

CIS501 – Lecture 14

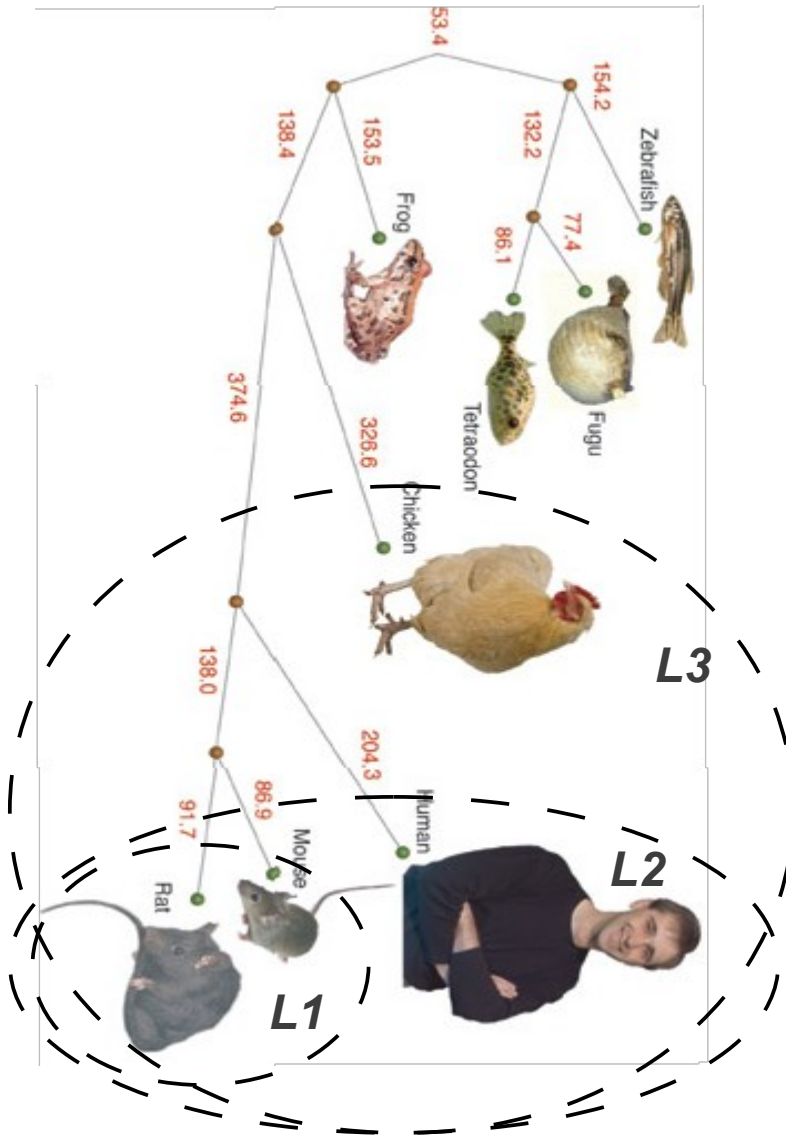
Woon Wei Lee

Fall 2013, 10am-11:15am,
Sundays and Wednesdays

For today:

- Unsupervised learning
 - Hierarchical clustering
 - Visualization

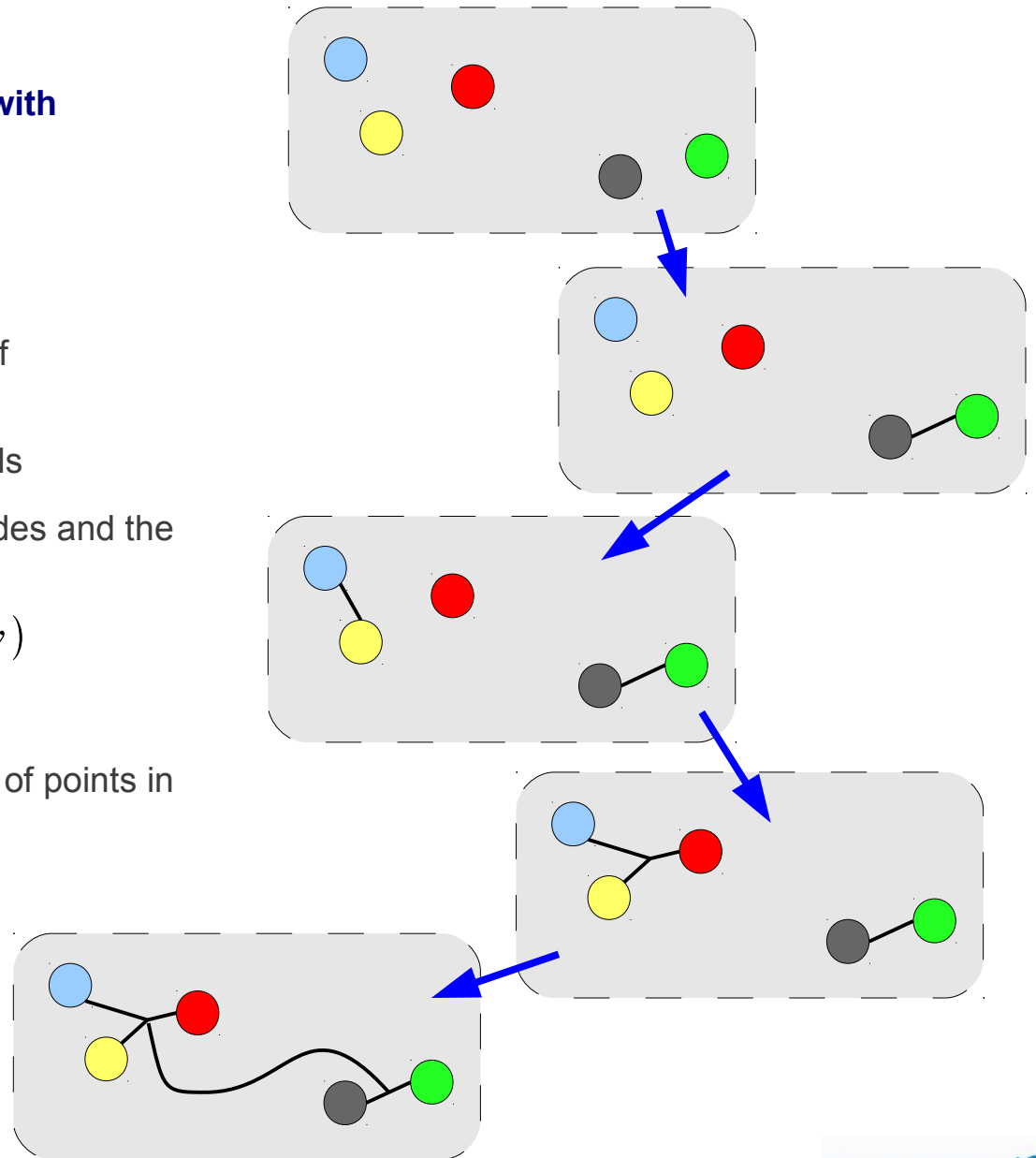
Hierarchical clustering



- **A form of clustering where the nodes are organized hierarchically**
 - Depiction can be rooted - as a “taxonomy”..
 - Or unrooted – simple depiction of relationships between nodes without super/sub-classes
- **Many hierarchical clustering algorithms have their roots in biology/bioinformatics**
 - Inference of *phylogenetic trees* (evolutionary history)
 - But also used for document clustering, image clustering, etc.
 - Typically works using only *distance* information
 - Clusters defined by number of levels in the tree.
- **Two classes**
 - Agglomerative clustering
 - Divisive clustering

Agglomerative clustering: UPGMA

- **Most basic of agglomerative techniques**
- **Stands for “Unweighted Pair Group Method with Arithmetic mean”**
 - “Greedy” algorithm
- **Algorithm:**
 - Inputs: Distance matrix between all pairs of nodes/individuals in the group
 - *Combine nearest pairs of nodes/individuals
 - Recalculate new distances between all nodes and the cluster using:
$$\frac{1}{|A||B|} \cdot \sum_{x \in A} \sum_{y \in B} d(x, y)$$
 - i.e. the average distance between all pairs of points in the two groups
 - Generate new distance matrix
 - Repeat from (*)

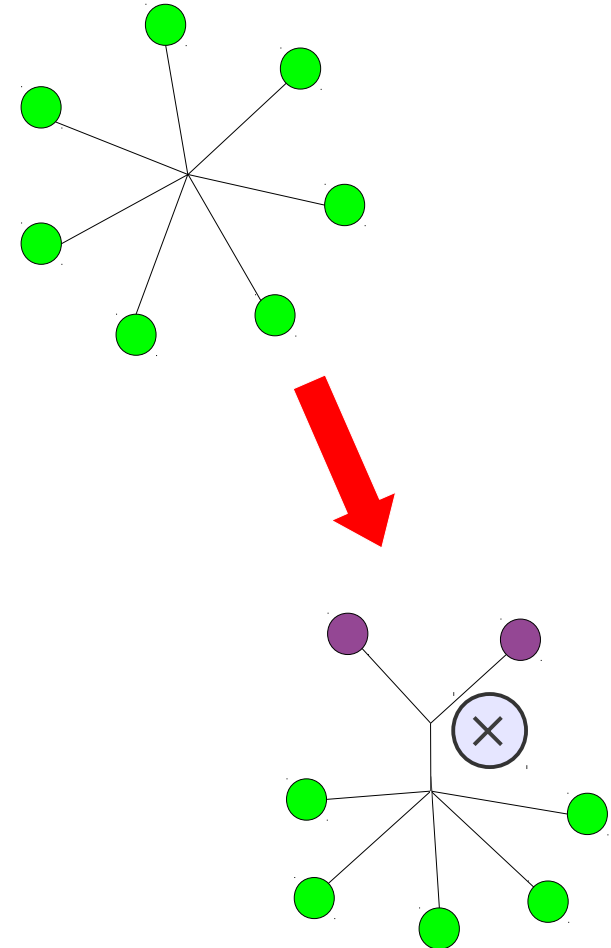


Agglomerative clustering: Neighbour Joining

- **UPGMA – good starting point but poor performance:**
 - “Molecular clock” problem
 - Heuristic technique → unreliable and frequently does not give good results
- **“Neighbour-Joining” algorithm**
 - Generates “unrooted” trees
 - Works via global estimates of branch length
- **Algorithm:**
 - Inputs: Distance matrix between all pairs of nodes/individuals in the group
 - Initiate with “star topology”
 - Choose pair of nodes which minimize:

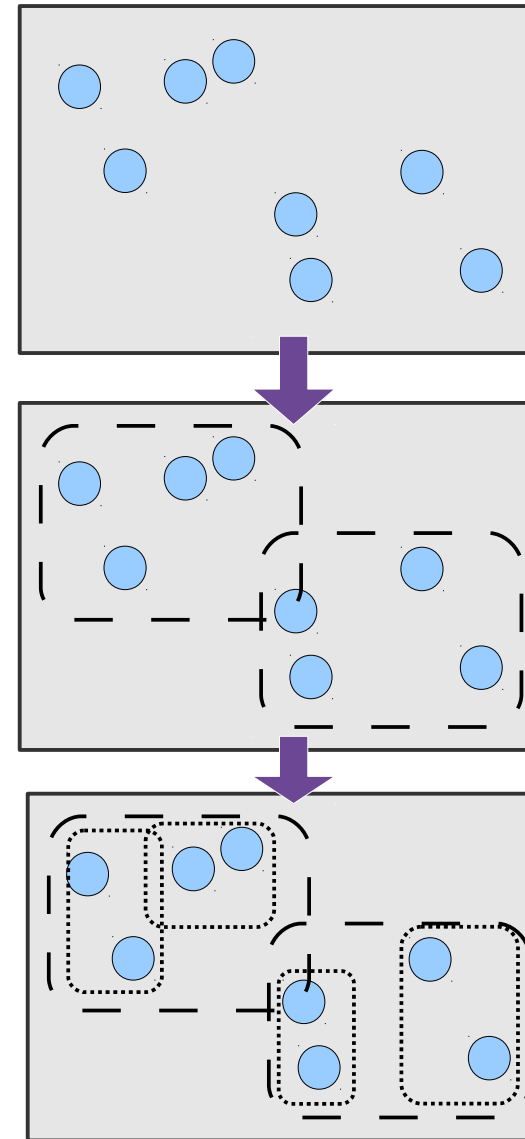
$$s_{12} = \frac{1}{2(N-2)} \sum_{k=3}^N (d(1,k) - d(2,k)) + \frac{1}{2} d(1,2) + \frac{1}{n-2} \sum_{3 \leq i \leq j} d(i,j)$$

- Pair up as shown on right
- Remove pair from distance matrix and replace with branch point (marked “X”)
- Repeat until only binary splits remain
- S_{ij} is the *total branch length* in the tree, which we attempt to minimize (principle of parsimony)



Divisive clustering: Hierarchical k -means algorithm

- **Also known as “top-down” clustering**
 - Any partitional clustering technique can be used as a divisive hierarchical clustering algorithm
- **Advantages:**
 - Many efficient/principled algorithms can be “recycled”
 - Clustering can be for a fixed number of levels
- **Example: “Hierarchical” k -means algorithm:**
 - Perform k -means as per normal
 - For each cluster with at least a minimum number of individuals, cluster using k -means again
 - Iterate until termination



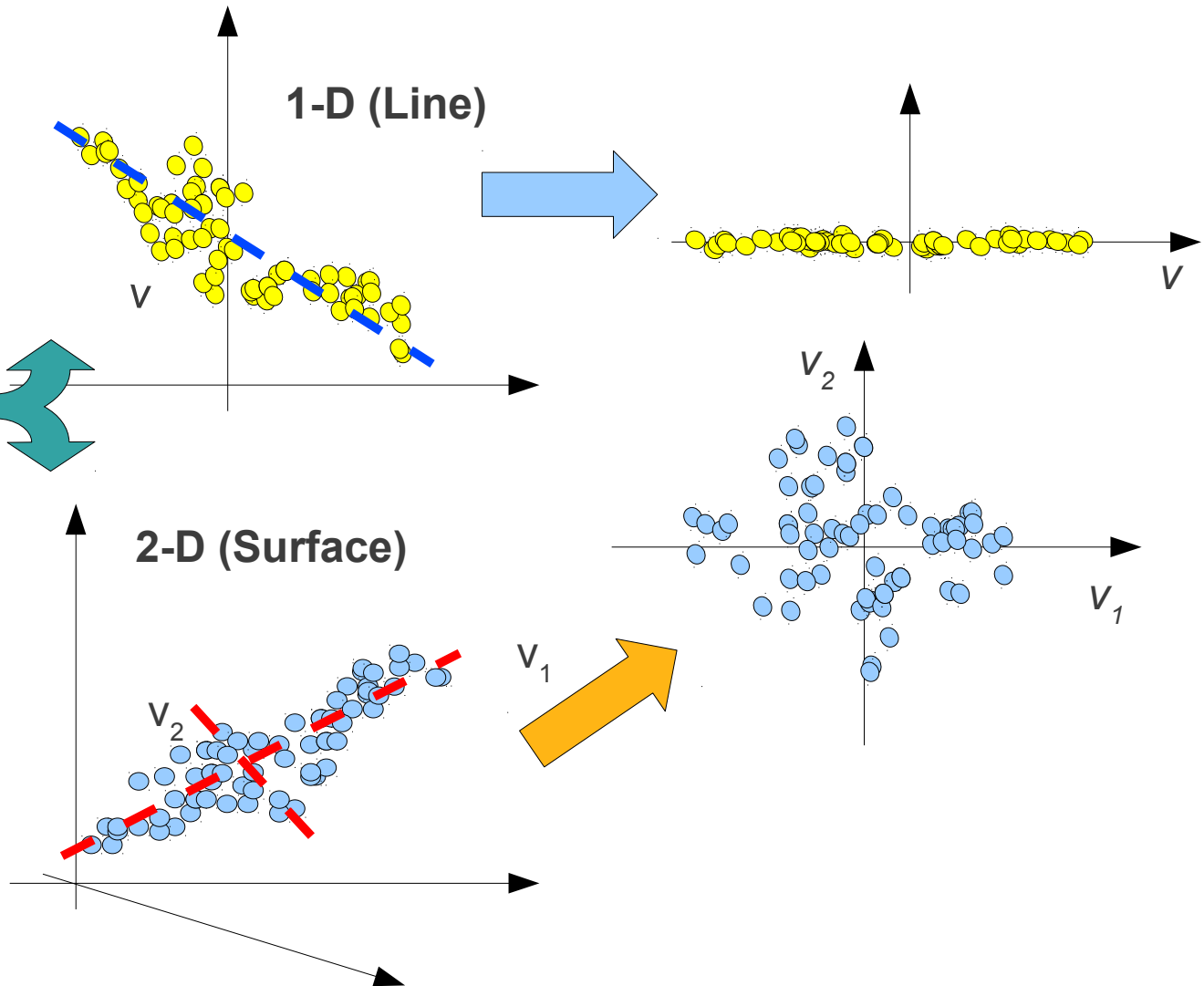
Unsupervised learning: Visualization



- **A cross between feature selection and dimensionality reduction!**
- **Basic motivation**
 - Convert high dimensional data set into a representation that is easily visualized/comprehended
 - Allows for manual clustering, classification, etc.
- **Two basic categories:**
 - Linear projections
 - PCA
 - ICA
 - NMF, etc..
 - Nonlinear projections
 - SOM
 - Multi-dimensional scaling (MDS)
 - Sammon Mapping

Linear visualization techniques

	A	B	C	D	E	F	G	H	I	J	K	L	M
	NAME	SKU	PROD GROUP	WAREHOUSE CODE	DATE	LENGTH	WIDTH	COLOR	WEIGHT LB	PACKAGING	COST	RETAIL	WHOLESALE
3	SHELF 1400	MAK400N	EPOXY WARE	75230		30	14	GREEN EPOXY	7	4	\$ 9.90	\$ 41.04	\$ 16.00
4	SHELF 1406	MAK406N	EPOXY WARE	75230		36	14	GREEN EPOXY	8	4	\$ 10.90	\$ 43.54	\$ 16.00
5	SHELF 1402	MAK402N	EPOXY WARE	75230		42	14	GREEN EPOXY	9.5	4	\$ 12.20	\$ 55.96	\$ 21.20
6	SHELF 1408	MAK408N	EPOXY WARE	75230		48	14	GREEN EPOXY	10.5	4	\$ 12.90	\$ 59.72	\$ 22.40
7	SHELF 1400	MAK400N	EPOXY WARE	75230		60	14	GREEN EPOXY	14	2	\$ 14.40	\$ 69.08	\$ 25.20
8	SHELF 1804	MAK1804N	EPOXY WARE	75230		24	18	GREEN EPOXY	7	4	\$ 10.90	\$ 49.54	\$ 18.00
9	SHELF 1800	MAK1800N	EPOXY WARE	75230		30	18	GREEN EPOXY	8	4	\$ 12.30	\$ 50.10	\$ 21.00
10	SHELF 1806	MAK1806N	EPOXY WARE	75230		36	18	GREEN EPOXY	9.5	4	\$ 12.30	\$ 55.96	\$ 21.20
11	SHELF 1802	MAK1802N	EPOXY WARE	75230		42	18	GREEN EPOXY	11	4	\$ 14.30	\$ 64.22	\$ 24.50
12	SHELF 1808	MAK1808N	EPOXY WARE	75230		48	18	GREEN EPOXY	12	4	\$ 15.30	\$ 69.01	\$ 26.20
13	SHELF 1804	MAK1804N	EPOXY WARE	75230		54	18	GREEN EPOXY	14.5	2	\$ 17.40	\$ 79.02	\$ 30.40
14	SHELF 1800	MAK1800N	EPOXY WARE	75230		60	18	GREEN EPOXY	17	2	\$ 17.90	\$ 80.73	\$ 30.80
15	SHELF 1802	MAK1802N	EPOXY WARE	75230		72	18	GREEN EPOXY	20	2	\$ 20.40	\$ 95.50	\$
16	SHELF 2100	MAK2100N	EPOXY WARE	75230		36	21	GREEN EPOXY	11	4	\$ 14.30	\$ 64.22	\$
17	SHELF 2102	MAK2102N	EPOXY WARE	75230		42	21	GREEN EPOXY	12	4	\$ 16.30	\$ 75.98	\$ 28.00
18	SHELF 2108	MAK2108N	EPOXY WARE	75230		48	21	GREEN EPOXY	14	4	\$ 16.30	\$ 79.31	\$ 28.30
19	SHELF 2100	MAK2100N	EPOXY WARE	75230		60	21	GREEN EPOXY	18	2	\$ 19.40	\$ 88.89	\$ 30.40
20	SHELF 2102	MAK2102N	EPOXY WARE	75230		72	21	GREEN EPOXY	24	2	\$ 23.80	\$ 109.17	\$ 41.60
21	SHELF 2404	MAK2404N	EPOXY WARE	75230		24	24	GREEN EPOXY	9	4	\$ 13.00	\$ 62.39	\$ 23.00
22	SHELF 2400	MAK2400N	EPOXY WARE	75230		30	24	GREEN EPOXY	11	4	\$ 14.00	\$ 65.97	\$ 25.50
23	SHELF 2406	MAK2406N	EPOXY WARE	75230		36	24	GREEN EPOXY	13	4	\$ 15.20	\$ 69.72	\$ 26.80
24	SHELF 2402	MAK2402N	EPOXY WARE	75230		42	24	GREEN EPOXY	15	4	\$ 17.50	\$ 80.73	\$ 30.80
25	SHELF 2408	MAK2408N	EPOXY WARE	75230		48	24	GREEN EPOXY	16	4	\$ 18.90	\$ 85.32	\$ 32.90
26	SHELF 2404	MAK2404N	EPOXY WARE	75230		54	24	GREEN EPOXY	19	2	\$ 21.00	\$ 95.33	\$ 36.75
27	SHELF 2400	MAK2400N	EPOXY WARE	75230		60	24	GREEN EPOXY	21	2	\$ 21.90	\$ 100.00	\$ 38.15
28	SHELF 2402	MAK2402N	EPOXY WARE	75230		72	24	GREEN EPOXY	26	2	\$ 25.30	\$ 117.43	\$ 44.00
29	POST 72	MS072N	EPOXY WARE	75230		72	0	GREEN EPOXY	4	4	\$ 4.50	\$ 20.64	\$ 7.80
30	POST 84	MS084N	EPOXY WARE	75230		84	0	GREEN EPOXY	5	4	\$ 5.40	\$ 24.77	\$ 9.45
31	14 SOL WALL BRKT	VM1400N	EPOXY WARE	75230		0	14	GREEN EPOXY	1.5	2	\$ 5.70	\$ 30.00	\$ 9.00
32	18 SOL WALL BRKT	VM1800N	EPOXY WARE	75230		0	18	GREEN EPOXY	3	2	\$ 5.40	\$ 30.00	\$ 9.40
33	18 TBL WALL BRKT	VM1800N	EPOXY WARE	75230		0	18	GREEN EPOXY	3.5	2	\$ 8.40	\$ 44.00	\$ 17.10



Problem is: how to find the optimal projection line/surface?

Principle Component Analysis (PCA)

- **Basic idea – find direction which capture “most” of the data**

- From figure on right, intuitively, v_a does this better
- One metric – variance of the projection
- Direction of maximal variance - known as the *principle components*

- **Finding the principle components**

- Linear projection defined as:

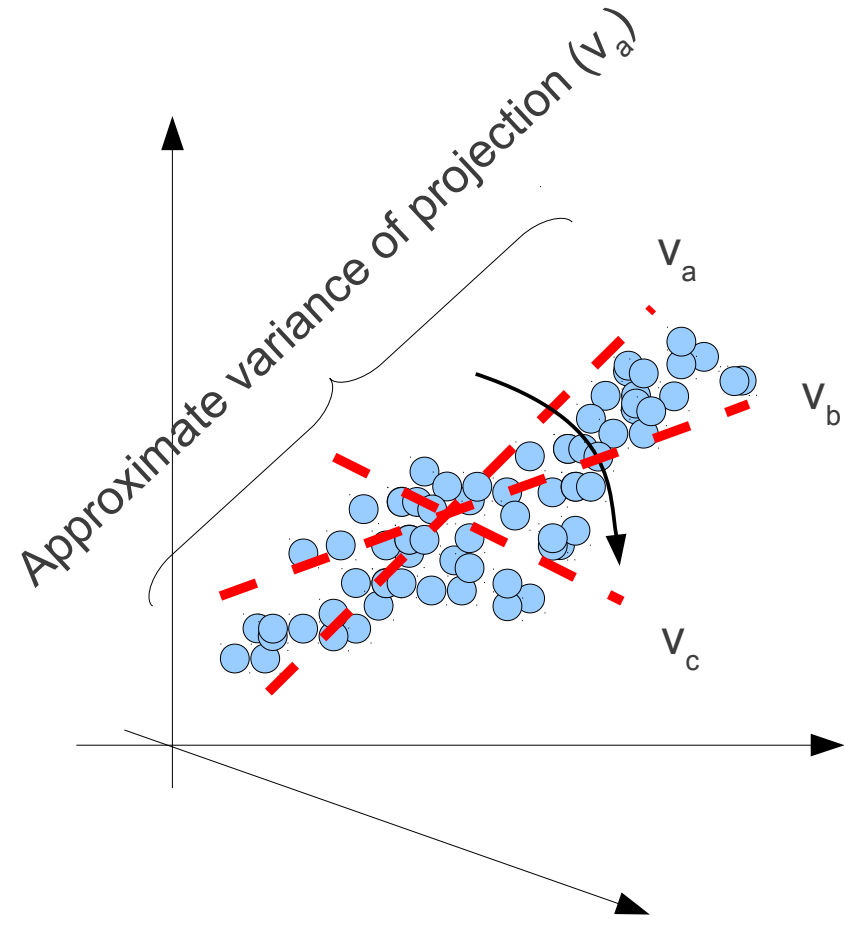
$$x' = v^T x$$

(x' is the transformed variable)

- Variance of projection is given by:

$$\begin{aligned}\sigma'^2 &= x' x'^T \\ &= v^T x (v^T x)^T \\ &= v^T x x^T v = v^T \Sigma v\end{aligned}$$

(where Σ is the data covariance matrix)



Principle Component Analysis (PCA)

- Hence, to find the principle components,


Maximize: $v^T \Sigma v$ (w.r.t. v)

- Trivial solution \rightarrow set v to ∞

- Constraint needed:

Set: $v^T v = 1$

$$L(\sigma, \lambda) = v^T \Sigma v - \lambda (v^T v - 1)$$

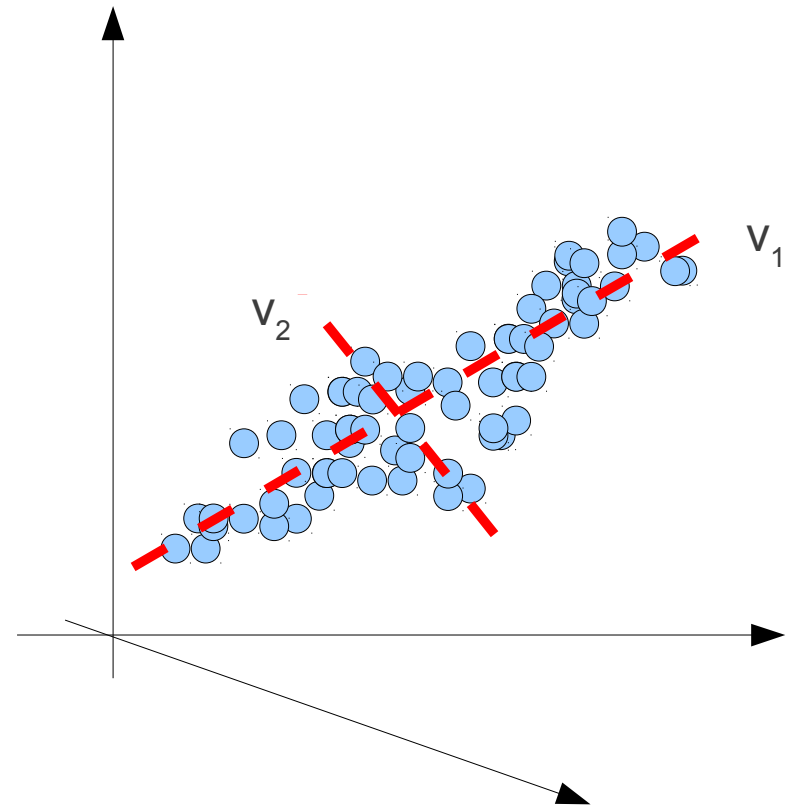
Lagrange multiplier

$$\frac{dL(\sigma, \lambda)}{dv} = 2 \Sigma v - 2 \lambda v = 0$$

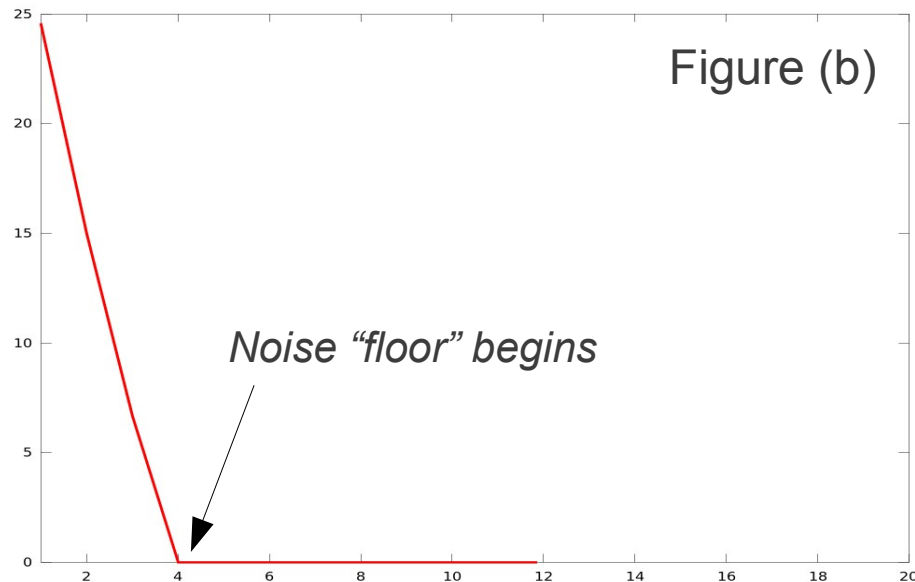
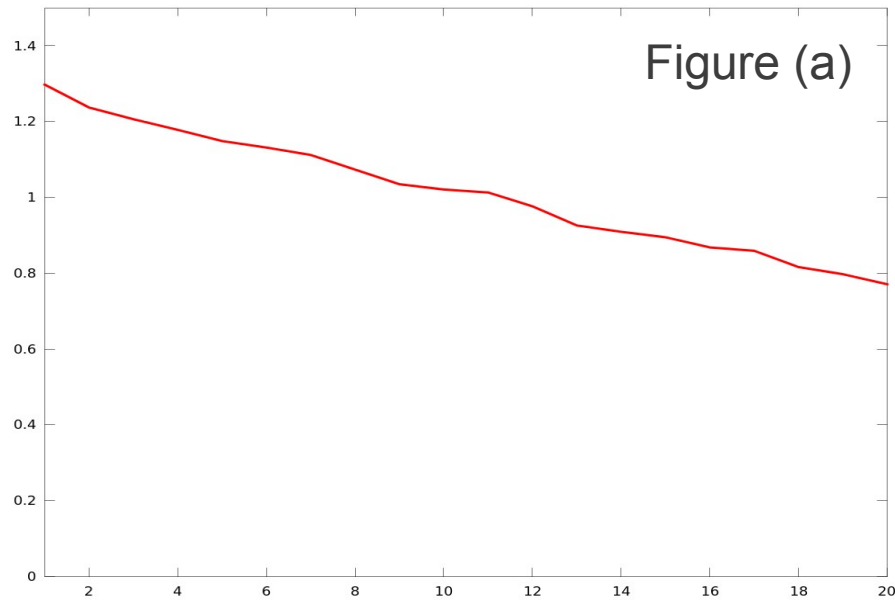
$$\Sigma v = \lambda v$$

- This can be re-written in the form of the “characteristic equation”

- Solution given by the eigenvectors of the covariance matrix
- The λ value gives the variance of the projection in this direction (the eigenvalues of the covariance matrix)



PCA (Cont'd)



i.e. The principle components are given by the *eigenvectors* of the covariance matrices

- For an n -dimensional dataset, there will be n such eigenvectors.
- Eigenvectors are mutually orthonormal
- Matrix of eigenvectors is hence a *rotation matrix*

Project upon a subset of these eigenvectors → dimensionality reduction

- The λ value → the eigenvalues of the covariance matrix
- Sorting these and plotting gives the *singular spectrum (SS)*

Figures (a) and (b):

(a) SS corresponding to 20 dimensional white noise

(b) SS for 3 dimensional white noise embedded in 20 dimensional space

- Note the noise "floor" in figure (b).