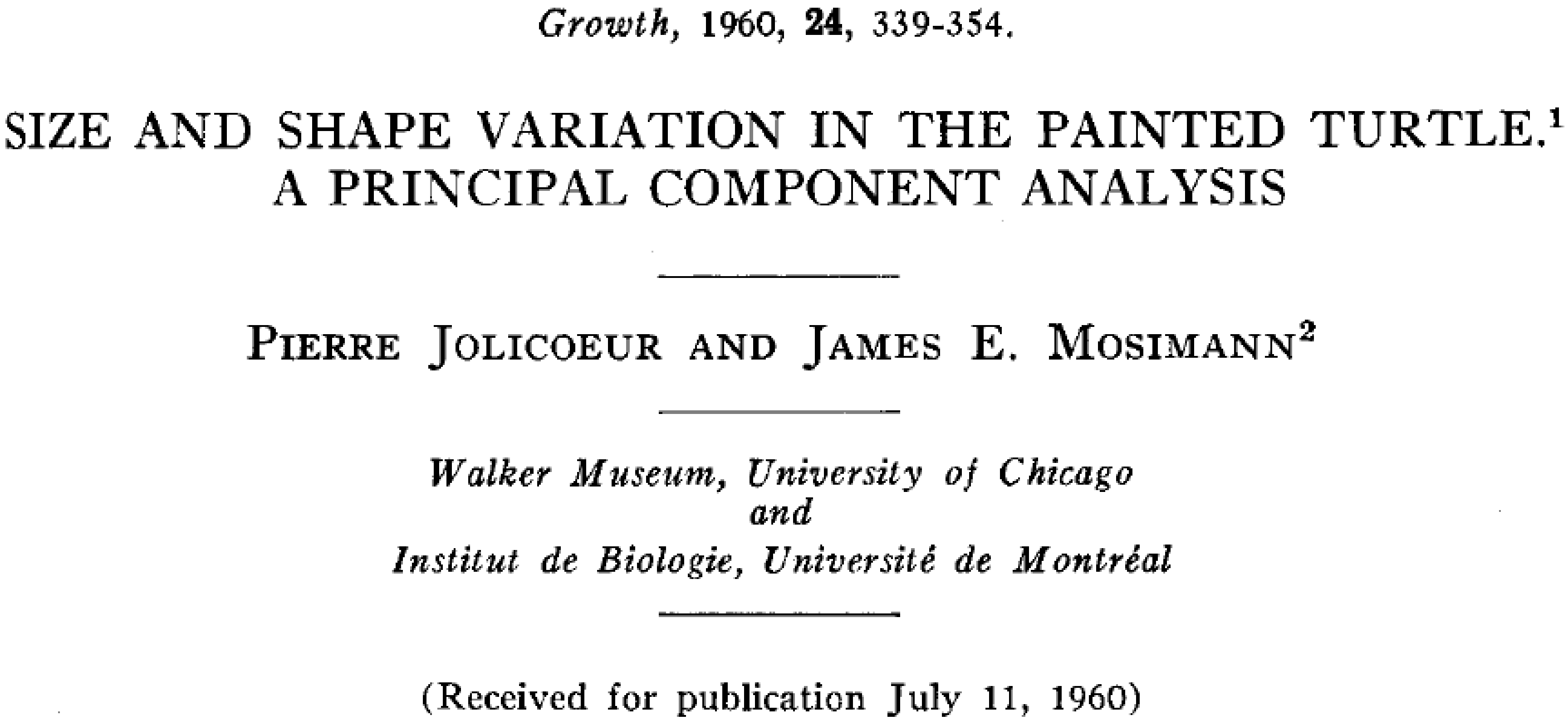
- http://wak2.web.rice.edu/bio/Kamakura_Analytic_Tools.html



This **FREE** suite of **Excel add-Ins** contains a collection of analytic tools I developed over the years for instructional purposes. These add-ins are now bundled under a sing Add-Ins tab. **This version of the suite works with Office 2010 or newer.** If you have Office 2007 or older, you can still install the older, independent Add-Ins. In this new s following tools, which work with data of reasonable sizes for instructional purposes:

- ***WordMap*** – produces a word cloud that represents both the frequency and the adjacency of words across multiple documents. First, it performs basic text-mining. Then it p adjacency measures and finally maps these adjacencies using Multidimensional Scaling or Multiple Correspondence Analysis.
- ***Data merge by keys*** – allows you to merge two sheets by up to three common key-columns. This is something I know how to do using VLOOKUP, but only with a single key colum
- ***Scatterplot with labels*** – does something I don't know how do with Excel, namely, scatter-plotting data with meaningful labels for each data point.
- ***3D Scatterplot*** – simple tool to produce and rotate 3D scatterplots; all you need is a label and three coordinates for each data point.
- ***K-means clustering*** – the good-old workhorse for classifying cases based on continuous data
- ***Latent Class Analysis*** – latent class (or finite mixture) analysis for categorical or ordinal data. It may also be used with interval-scaled (i.e., Likert scale) data.
- ***Correspondence Analysis*** – very popular space-reduction technique in Europe (particularly in France) for categorical data.
- ***Principal Components Analysis*** – another extremely popular space-reduction technique, for continuous data.
- ***Dynamic Factor Analysis*** – similar to *Principal Component Analysis*, except that the factor scores represent smooth (after filtering out noise) latent trends over time. Great for Trendspotting!.

Remember to show quick demo



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

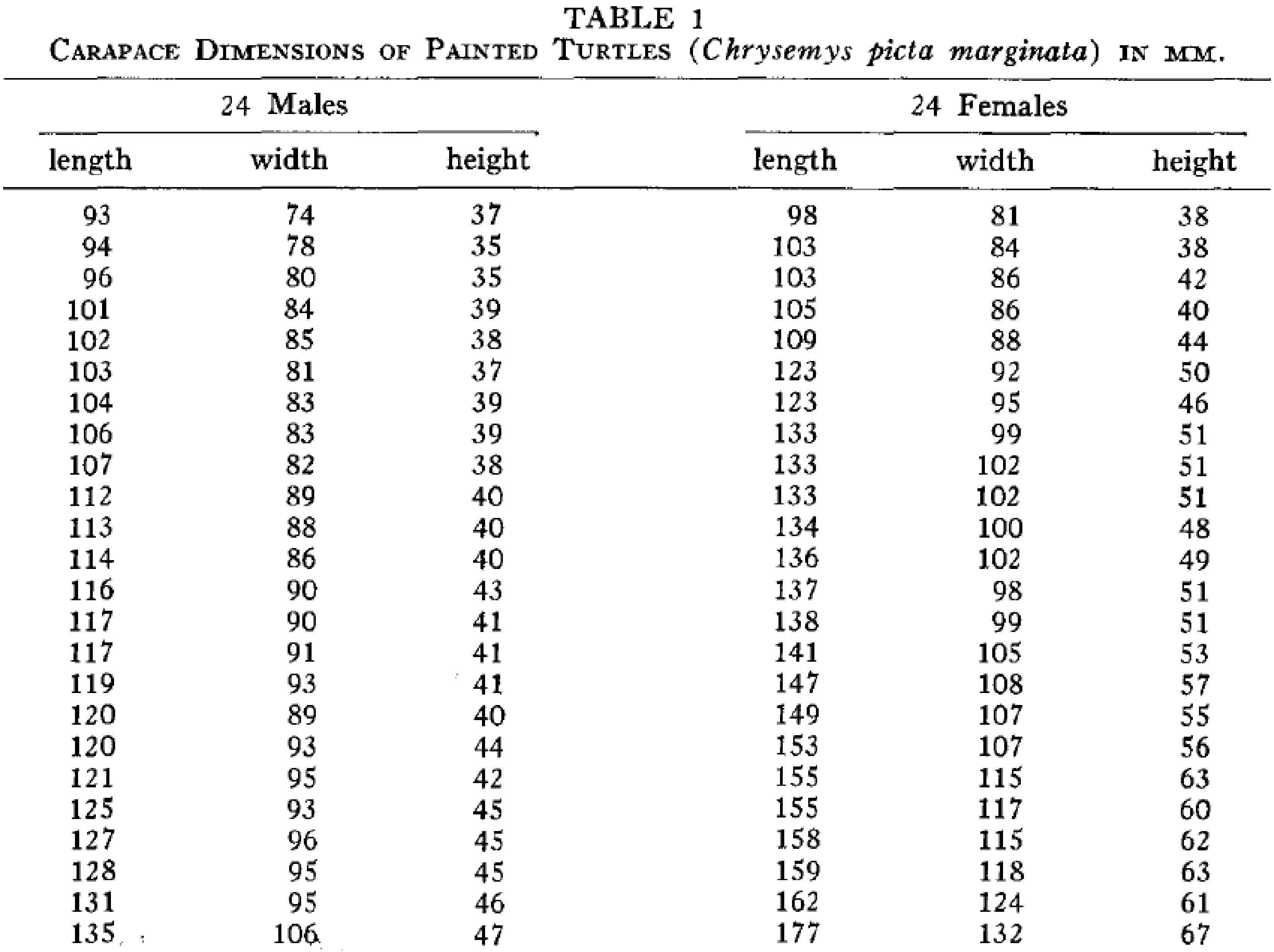
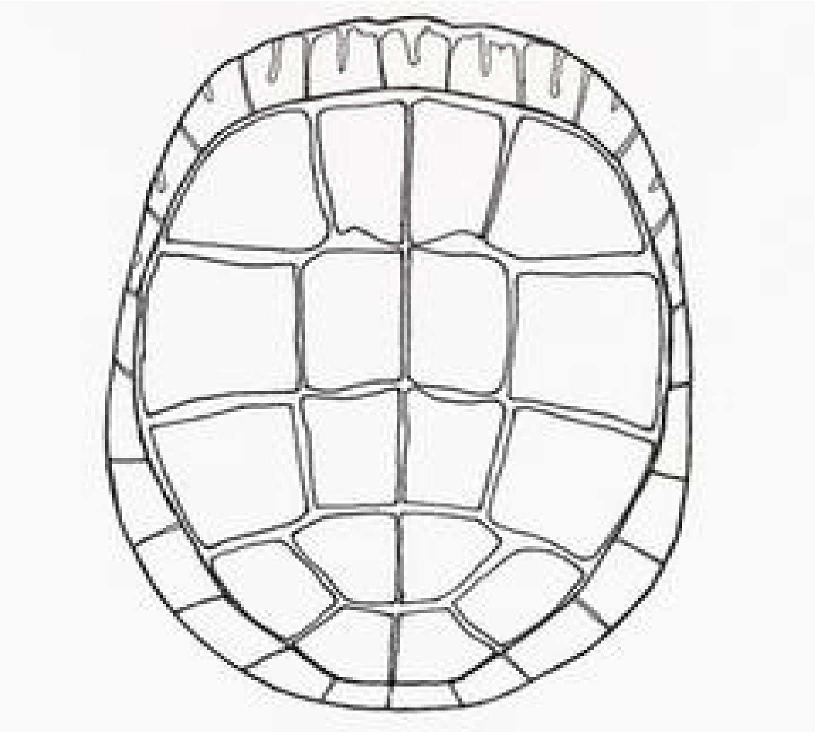


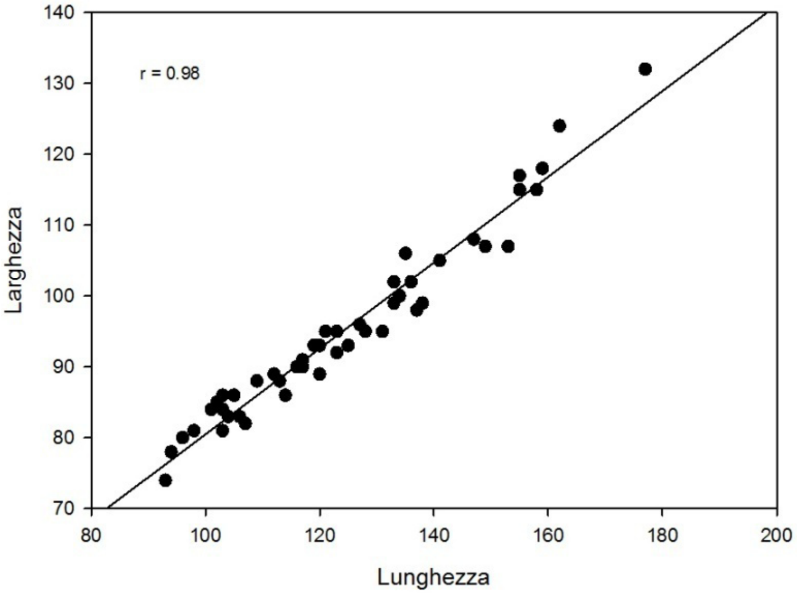
TABLE 1

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

$$\text{PC1} = 33.78*\text{Length} + 33.73*\text{Width} + 33.57*\text{Height}$$

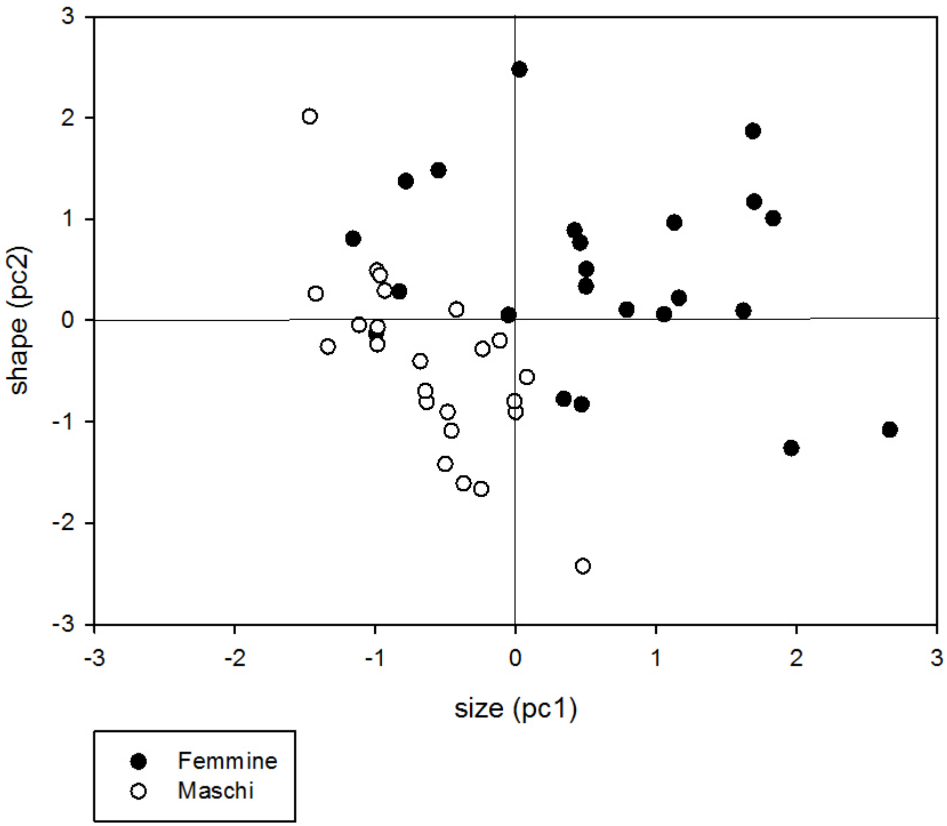
$$\text{PC2} = -1.57*\text{Length} - 2.33*\text{Width} + 3.93*\text{Height}$$

- 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

Credit: Alessandro Giuliani

Female turtles are larger and have more exaggerated height ☺

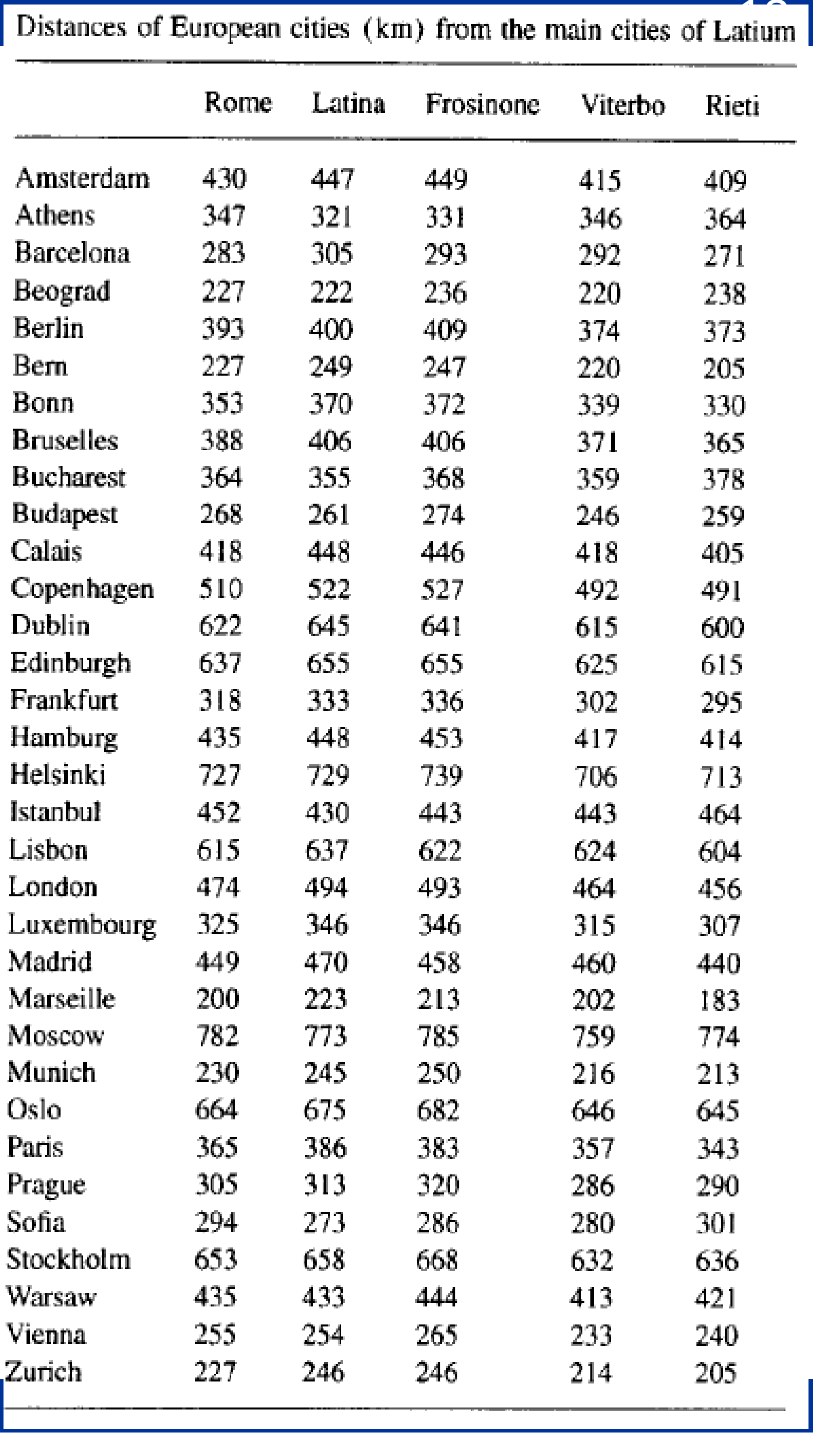


Exercise

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

  Are Madrid and Warsaw near each other?

Giuliani et al., Physics Letters A, 247:47-52, 1998



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

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

PCA IN BIOMARKER SELECTION

PCA in biomarker selection

When PCA is applied e.g. on gene expression data,

- **PCs w/ large variance ≈ diff expressed pathways**
- **Variables with large coefficients in a PC ≈ key genes in the pathway associated with that PC**

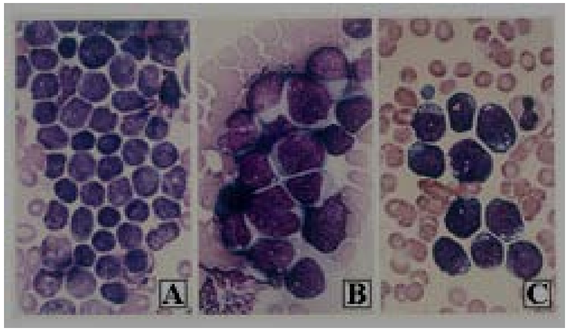
PCA can be a useful biomarker-selection approach

- **E.g., biomarkers ≈ genes w/ high loading**
  - Loading of gene x = $\Sigma_j \mid \alpha_{xj} * \sigma_j^2 \mid$, where α_{xj} is coefficient of x in PC_j, and σ_j^2 is variance of PC_j

Example

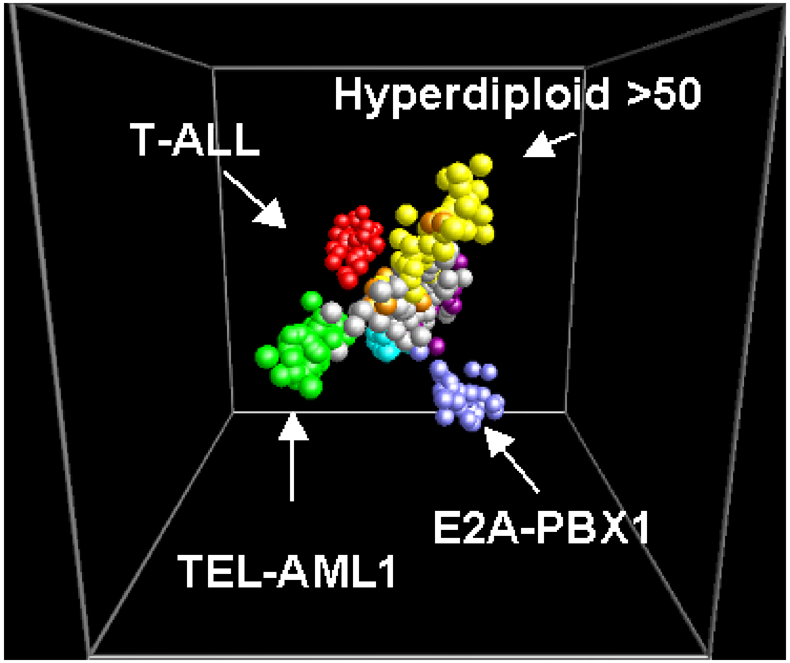
- **Major subtypes: T-ALL, E2A-PBX, TEL-AML, BCR-ABL, MLL genome rearrangements, Hyperdiploid>50**
- **Diff subtypes respond differently to same Tx**
- **Over-intensive Tx**
  - Development of secondary cancers
  - Reduction of IQ
- **Under-intensiveTx**
  - Relapse

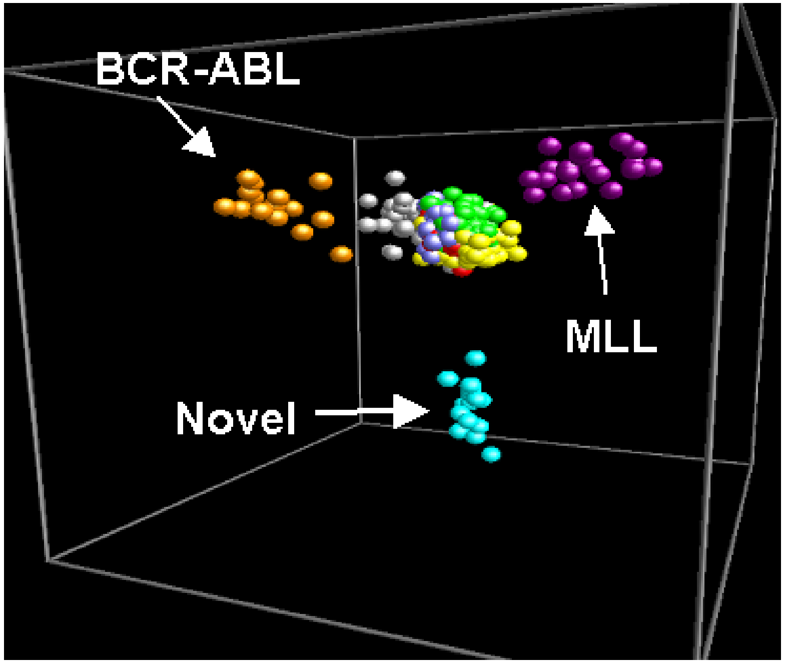
- **The subtypes look similar**



- **Can we diagnosis the subtypes based on gene expression profiling?**

PCA in ALL subtype diagnosis





- **Steps:**
  - Identify genes with high variance
  - Perform PCA on them
  - Plot using PC1 to 3

Induction of hypothesis

- The PCs capture different biological pathways. The values of PCs capture different states of these pathways

- Hypothesis: If patient X has ALL subtype T, X’s biological pathways are in state S_T

… and abduction during diagnosis

- Observation: John’s biological pathways are in state S_T

- Abduction: John has ALL subtype T

BATCH EFFECTS

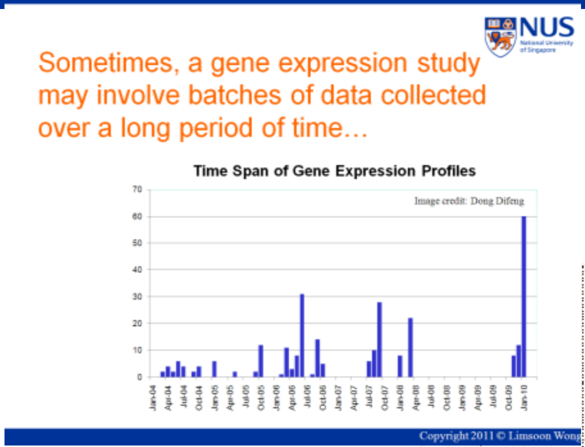
What are batch effects?

- **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**

Visualizing batch effects

- **Rank variables / genes by variance**
- **Keep those with high variance (e.g. top 30-50%)**
- **Perform PCA on them**
- **Make scatter plot of the first 2-3 PCs**
  - Do the subjects clusters by batch?
- **Make paired boxplot of each PC wrt class and batch variables**
  - Is PC more correlated with batch?

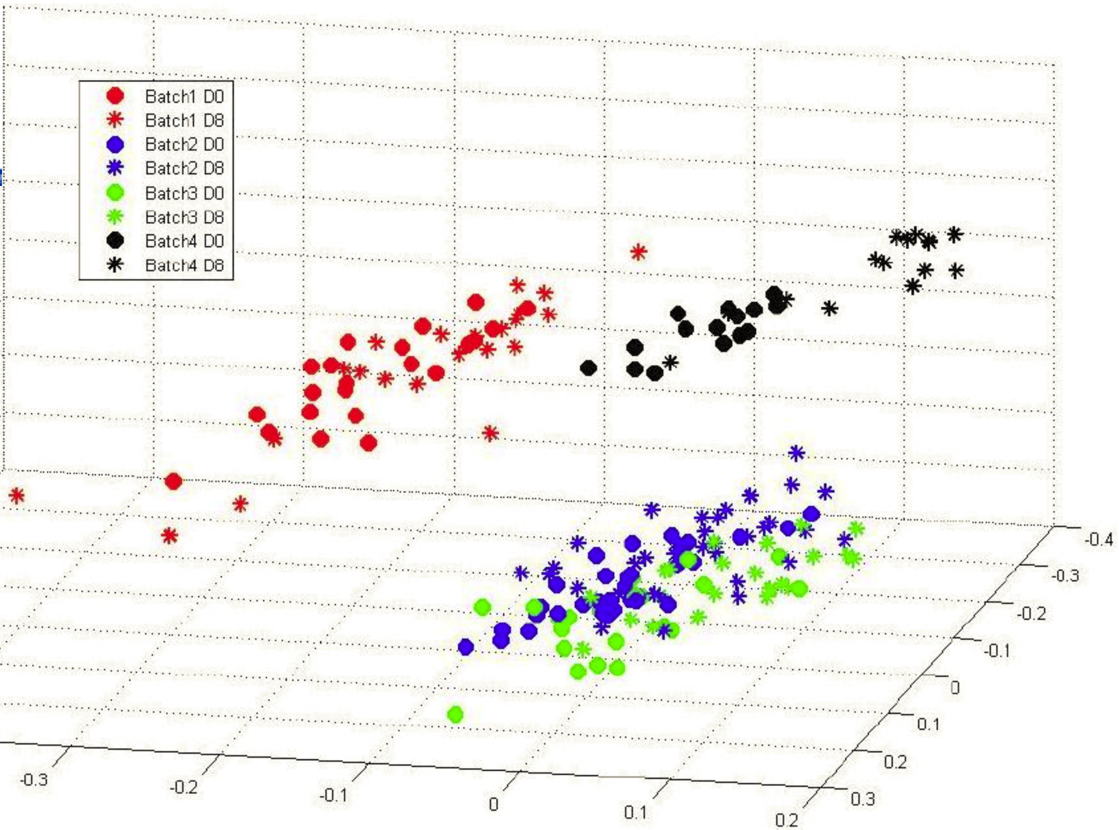


Sometimes, a gene expression study may involve batches of data collected over a long period of time…

Time Span of Gene Expression Profiles

Image credit: Dong Difeng

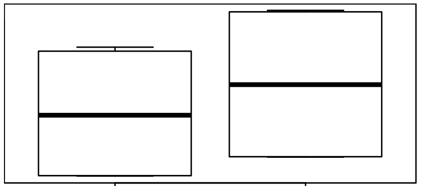
PCA scatter plot

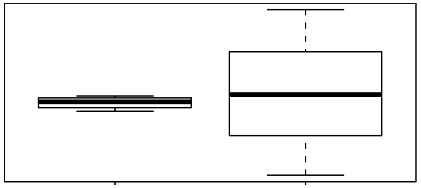


- **Samples from diff batches are grouped together, regardless of subtypes and treatment response**

Image credit: Difeng Dong’s PhD dissertation, 2011

Paired boxplots of PCs





- **It is easier to see which PC is enriched in batch effects by showing, side by side, the distribution of values of each PC stratified by class and suspected batch variables**

Normalization

- **Aim of normalization: Reduce variance w/o increasing bias**

- **Scaling method**
  - Intensities are scaled so that each array has same ave value
  - E.g., Affymetrix's

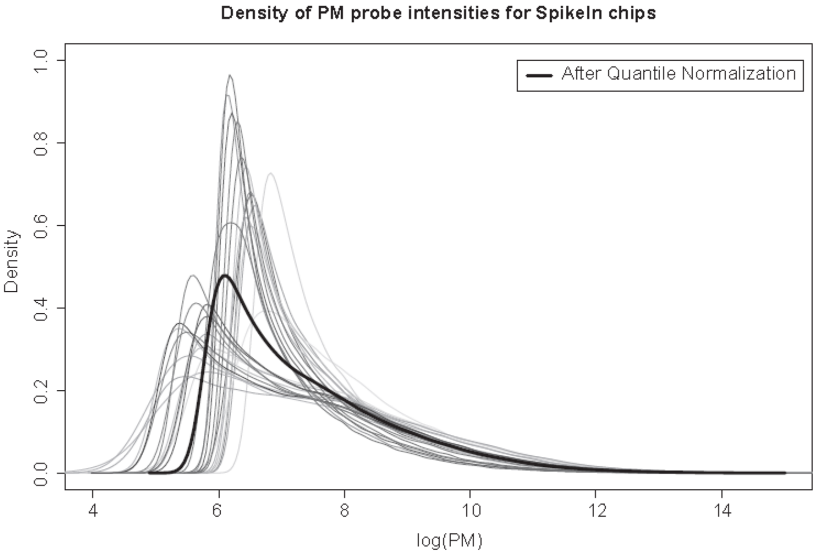
- **Transform data so that distribution of probe intensities is same on all arrays**
  - E.g., $(x - \mu) / \sigma$

- **Quantile normalization**

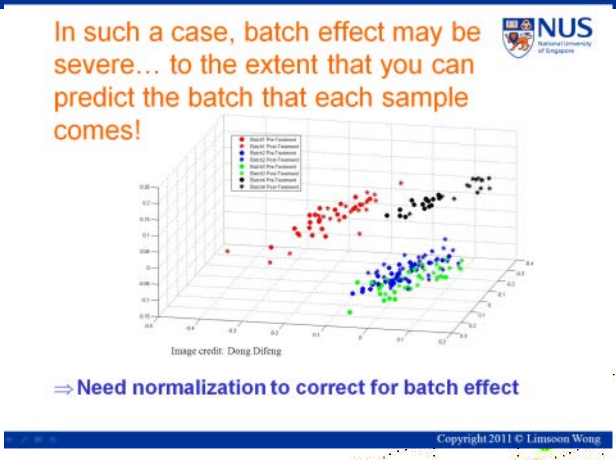
- **Gene fuzzy score, GFS**

Quantile normalization

- **Given *n arrays of length p, form X of size p × n where each array is a column***
- **Sort each column of *X to give X*$_{sort}$**
- **Take means across rows of *X*$_{sort}$ *and assign this* mean to each elem in the row to get *X'*$_{sort}$**
- **Get *X*$_{normalized}$ *by arranging each column of X'*$_{sort}$ to have same ordering as *X***



- Implemented in some microarray s/w, e.g., EXPANDER



In such a case, batch effect may be severe… to the extent that you can predict the batch that each sample comes!

Image credit: Dong Difeng

⇒Need normalization to correct for batch effect

Copyright 2011 © Limsoon Wong

After quantile normalization

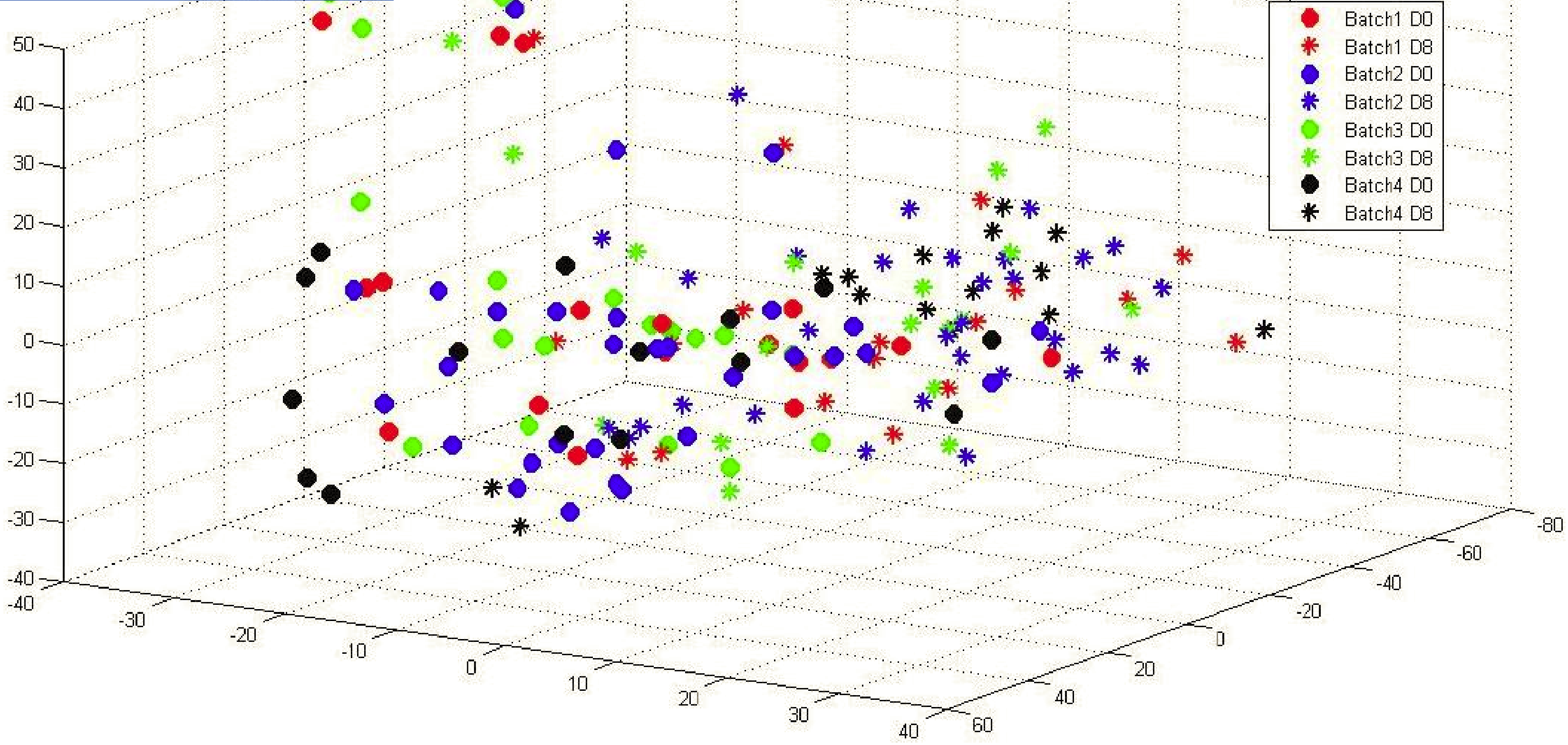
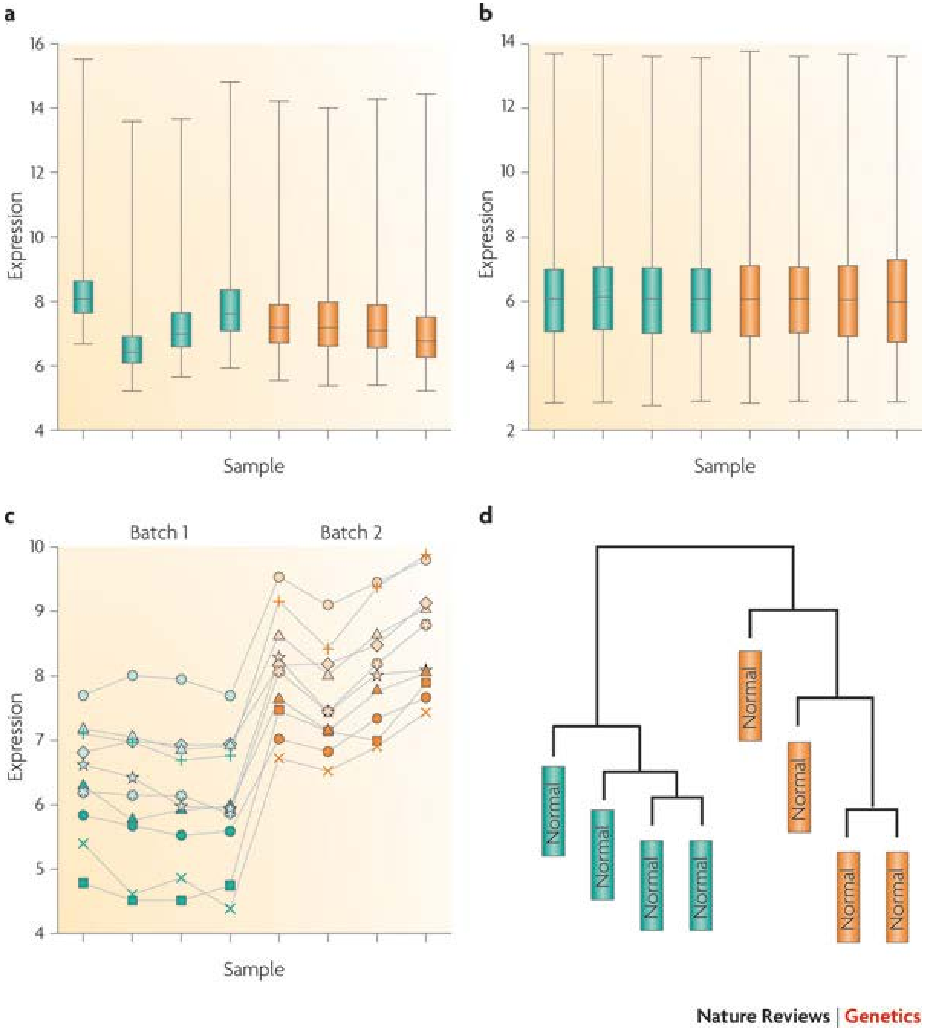


Image credit: Difeng Dong's PhD dissertation, 2011

Caution: It is difficult to eliminate batch effects effectively



Leek et al, Nature Reviews Genetics, 11:733-739, 2010

Green and orange are normal samples differing in processing date

a: Before normalization

b: Post normalization

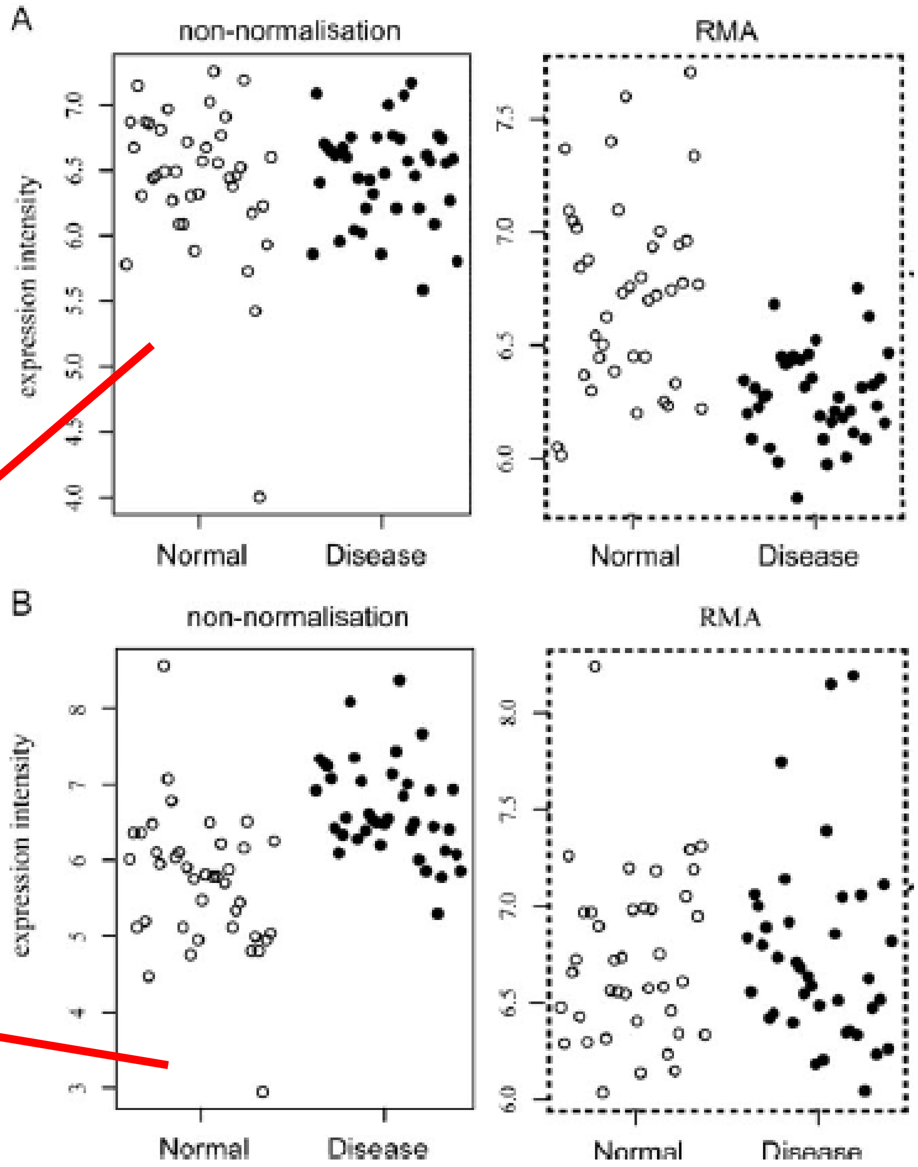
c: Checks on individual genes susceptible to batch effects

d: Clustering after normalization (samples still cluster by processing date)

Caution: “Over normalized” signals in cancer samples

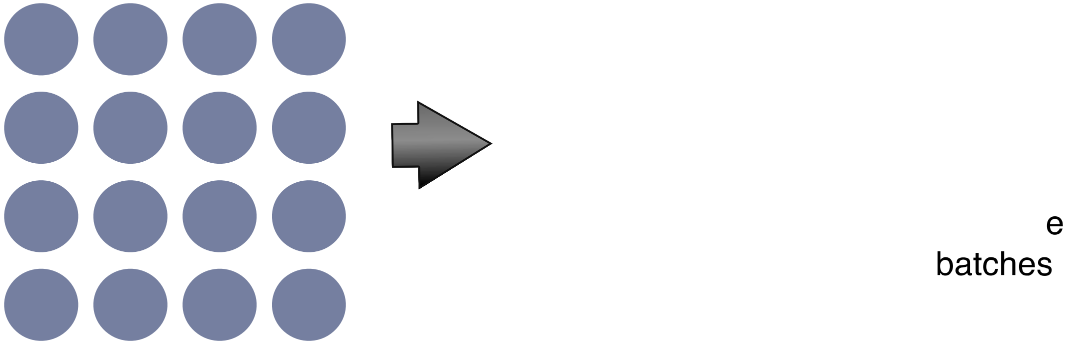
A gene normalized by quantile normalization (RMA) was detected as down-regulated DE gene, but the original probe intensities in cancer samples were not diff from those in normal samples

A gene was detected as an up-regulated DE gene in the non-normalized data, but was not identified as a DE gene in the quantile-normalized data



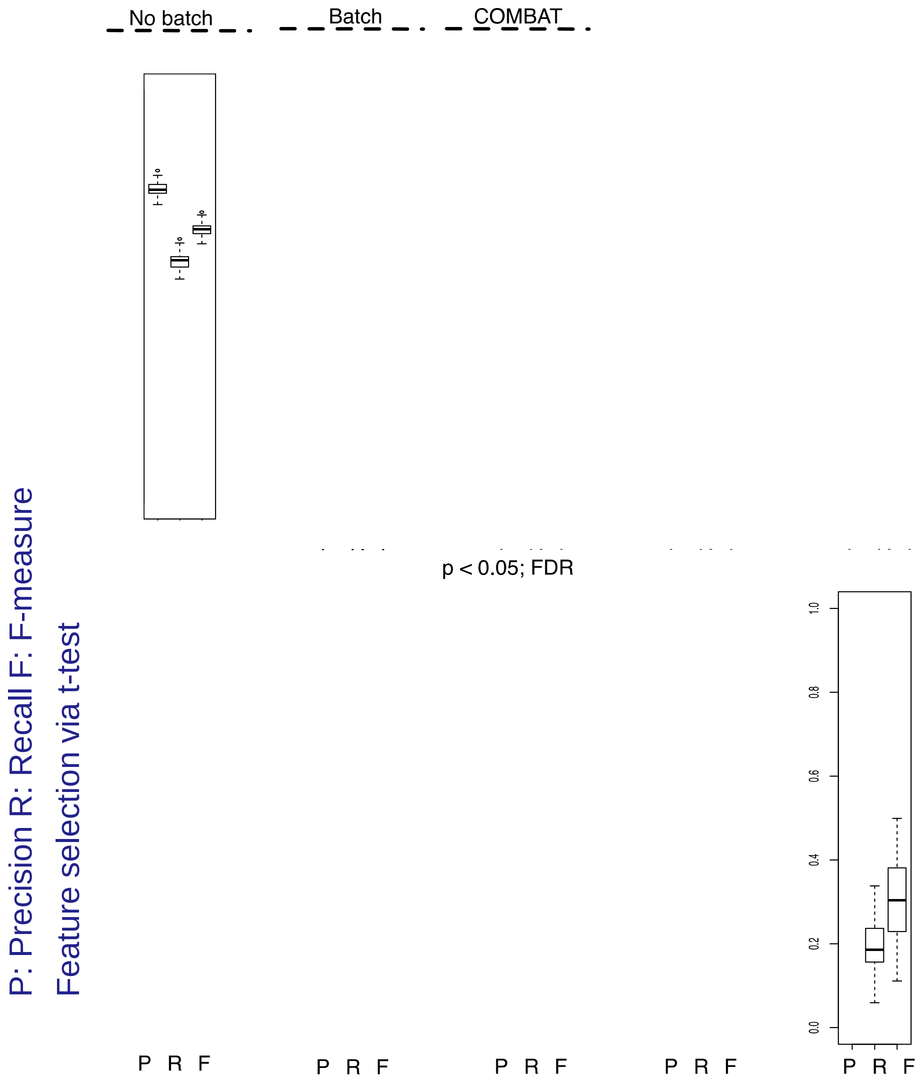
Wang et al. *Molecular Biosystems*, 8:818-827, 2012

Simulated data



- Real one-class data from a multiplex experiment (no batches); n = 8
- Randomly assigned into two phenotype classes D and D*, 100x
- 20% biological features are assigned as differential, and a randomly selected effect size (20%, 50%, 80%, 100% and 200%) added to D*
- Half of D and D* are assigned to batch 1, and the other half assigned to batch 2. A randomly selected batch effect (20%, 50%, 80%, 100% and 200%) is added to all features in batch 1

Batch-effect correction can introduce false positives



- **Precision is strongly affected by batch correction via COMBAT**
- **This means that false positives are added post-batch correction. Data integrity is affected**
- **Moreover, post-batch correction does not restore performance to where no batch is present**

Exercise

- **Why normalization methods like mean scaling, z-score, and quantile normalization sometimes do not work well?**

PCA FOR ISOLATING BATCH EFFECTS

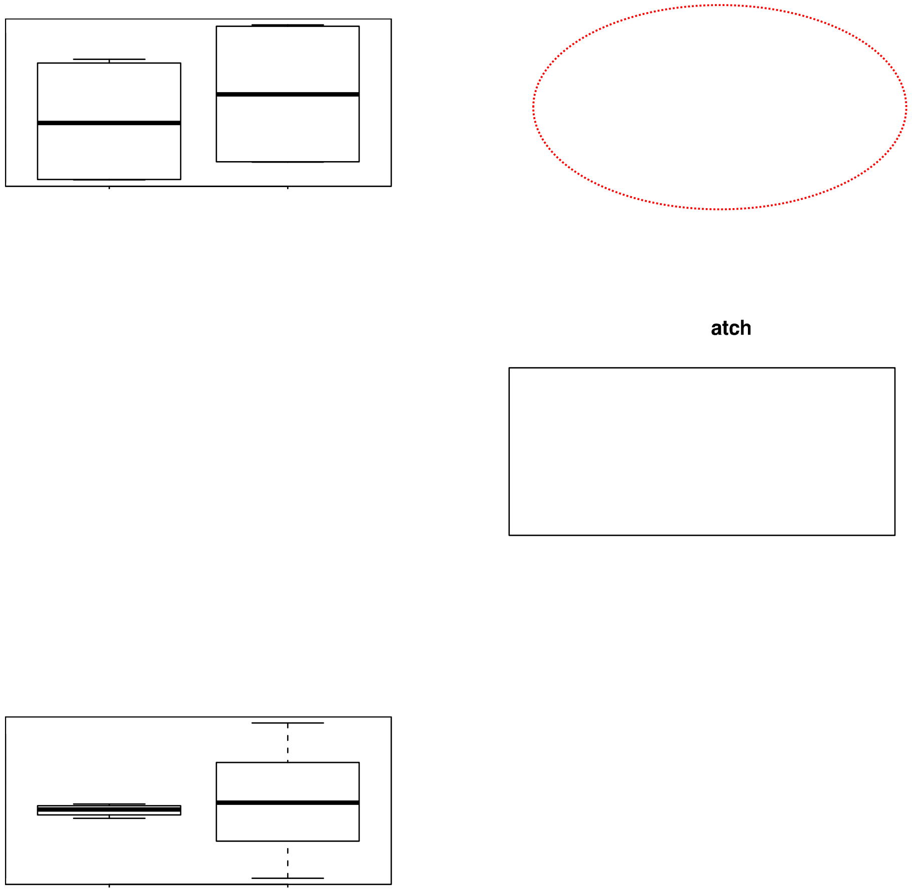
PCA for isolating batch effects

- When a batch effect is observed, it is common practice to apply a batch effect-removal or -correction method

- However, this does not necessarily work well in practice. Moreover, if the data does not fit the correction method's assumptions, it may lead to false positives

- Instead, we may opt for a more direct strategy by simply removing PCs (usually PC1) enriched in batch effects, and deploying the remaining PCs as features for analysis

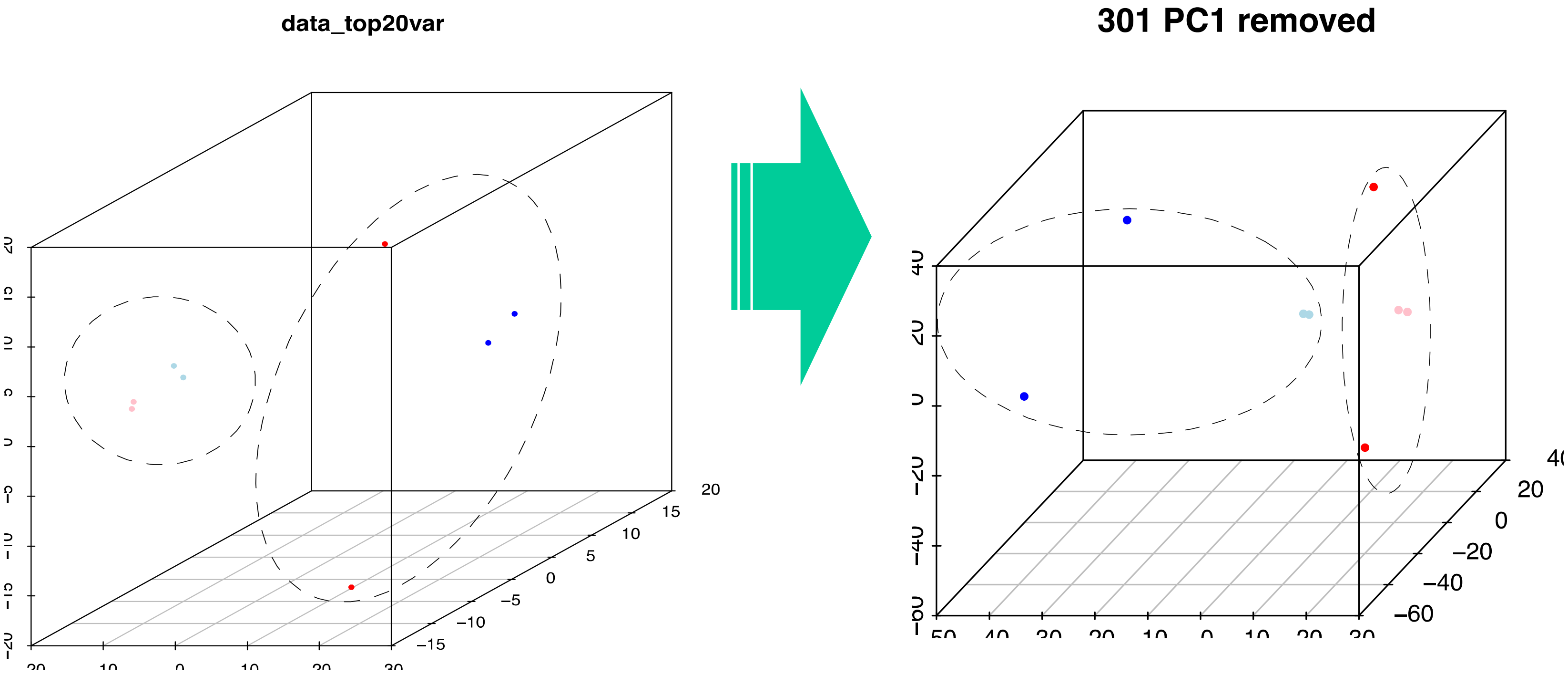
Goh & Wong, “Protein complex-based analysis is resistant to the obfuscating consequences of batch effects”, *BMC Genomics*18(Suppl2):142, 2017



Determine PCs associated with batch using paired boxplots of PCs

- **Batch effects dominate in PC1**

Removal of batch effect-laden PCs removes most batch effects



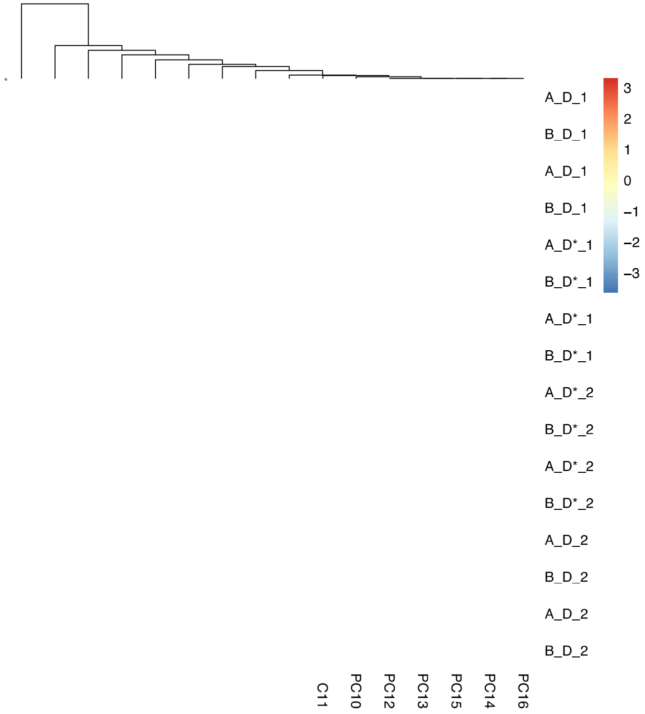
data_top20var

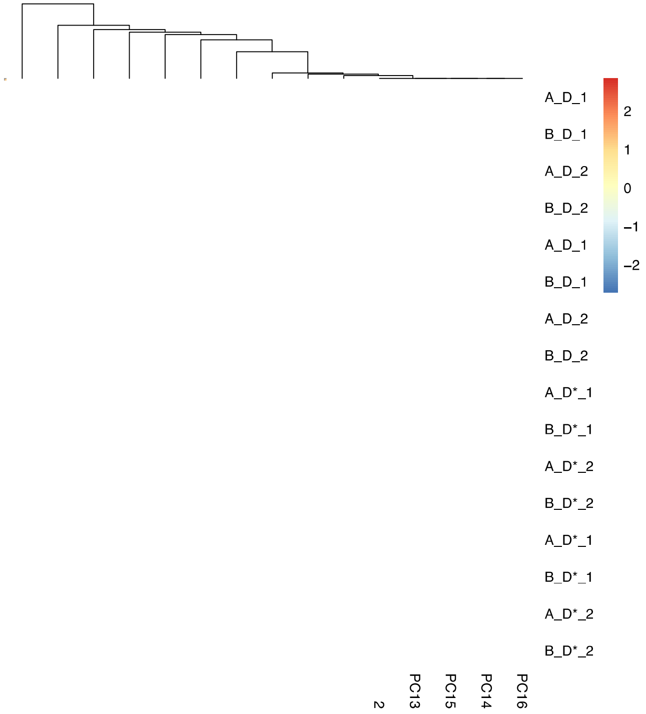
301 PC1 removed

● D, Rep 1 ● D*, Rep 1 ● D, Rep 2 ● D*, Rep 2

Samples separate by class post PC1 removal, no batch subgrouping

A and B are different datasets with different batch effects inserted



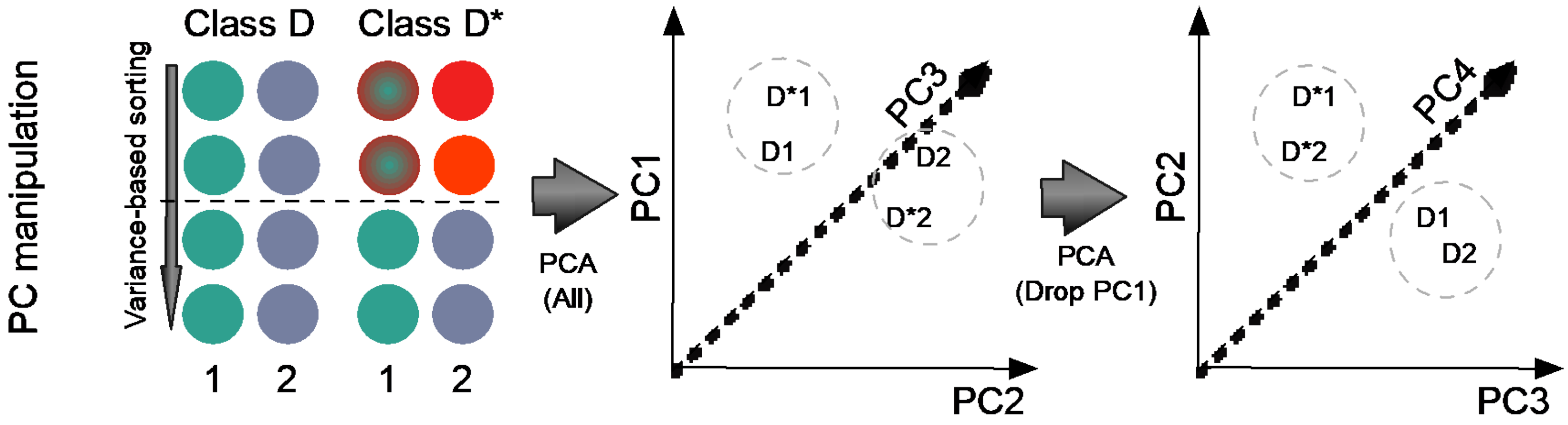


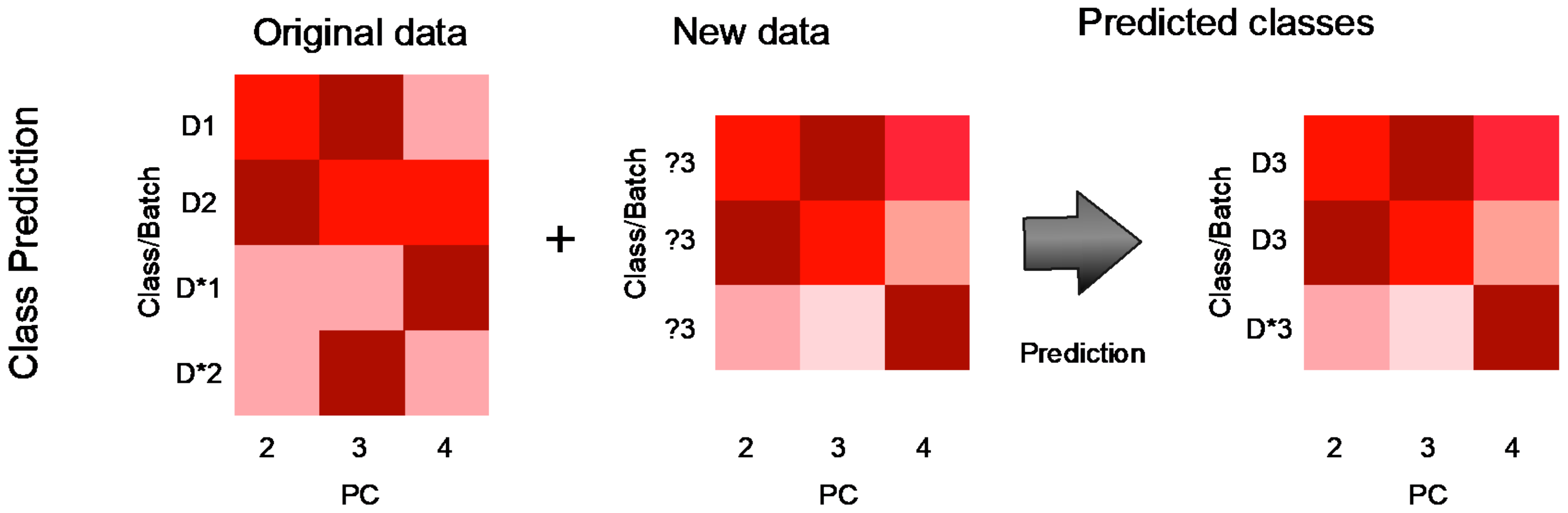
Batch effects dominate

Class-effect discrimination recovered

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

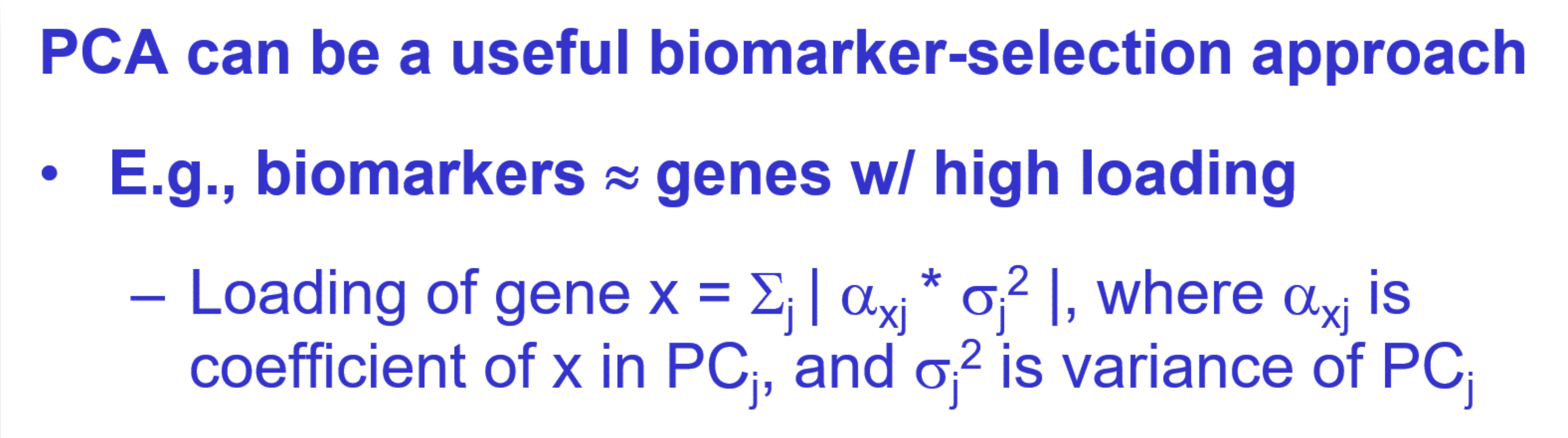
In short, PC manipulation is helpful for dealing w/ batch effects





Exercise

- **Suggest a modification to the formula below to avoid selecting genes laden with batch effects**



PCA can be a useful biomarker-selection approach

- **E.g., biomarkers ≈ genes w/ high loading**
  - Loading of gene x = $\Sigma_j \mid \alpha_{xj} * \sigma_j^2 \mid$, where α_{xj} is coefficient of x in PC_j, and σ_j^2 is variance of PC_j

BATCH EFFECT-RESISTANT FEATURE SELECTION

What if class and batch effects are strongly confounded?

- **Neither batch-effect correction nor PCA work well**

- **We also do not want to inadvertently lose information on disease subpopulations (which look like batch effects but are meaningful)**

⇒ Consider using protein complexes / subnetworks of biological pathways as biomarkers / context for biomarker selection

FSNET

- **FSNET --- a protein complex-based feature-selection methods. Use expression rank-based weighting method (viz. GFS) on individual proteins, followed by intra-class-proportion weighting**

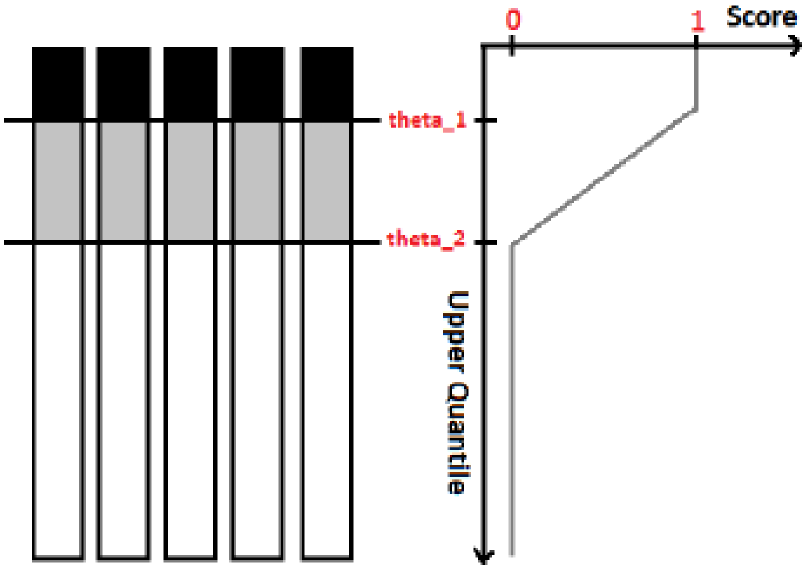
And for comparison …

- **SP is the protein-based two-sample t-test**

Goh & Wong, “Protein complex-based analysis is resistant to the obfuscating consequences of batch effects”, *BMC Genomics*, 18(Suppl 2):142, 2017

FSNET

- **β(g,C)**
  - Proportion of tissues in class C that have protein g among their most-abundant proteins
- **Score(S,p,C)**
  - Score of protein complex S and tissue p weighted based on class C
- **f_{SNET}(S,X,Y,C)**
  - Complex S is differentially high in sample set X and low in sample set Y, weighted based on class C, when f_{SNET}(S,X,Y,C) is at largest extreme of t-distribution

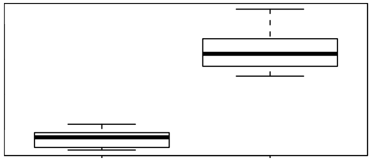


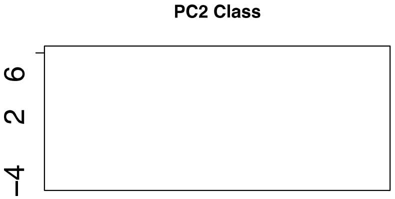
$$\beta(g_i, C_j) = \sum_{pk \epsilon c_j} \frac{fs(g_i, p_k)}{|C_j|}$$

$$\text{score}(S, p_k, C_j) = \sum_{g_i \in S} fs(g_i, p_k) * \beta(g_i, C_j)$$

$$f_{\text{SNET}}(S, X, Y, C_j) = \frac{\text{mean}(S, X, C_j) - \text{mean}(S, Y, C_j)}{\sqrt{\frac{\text{var}(S,X,C_j)}{|X|} + \frac{\text{var}(S,Y,C_j)}{|Y|}}}$$

Network-based methods are enriched for class-related variation (Real data)

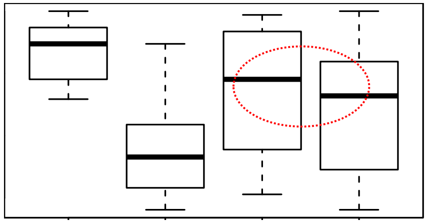




- **PCA on SP-selected genes: Class & batch effects are confounded; cf. PC2**
- **PCA on FSNET-selected complexes: Class & batch effects are less confounded in top PCs**

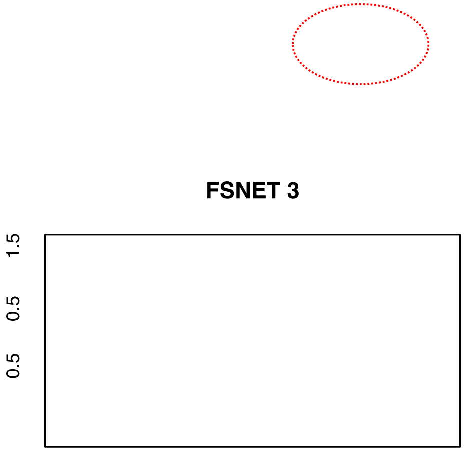
Top complex-based features are strongly associated with class, not batch

Rank 1



Rank 2

Rank 3



- **FSNET captures class effects while being robust against batch effects. In contrast, both class and batch variability are present in the top variables selected by SP**

CONCLUDING REMARKS

What have we learned?

- PCA is a useful paradigm for biomarker selection
- PCA is not just a visualization tool; it can also be used for dealing with batch effects
- When class & batch effects are deeply confounded at the level of proteins / genes, it is might be better to analyze at the level of protein complexes / pathway subnetworks

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
- [GFS] Belorkar & Wong. *BMC Bioinformatics*, 17(Suppl 17):540, 2016
- [FSNET] Goh & Wong, *BMC Genomics*, 18(Suppl 2):142, 2017