Machine Learning: Dimensionality

Dr Chris Yeomans

c.m.yeomans@exeter.ac.uk

Camborne School of Mines, CEMPS, Penryn Campus, University of Exeter

Acknowledgements

Dr JJ Valetta

- Now at University of St Andrews

Dr Jiangjiao Xu

- University of Exeter

Additional resources:

https://github.com/GeostatsGuy

https://www.youtube.com/channel/UCLqEr-xV-ceHdXXXrTId5ig

Intended Learning Outcomes

By the end of this lecture you will:

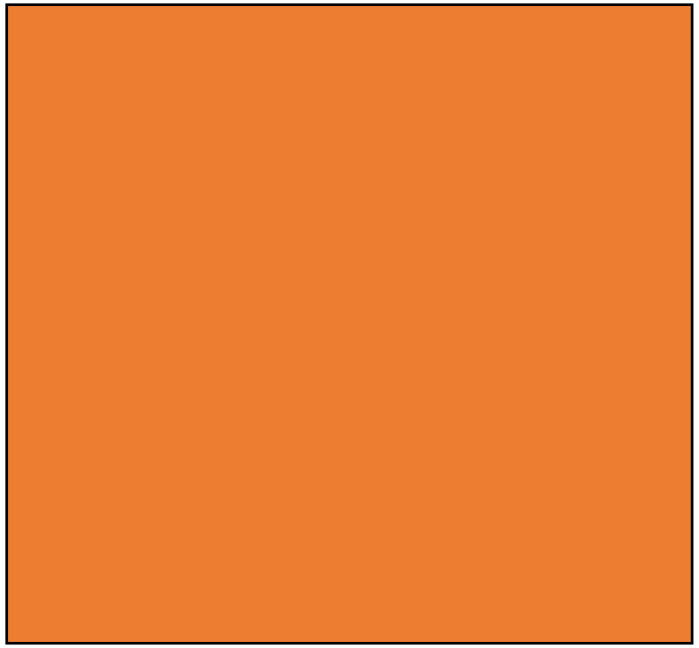
- Appreciate the need for dimensionality reduction
- Understand basic feature extraction and feature selection
- Be aware of various methods for dimensionality reduction

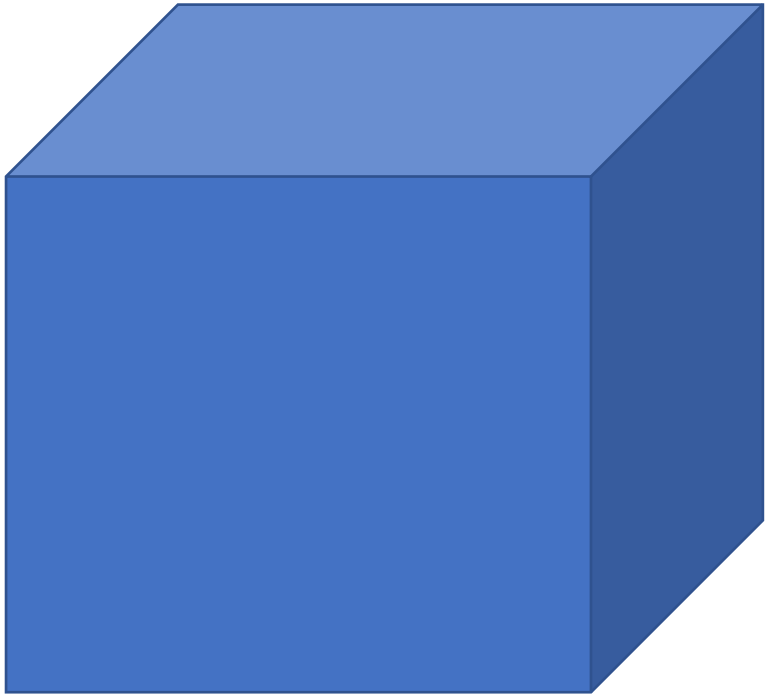
Overview

- Understanding dimensions
- Why reduce the number of dimensions?
- Types of dimensionality reduction
  - Pearson Correlation Coefficients
  - Principal Component Analysis
  - Independent Component Analysis
  - t-distributed Stochastic Neighbour Embedding (t-SNE)
- Feature selection

Data and dimensions

- Dimensions are easy, right?

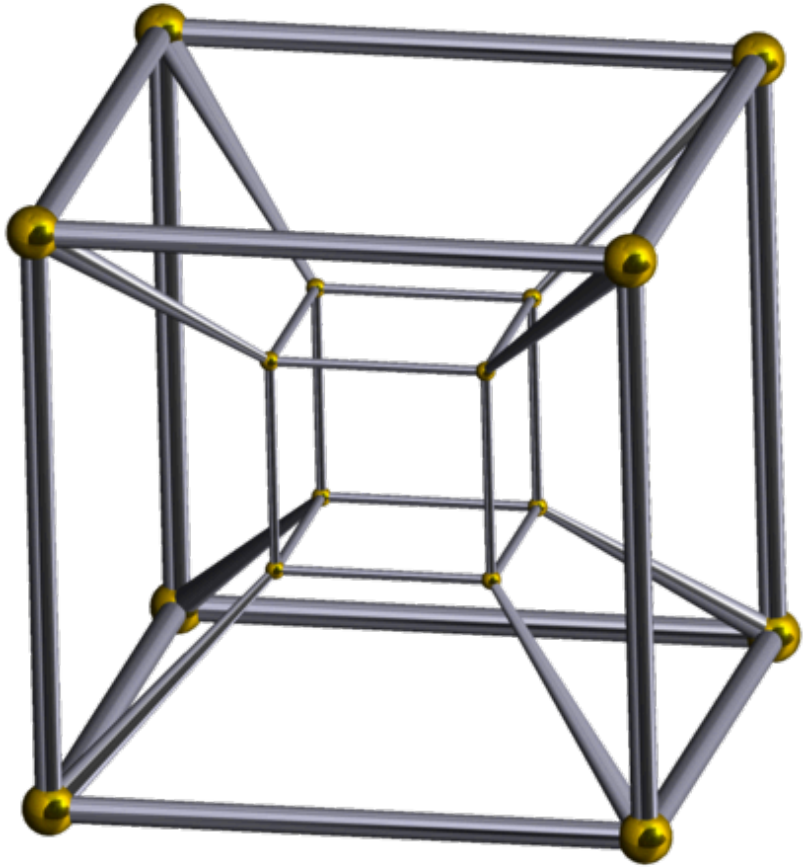




+ time

Data and dimensions

- What about a four dimensional cube? AKA, the tesseract



Data and dimensions

- With data frames, for every additional variable you are adding a new dimension

Location / Author description	TiO2	Al2O3	Fe2O3t	Fe2O3	FeO	FeOT	MnO	MgO	CaO	Na2O	K2O	P2O5	F	Cl	B	P	B2O3	Li2O	S	H2O+	H2O-	Ag	As	Ba	Be	Bi	Cd	Co	Cr	Cs	Cu	Ga	Ge	Hf
Holmans	0.22	15.78		1.21	0.62	1.71	0.05	0.43	0.70	3.03	5.40	0.24																							2.75
Granite West Megilligar	0.03	17.10	0.66					0.19	0.23	4.30	3.70												3.00								13.00	49.00		
Pegmatite East Megilligar	0.04	16.70	1.04					0.21	0.31	3.90	3.85												198.00								10.00	47.00		
Lamorna	0.38	14.81		0.32	2.19	2.48	0.07	0.65	0.85	2.52	5.57	0.29	0.24	0.06			0.02		0.00	1.17	0.27	0.30	7.00	280.00	6.00		0.30	5.30	13.10	31.00	7.00	22.60	2.90	4.60
Lamorna	0.38	14.92		0.16	2.35	2.49	0.05	0.71	0.88	2.67	5.43	0.29	0.23	0.06			0.02		0.00	1.07	0.20	0.30	6.00	275.00	6.00		0.20	5.20	12.90	28.00	4.00	22.60	2.70	4.40
Quarry nr Zennor	0.35	14.72		0.31	2.07	2.35	0.05	0.61	0.85	2.65	5.40	0.28	0.24	0.05			0.04		0.00	1.11	0.26	<0.1	10.00	250.00	5.00		0.30	4.80	9.40	38.00	3.00	22.20	2.90	4.20
Newmill Quarry	0.29	14.75		0.47	1.53	1.95	0.03	0.46	0.80	2.61	5.69	0.30	0.26	0.05			0.06		0.00	1.02	0.28	0.30	54.00	230.00	4.00		0.40	7.60	8.10	36.00	4.00	23.20	3.50	3.90
Quarry nr Pendeen	0.15	14.13		0.52	0.77	1.24	0.02	0.22	0.65	2.69	5.24	0.31	0.62	0.05			0.09		0.00	0.80	0.30	0.10	3.00	48.00	2.00		0.30	0.80	2.70	48.00	6.00	25.80	4.50	2.10
Castle-an-Dinas	0.16	14.67		0.32	1.15	1.44	0.07	0.32	0.64	3.05	4.99	0.27	0.25	0.05			0.06		0.00	0.90	0.30	0.30	35.00	135.00	12.00		0.30	2.30	4.80	52.00	3.00	23.00	3.40	2.10
Castle-an-Dinas	0.16	14.92		0.39	0.98	1.33	0.03	0.35	0.51	2.98	5.19	0.28	0.16	0.03			0.06		0.00	1.03	0.29	<0.1	3.00	135.00	13.00		0.30	1.70	4.20	48.00	7.00	22.80	4.00	2.10
Penrew Quarry	0.25	15.18		0.30	1.32	1.59	0.04	0.49	0.99	3.01	5.03	0.24	0.22	0.02			0.02		0.00	1.14	0.21	0.10	24.00	200.00	9.00		0.20	4.30	11.60	40.00	1.00	27.20	3.40	3.20
Trevone Quarry	0.28	15.15		0.21	1.53	1.72	0.04	0.51	0.98	2.98	5.08	0.26	0.24	0.01			0.02		0.00	0.94	0.21	0.10	16.00	195.00	9.00		0.20	4.00	11.60	44.00	1.00	26.20	3.40	3.40
Boscahan Quarry	0.27	14.48		0.12	1.54	1.65	0.04	0.48	0.99	3.02	4.70	0.24	0.23	0.02			0.02		0.01	0.86	0.17	0.10	16.00	175.00	7.00		<0.1	4.10	11.40	38.00	2.00	25.40	3.10	3.30
Chywoon Quarry	0.25	15.44		0.19	1.44	1.61	0.04	0.47	0.96	3.13	5.10	0.25	0.23	0.02			0.03		0.00	1.02	0.23	<0.1	23.00	195.00	9.00		0.20	3.50	10.30	44.00	1.00	27.20	3.80	3.30
Pelestine Quarry	0.24	14.82		0.14	1.38	1.51	0.04	0.44	0.88	3.05	4.85	0.23	0.28	0.01			0.03		0.00	0.82	0.16	0.10	20.00	170.00	8.00		0.30	3.60	8.90	52.00	<1	26.00	3.60	2.90
400 decline S. Crofty	0.25	15.44		0.25	1.42	1.64	0.04	0.52	0.91	3.16	5.11	0.25	0.25	0.01			0.02		0.00	1.06	0.19	0.10	22.00	195.00	10.00		0.20	3.30	9.80	48.00	1.00	26.80	3.90	3.10
Rosemanowas	0.21	14.99		0.21	1.25	1.44	0.04	0.40	0.83	3.11	5.05	0.23	0.27	0.01			0.03		0.00	0.92	0.21	0.20	31.00	170.00	10.00		<0.1	3.70	8.80	44.00	1.00	26.20	3.90	2.90
Luxulyan Quarry	0.32	14.08		0.12	1.92	2.03	0.06	0.49	0.99	2.88	4.90	0.25	0.24	0.04			0.02		0.00	0.84	0.17	0.10	3.00	210.00	7.00		0.30	3.50	5.60	32.00	3.00	21.00	3.10	4.00
Goonbarrow	0.22	14.43		0.28	1.33	1.58	0.04	0.36	0.68	2.95	5.21	0.28	0.25	0.04			0.05		0.00	0.78	0.18	0.10	27.00	140.00	7.00		0.30	3.40	5.30	52.00	4.00	23.60	4.10	3.00
Craddock Moor	0.23	14.88		0.16	1.37	1.51	0.04	0.42	0.88	2.73	5.34	0.27	0.29	0.01			0.03			1.17	0.13	0.10	18.00	180.00	7.00		0.30	4.50	8.20	40.00	2.00	23.60	3.00	3.00
De Lank Quarry	0.22	15.22		0.21	1.22	1.41	0.05	0.41	0.89	3.10	4.96	0.29	0.28	0.01			0.02		0.00	1.01	0.17	0.10	16.00	175.00	13.00		0.30	3.00	6.70	42.00	2.00	27.00	3.00	3.20
Bolventor road cutting	0.19	14.54		0.31	0.97	1.25	0.05	0.27	0.74	2.84	5.19	0.29	0.39	0.01			0.04		0.00	0.97	0.18	0.10	8.00	110.00	9.00		0.30	2.40	3.60	78.00	3.00	23.20	4.50	2.80
Blackenstone Quarry	0.42	13.85		0.23	2.48	2.69	0.08	0.64	1.48	3.00	4.78	0.23	0.16	0.05			0.02		20.00	0.73	0.21	0.20	11.00	195.00	12.00		0.40	5.60	5.80	41.00	6.00	20.20	2.80	4.40
Quarry nr Haytor	0.30	13.83		0.23	1.79	2.00	0.06	0.57	1.17	3.13	4.57	0.18	0.18	0.06			0.02		10.00	0.76	0.16	0.10	7.00	235.00	7.00		<0.1	4.50	8.40	36.00	1.00	20.80	3.30	4.00
Haytor Quarry	0.26	13.70		0.17	1.69	1.84	0.07	0.44	0.75	2.99	5.03	0.19	0.10	0.03			0.01		15.00	0.85	0.16	0.30	3.00	140.00	14.00		<0.1	3.40	5.90	33.00	1.00	19.60	3.20	3.20
Haytor Quarry	0.22	13.63		0.16	1.62	1.76	0.08	0.40	0.54	2.98	5.01	0.18	0.06	0.03			0.01		40.00	1.00	0.22	0.20	5.00	110.00	9.00		0.40	3.40	4.90	28.00	3.00	19.40	3.60	2.80
Marrivale Quarry	0.18	13.35		0.33	1.41	1.71	0.08	0.25	0.51	3.01	4.94	0.21	0.26	0.04			0.05		10.00	0.77	0.15	0.10	5.00	68.00	16.00		0.10	2.40	2.70	69.00	5.00	22.20	4.70	3.10
Prison Quarry	0.21	13.35		0.48	1.41	1.84	0.07	0.30	0.60	2.96	4.89	0.21	0.25	0.04			0.07		10.00	0.66	0.19	0.30	6.00	78.00	14.00		0.40	2.40	3.00	74.00	4.00	22.00	4.70	3.30
Prison Quarry	0.17	13.23		0.31	1.41	1.69	0.07	0.27	0.44	2.74	5.22	0.21	0.28	0.04			0.07		10.00	0.73	0.25	0.30	9.00	52.00	18.00		0.30	1.60	2.40	93.00	7.00	22.40	4.40	3.00
Lamorna	0.41	14.68		0.42	2.12	2.50	0.06	0.70	0.92	2.72	5.39	0.28	0.24				0.02						12.00	278.00										
Penryn Granite Quarry	0.26	15.00		0.24	1.36	1.58	0.04	0.37	0.68	3.09	5.28	0.25	0.21				0.01						37.00	196.00										
Penryn Granite Quarry	0.29	21.22		0.04	0.06	0.10	0.02	0.08	2.88	6.92	1.06	0.05	0.08				<0.002						6.50	31.00										
Quarry	0.28	14.72		0.28	1.40	1.65	0.04	0.28	0.72	3.04	5.02	0.23	0.22				0.01						16.00	201.00										
Chywoon Quarry	0.25	14.83		0.25	1.31	1.53	0.04	0.28	0.70	3.11	5.19	0.31	0.22				0.02						25.00	191.00										
Quarry nr Herniss	0.24	14.48		0.25	1.26	1.48	0.05	0.34	0.66	3.12	4.94	0.29	0.25				0.03						39.00	171.00										
Trevone Quarry	0.27	14.56		0.15	1.48	1.61	0.05	0.42	0.78	3.02	4.90	0.36	0.23				0.02						22.00	200.00										
Trevone Quarry	0.07	15.02		0.35	0.50	0.81	0.08	0.10	0.17	3.99	3.77	0.29	0.75				0.03						9.00	30.00										
Bosahan Quarry	0.32	15.04		0.44	1.50	1.90	0.05	0.57	0.80	3.14	4.83	0.23	0.22				0.02						19.00	195.00										
Quarry nr Coverack Bridges	0.22	15.00		0.45	1.18	1.58	0.05	0.23	0.49	3.23	4.95	0.27	0.36				0.04						55.00	151.00										
Carn Marth	0.16	14.83		0.49	0.76	1.20	0.04	0.16	0.41	3.29	4.97	0.33	0.45				0.07						94.00	108.00										
Lands End MGG - Porth Nanven	0.12	13.53	1.49				0.03	0.14	0.35	3.27	5.00	0.23											14.00	8.00			2.00	95.00		47.00	4.00	28.00		
Lands End MGG - Progo Cove	0.16	13.92	1.57				0.04	0.23	0.35	3.20	5.21	0.24											15.00	90.00				55.00		32.00	7.00	25.00		
Lands End - Lamorna	0.38	14.86	2.88				0.05	0.59	0.72	2.82	5.69	0.26											14.00	302.00				81.00		31.00	10.00	23.00		
Lands End	0.21	14.07	1.99				0.05	0.28	0.67	3.43	4.59	0.32											27.00	76.00				83.00		93.00	3.00	29.00		
Lands End	0.31	14.61	2.23				0.05	0.41	0.84	2.88	5.83	0.30											22.00	238.00				66.00		34.00	6.00	24.00		
Lands End	0.28	14.56	2.30				0.05	0.43	0.70	2.82	5.67	0.27										1.00	52.00	247.00				51.00		40.00	6.00	23.00		
Lands End - Castle an Dinas	0.14	14.60	1.77				0.06	0.23	0.58	3.42	5.09	0.24											33.00	140.00				36.00		51.00	4.00	24.00		
Carnmenellis Ga - Rosemanowes	0.22	15.06	1.71				0.04	0.33	0.79	3.39	4.91	0.23											31.00	180.00			1.00	48.00		50.00	3.00	25.00		
Carnmenellis Ga - Rosemanowes	0.21	14.76	1.67				0.04	0.42	0.91	3.24	4.90	0.24											21.00	165.00				70.00		48.00	5.00	26.00		
Carnmenellis Ga - Carnsew	0.25	14.83	1.76				0.08	0.45	0.93	2.93	5.06	0.23											14.00	212.00				66.00		38.00	2.00	27.00		
Carnmenellis Ga - Carnsew	0.26	14.47	1.93				0.05	0.52	0.88	2.71	5.23	0.31											29.00	194.00				74.00		42.00	6.00	27.00		
Carnmenellis Ga - Carnsew	0.26	14.73	1.83				0.06	0.51	1.01	3.20	5.42	0.23											11.00	204.00				63.00		39.00	3.00	28.00		
Carnmenellis Ga - Trevone Quarry	0.21	14.84	1.73				0.06	0.54	0.87	2.84	5.29	0.24											13.00	213.00				68.00		44.00	2.00	26.00		

Why reduce the number of dimensions?

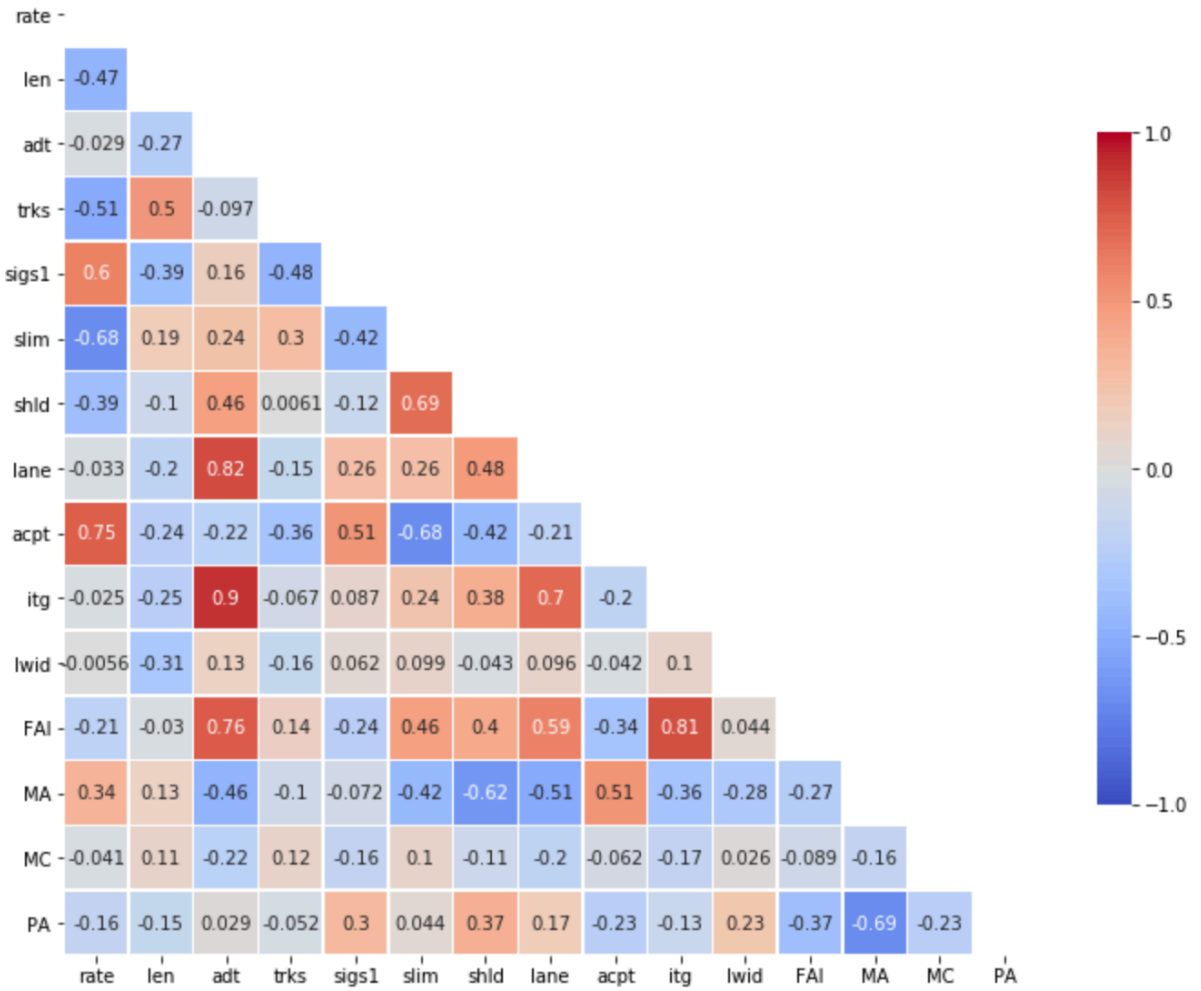
- Achieved through feature extraction and feature selection
- Assist in visualising the data to understand data structure
- Identify best predictors (before further feature enhancement – iterative)
  - E.g. plausible causal drivers under an experimental setup
  - E.g. what controls platinum concentration in the crust, melting (MgO#) or contamination (Ti)
- Enhancing features
- Remove redundant data
- Improve model efficiency (reduce modelling time)

Pearson Correlation Coefficients

- A simple way to determine correlated variables

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}$$

- Often discard one of the correlated variables when $r > 0.8$
- Only considers bivariate correlation
- It is simple but a bit of a sledgehammer



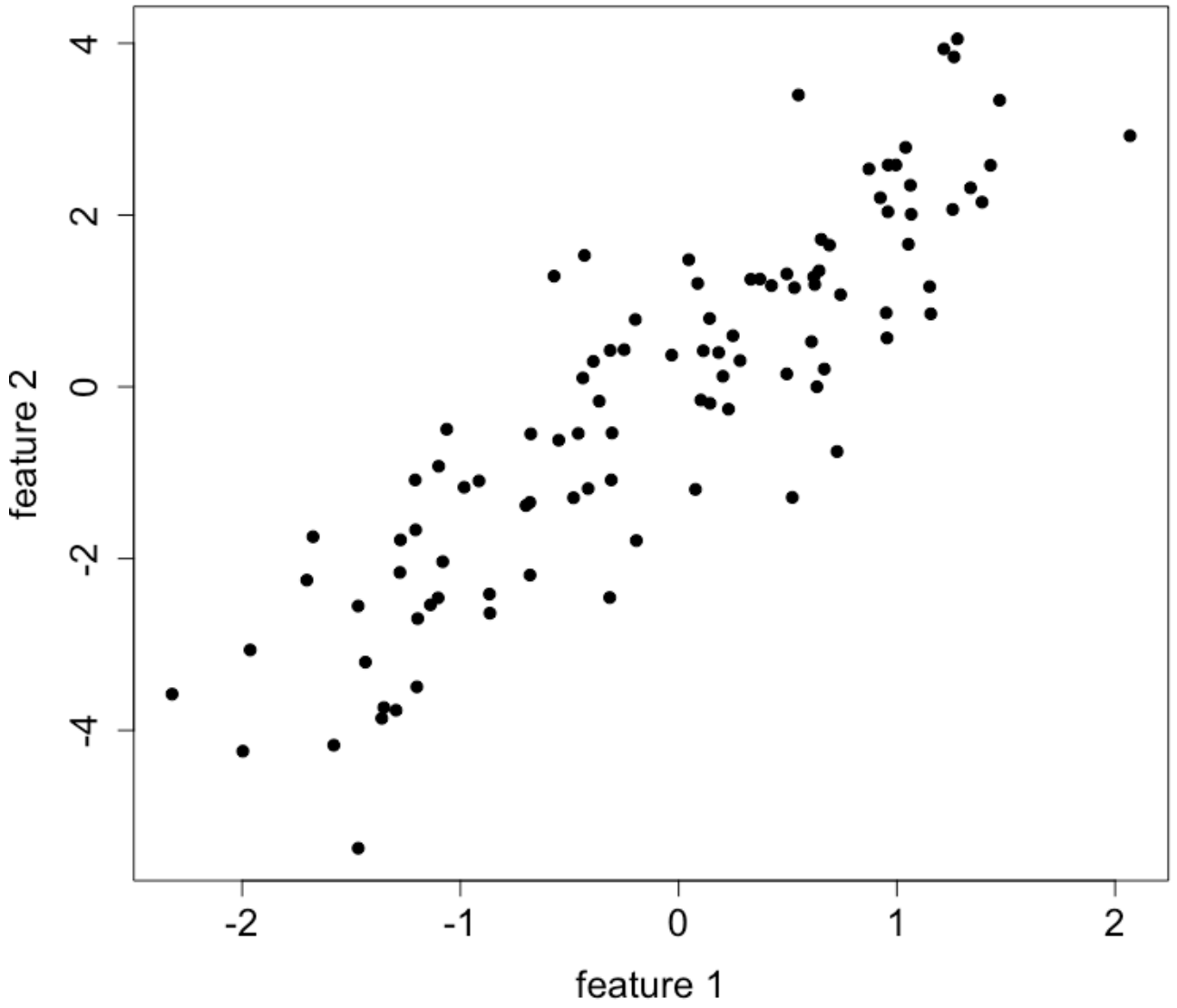
Principal Component Analysis

- A linear dimensionality reduction method
- The new uncorrelated features (PCA 1, PCA 2,...) are weighted (w's) linear combinations of the original data (x's)

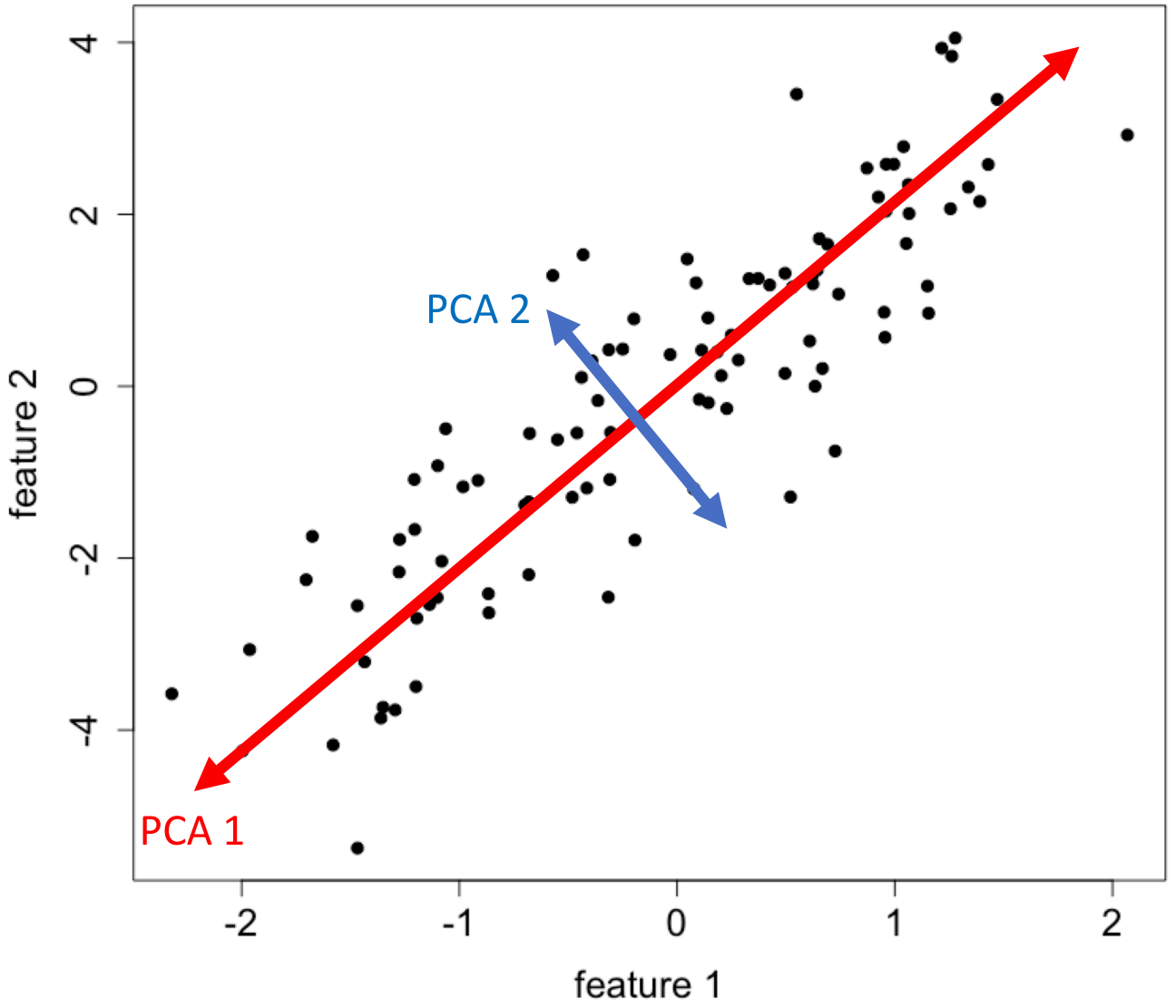
$$\begin{aligned}
\text{PCA 1} &= w_{11}x_1 + w_{12}x_2 + \ldots + w_{1p}x_p \\
\text{PCA 2} &= w_{21}x_1 + w_{22}x_2 + \ldots + w_{2p}x_p \\
&\vdots \\
\text{PCA p} &= w_{p1}x_1 + w_{p2}x_2 + \ldots + w_{pp}x_p
\end{aligned}$$

- Objective is to find directions, called principal components, that maximise the variance of the data

Principal Component Analysis



Principal Component Analysis



t-SNE

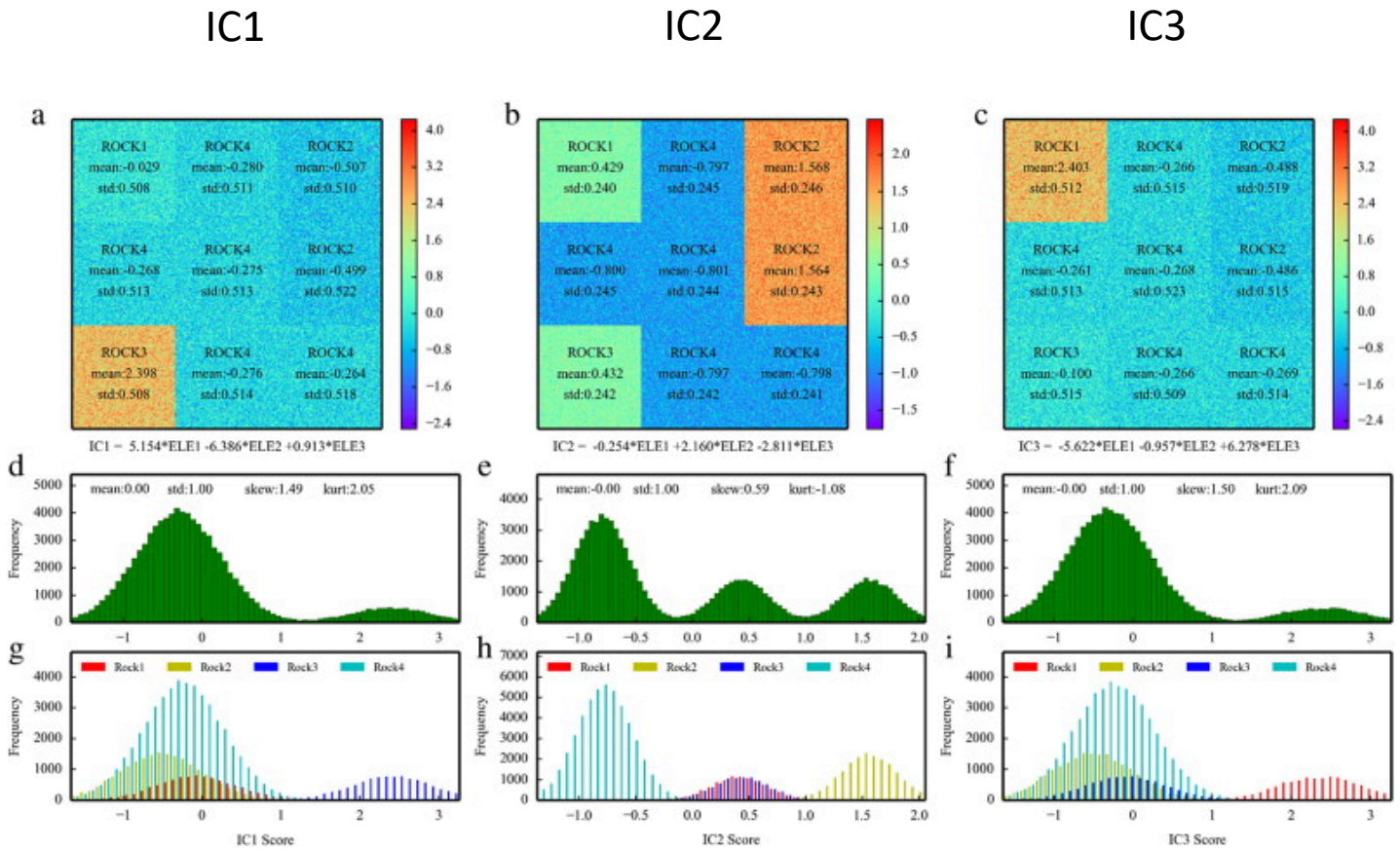
- A non-linear dimensionality reduction method
- Projects data into a lower-dimensional space/embedding such that the original high-dimensional clustering is preserved

e.g. Van der Maaten and G. Hinton. (2008) *Journal of Machine Learning Research*

e.g. Horrocks et al. (2019) *Computers & Geosciences*

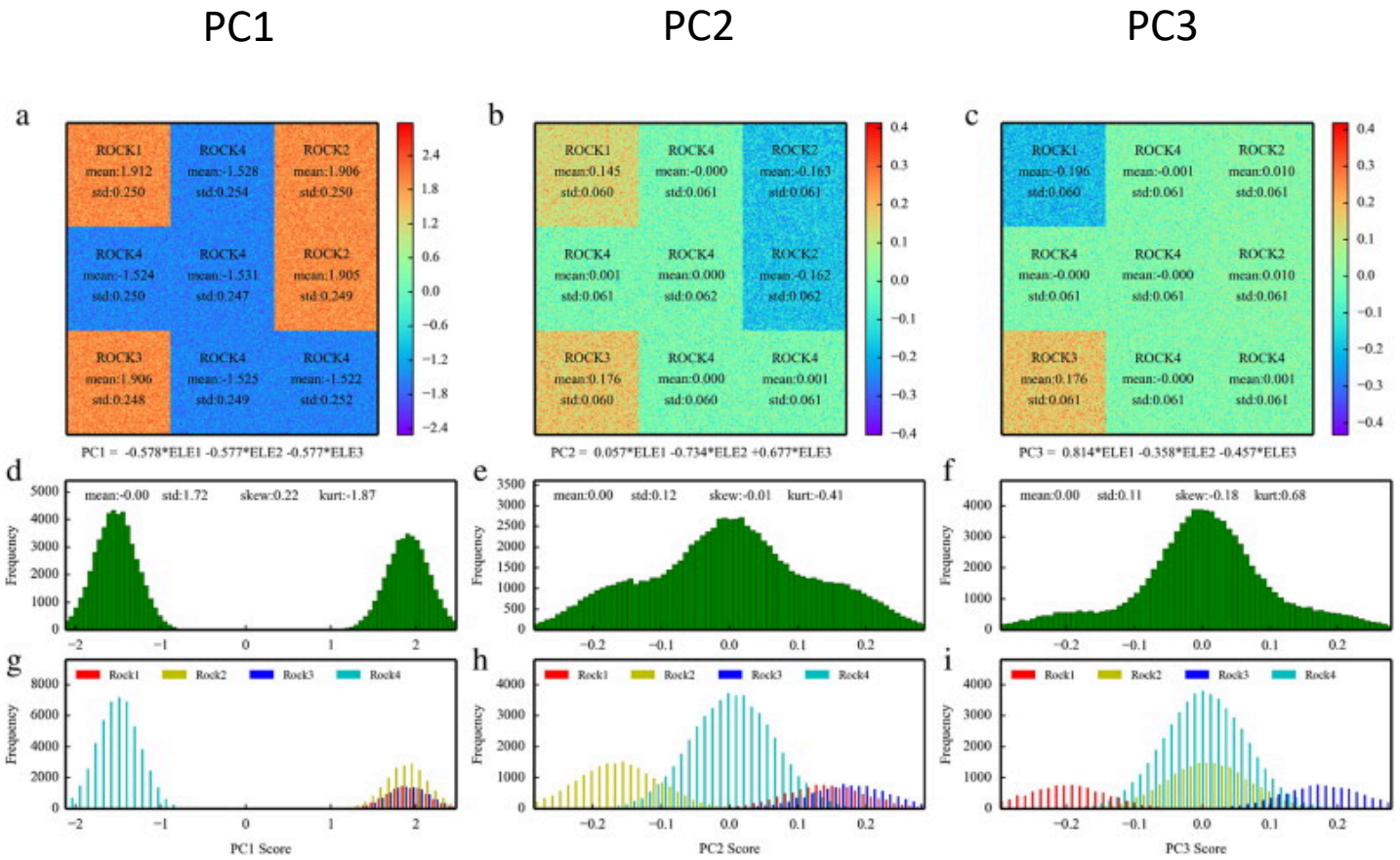
Independent Component Analysis

- ICA finds independent components with non-Gaussian distributions
- ICA is used to separate source signals from mixed signals without or with little prior information about the source signals or the mixing process

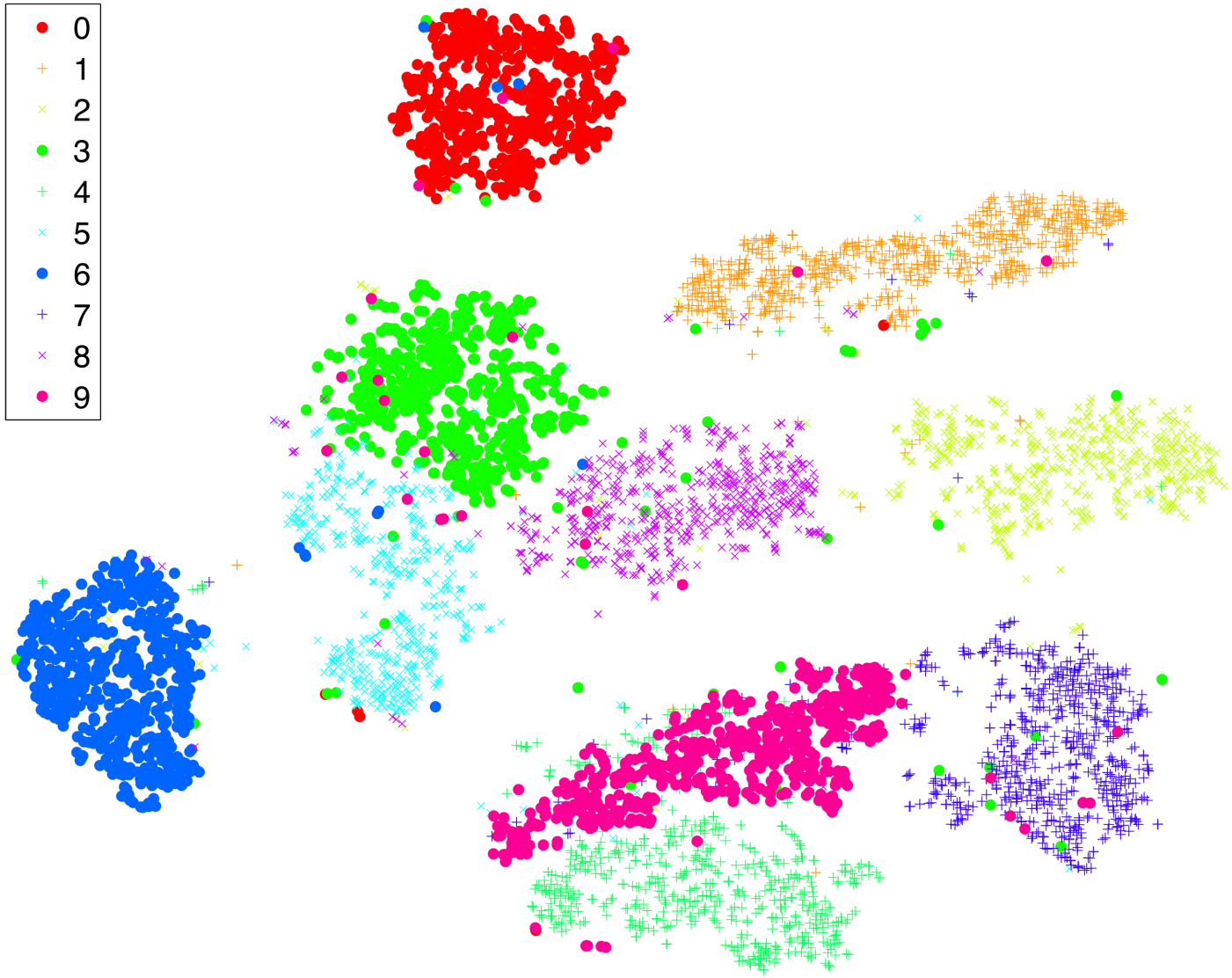


Independent Component Analysis

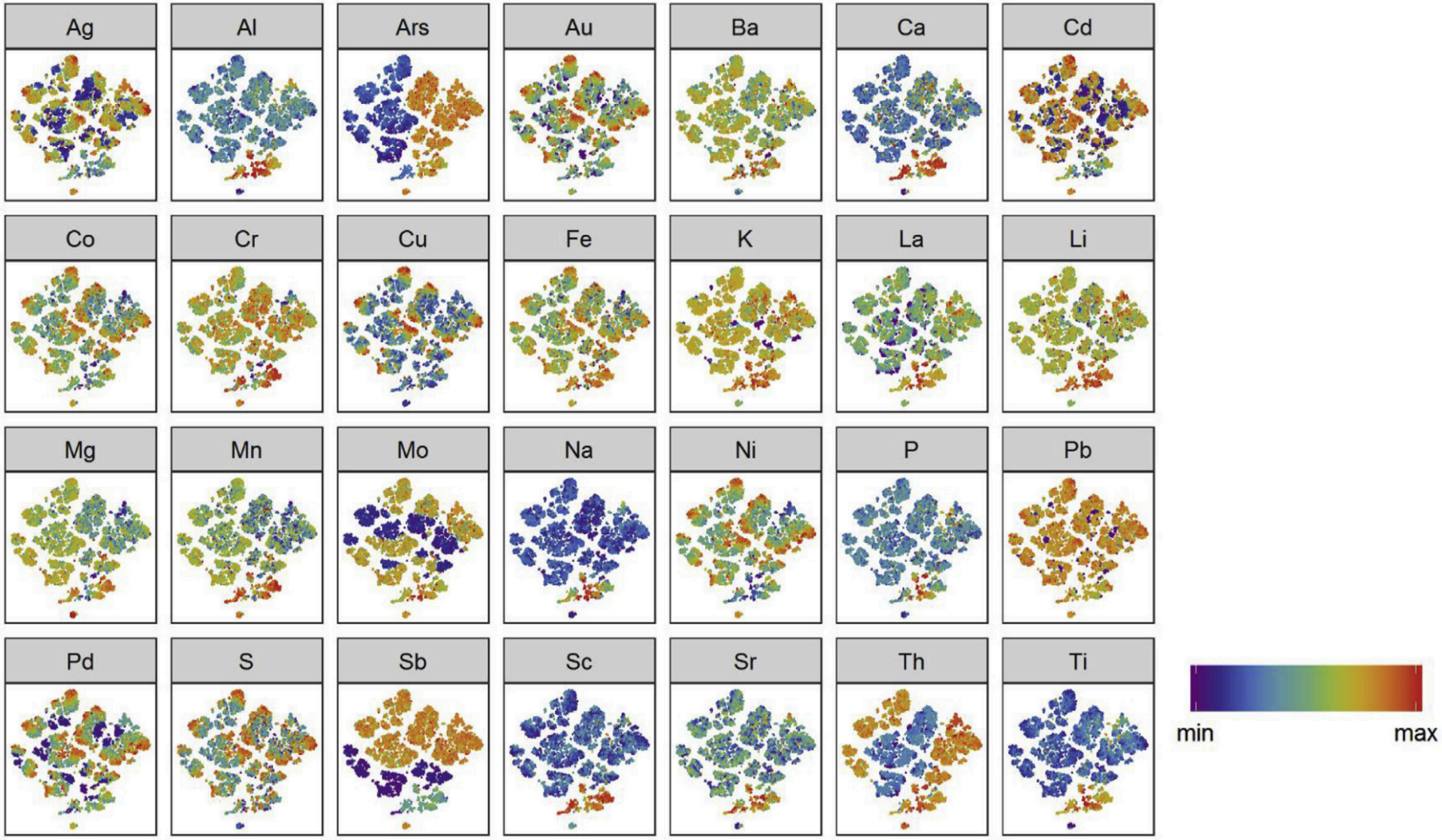
- Comparison with PCA
- Signals are less well separated after the first component
- But ICA is unranked so all components must first be interrogated



t-SNE

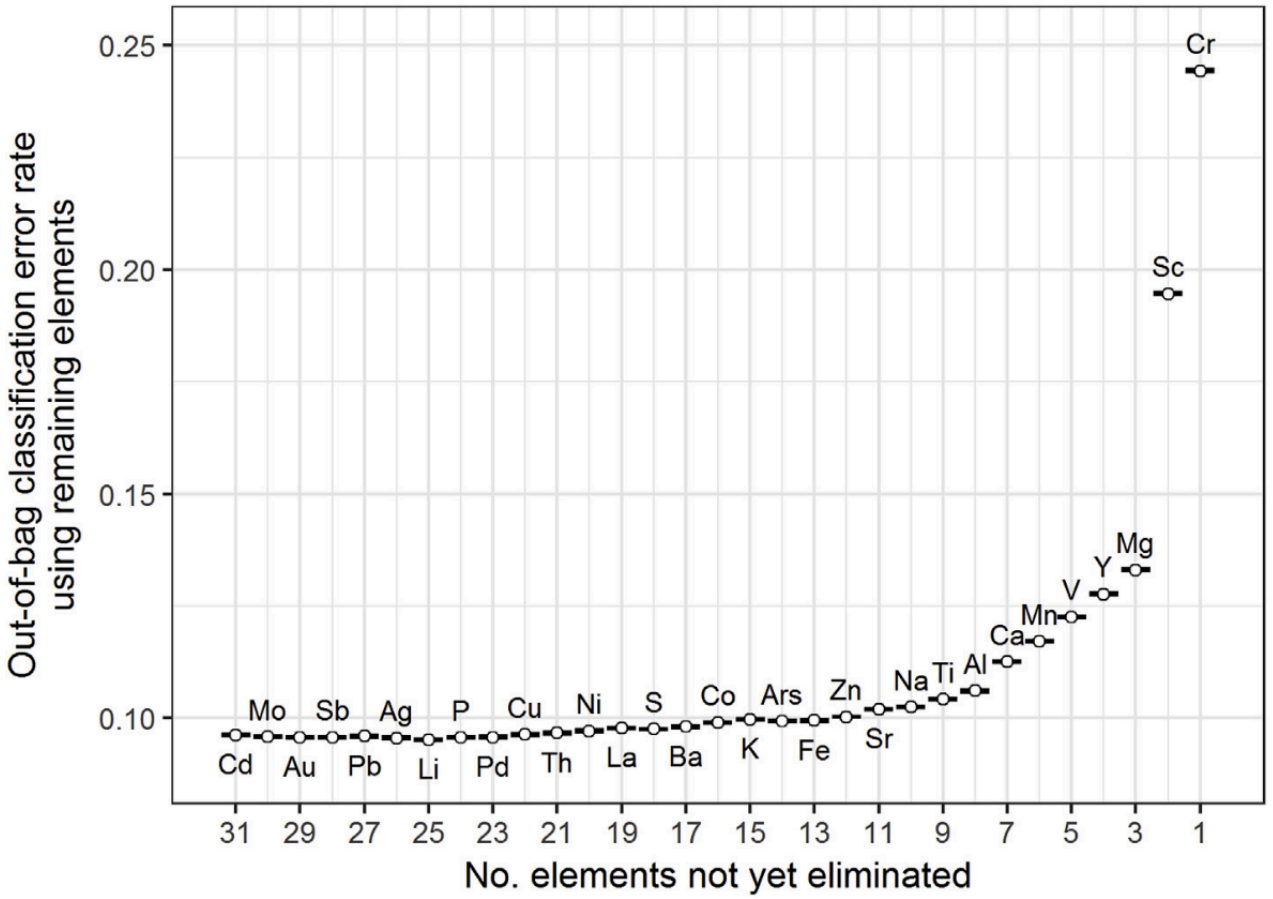


t-SNE


Ag
Al
Ars
Au
Ba
Ca
Cd
Co
Cr
Cu
Fe
K
La
Li
Mg
Mn
Mo
Na
Ni
P
Pb
Pd
S
Sb
Sc
Sr
Th
Ti
min
max

Feature selection

- Now you've reduced the data dimensionality, are the features you have the best predictors?
- PCA uses a scree plot
- In-algorithm selection (e.g. out-of-bag error)
- Do you need to reprocess?



Horrocks et al. (2019)

Feature selection