

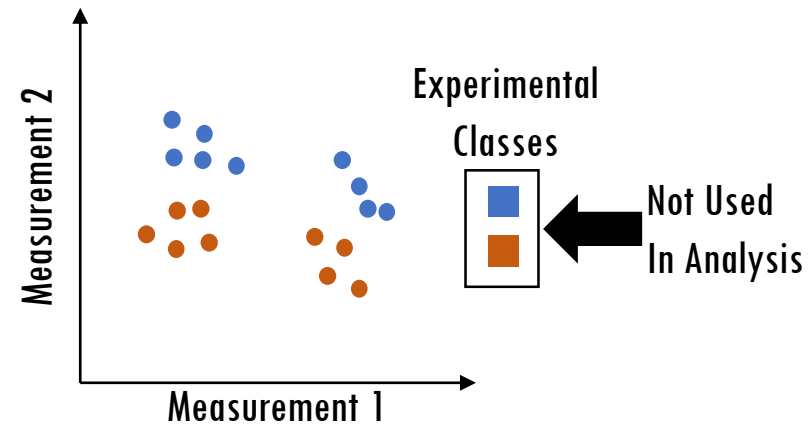
Unsupervised Machine Learning

Satwik Rajaram

Usage Scenario: What's going on with my data?

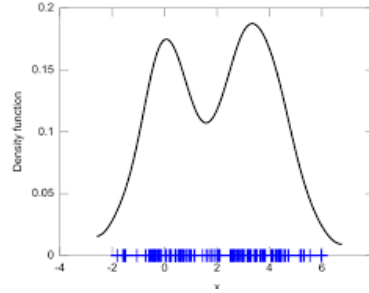
- High dimensional data
- No pre-defined groups to classify points
- Often first pass analysis
- Want to know about structure of data
 - What are the groupings in the data?
 - What are the dominant effects?
- Supervised:
 - Does the data contain this info?
- Unsupervised:
 - What information does the data tell us?
 - Utility is more subjective

Samples	Measurement 1	Measurement 2	Measurement 3	Mea
1	123	0.2	10	
2	324	0.3	13	...
3	3131	0.1	97	
4	146	0.5	5	

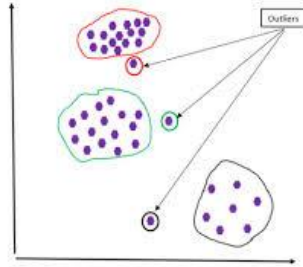


Different Types of Unsupervised Learning

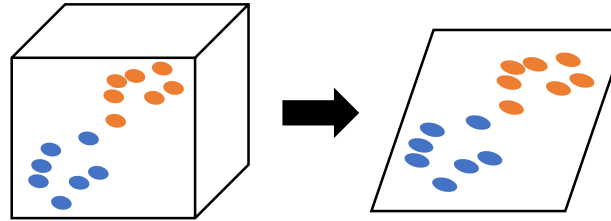
- Density Estimation



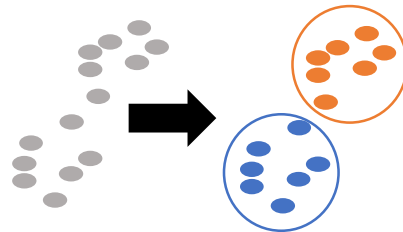
- Outlier Detection



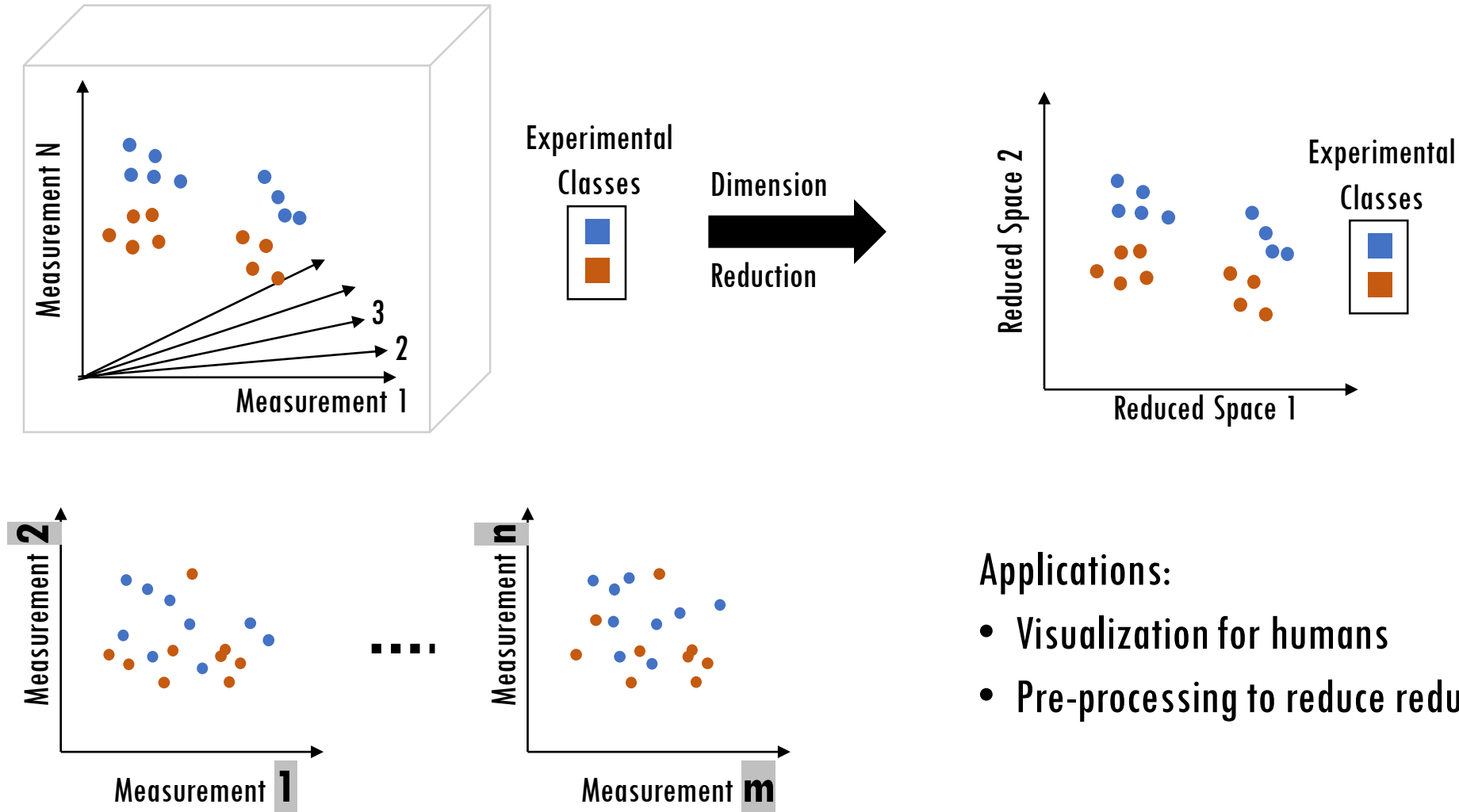
- Dimensional Reduction



- Clustering



Dimension Reduction

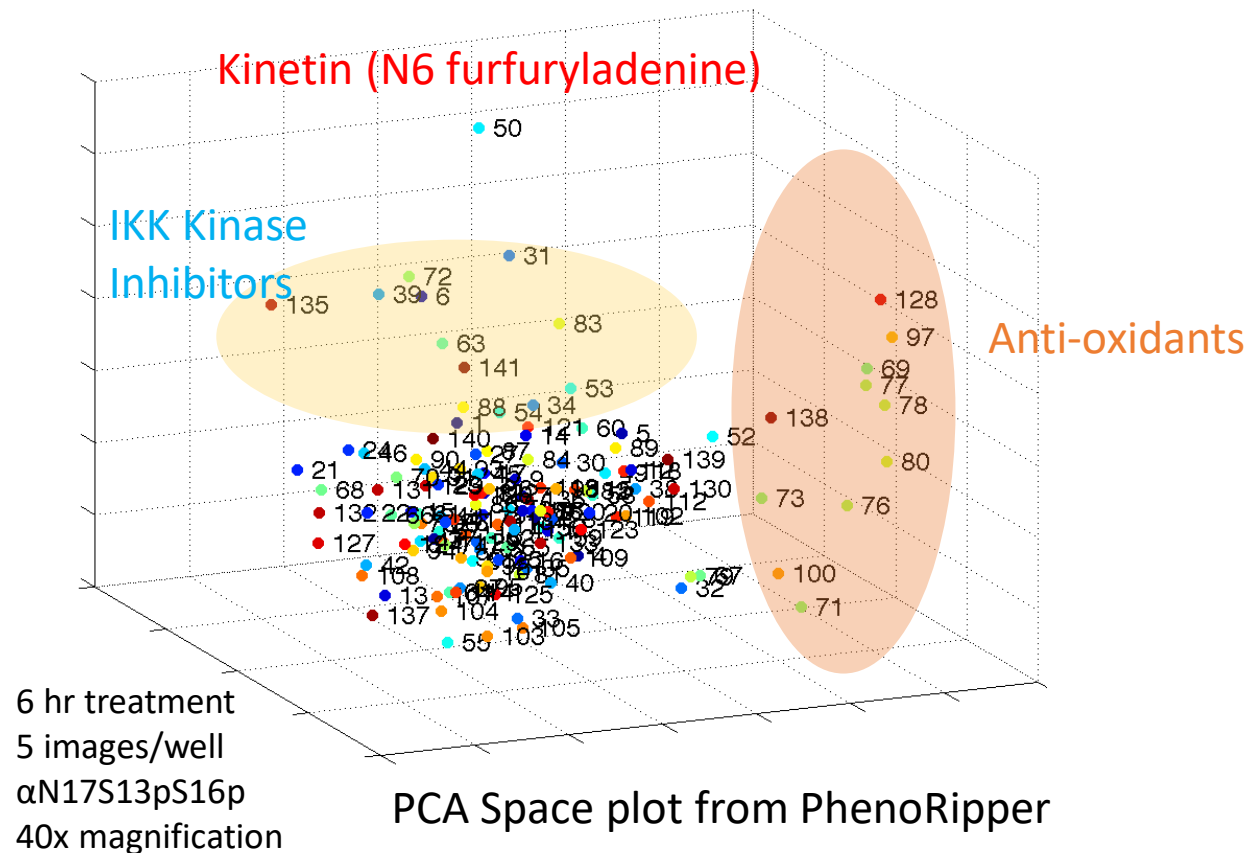


Applications:

- Visualization for humans
- Pre-processing to reduce redundancy/noise

Example from my Work

Natural Compounds Library (N=133) Screen:
System: Mouse striatal neuronal cell line (STHdh)
Readout: phosphorylation in N17 region of huntingtin



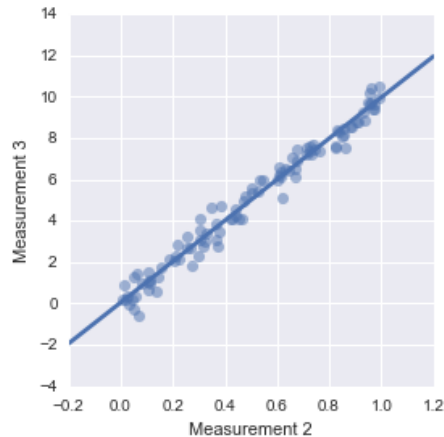
Laura Bowie
@ Truant Lab

Why is dimensional reduction possible?

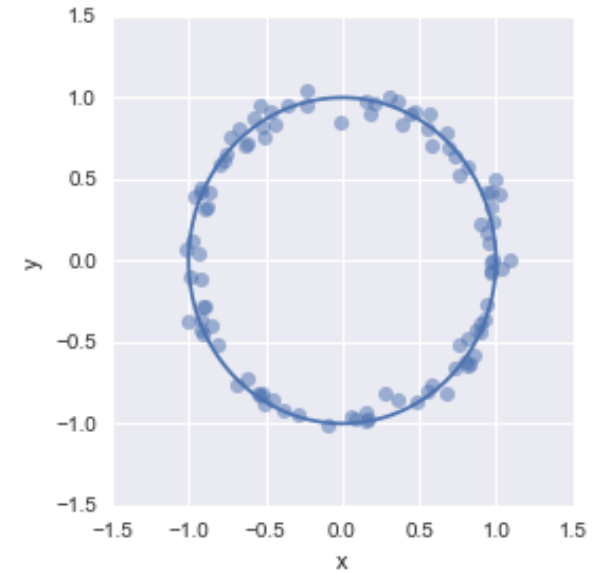
Ans: The whole input space is not used

Correlated variables don't offer independent information

Samples	Measurement 1	Measurement 2	Measurement 3	Measurement 4
1	123	0.2	21	0.01
2	324	0.3	30	0.01
3	3131	0.1	12	0.02
4	146	0.5	49	0.5



Constraints can cause data to inhabit sub-space



Dimension Reduction is not always successful!

Approaches to dimensional reduction

Variable Based

Samples	Measurement 1	Measurement 2	Measurement 3	Measurement 4
1	123	0.2	21	0.01
2	324	0.3	30	0.01
3	3131	0.1	12	0.02
4	125	0.19	25	0.015

Correlated Variables
Form Single Dimension

Examples:

1. Principal Components Analysis
2. Non-negative Matrix Factorization

Distance Based

Samples	Measurement 1	Measurement 2	Measurement 3	Measurement 4
1	123	0.2	21	0.01
2	324	0.3	30	0.01
3	3131	0.1	12	0.02
4	125	0.19	25	0.015

Similar Original Data



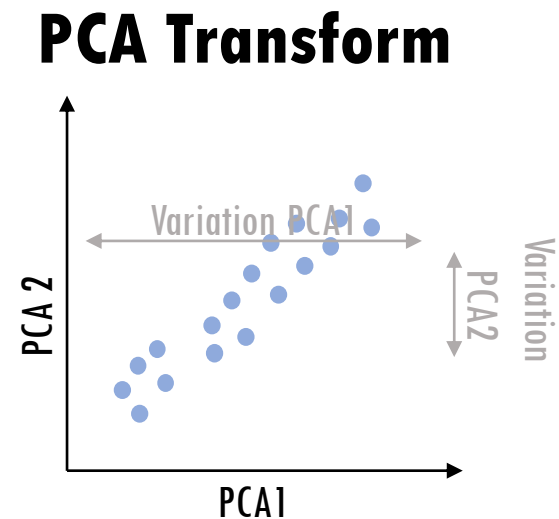
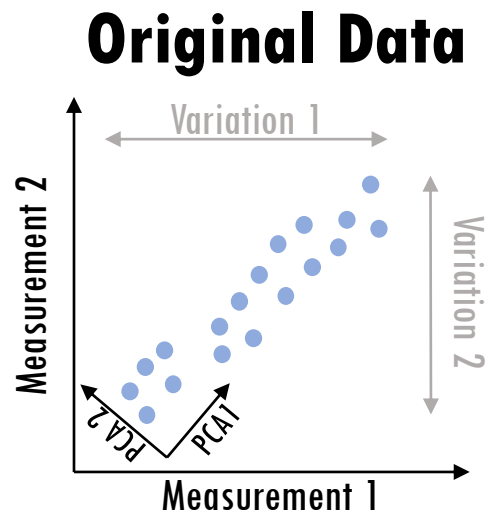
Similar After Reduction

Examples:

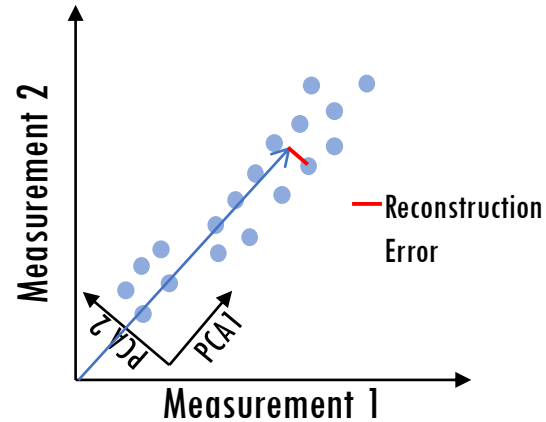
1. Multidimensional Scaling
2. t-Distributed Stochastic Neighbor Embedding

Principal Component Analysis (PCA)

Intuition: Rotate data so most variation is brought to the front



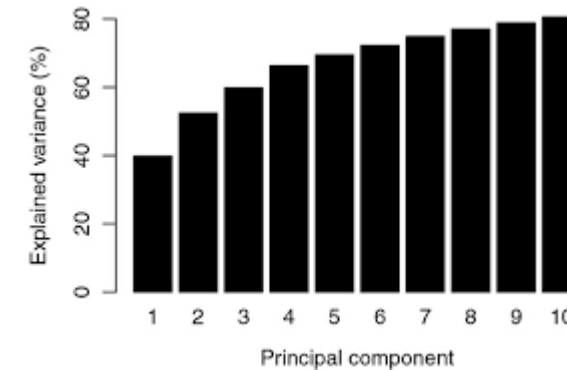
Dimensional Reduction With PCA



Since Total Variance is Constant
Maximizing Variance Captured
Minimizes Reconstruction Error

$$\mathbf{w}_{(1)} = \arg \max_{\|\mathbf{w}\|=1} \{\|\mathbf{X}\mathbf{w}\|^2\} = \arg \max_{\|\mathbf{w}\|=1} \{\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}\}$$

Generalize to More Dimensions



PCA Examples

Genes Mirror Geography in Europe (*Nature* 2009, 456, 98)

1. QUESTION:

What factors
most contribute
to genetic variation
in Europe?

2. EXPR SET-UP:

measurements:

197,146 genes

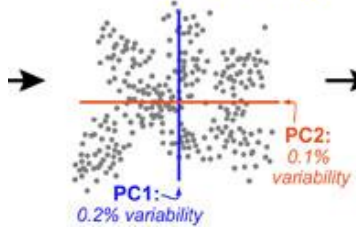
data points:

1,387 Europeans

3. DATA: 197,146D shape



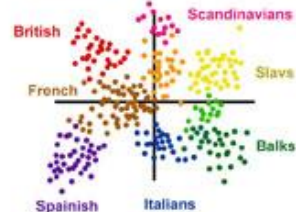
4. PCA: 1D "fit" lines



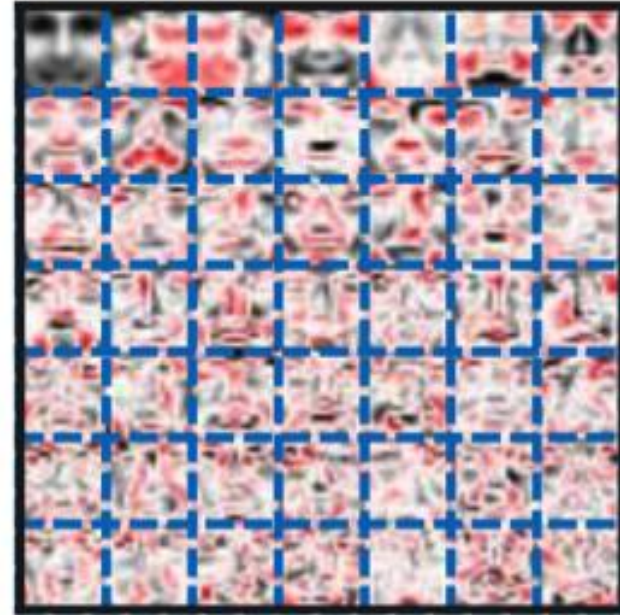
5. HYPOTHESIS?



6. CONFIRMATION:



PCA

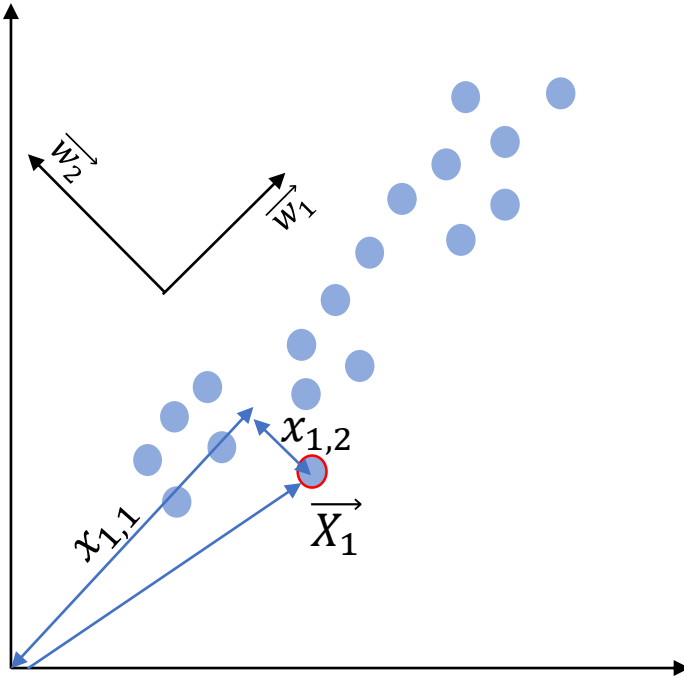


||



×

Generalizing the Idea



1st Data Point $\rightarrow \vec{X}_1 = x_{1,1}\vec{w}_1 + x_{1,2}\vec{w}_2 + x_{1,3}\vec{w}_3 + \dots$

Basis Vectors

Coordinates
Of 1st Data Point

$$\vec{X}_2 = x_{2,1}\vec{w}_1 + x_{2,2}\vec{w}_2 + x_{2,3}\vec{w}_3 + \dots$$

$$\vec{X}_3 = x_{3,1}\vec{w}_1 + x_{3,2}\vec{w}_2 + x_{3,3}\vec{w}_3 + \dots$$

⋮

$$\vec{X}_N = x_{N,1}\vec{w}_1 + x_{N,2}\vec{w}_2 + x_{N,3}\vec{w}_3 + \dots$$

$$\min_w \|X - xw\|$$

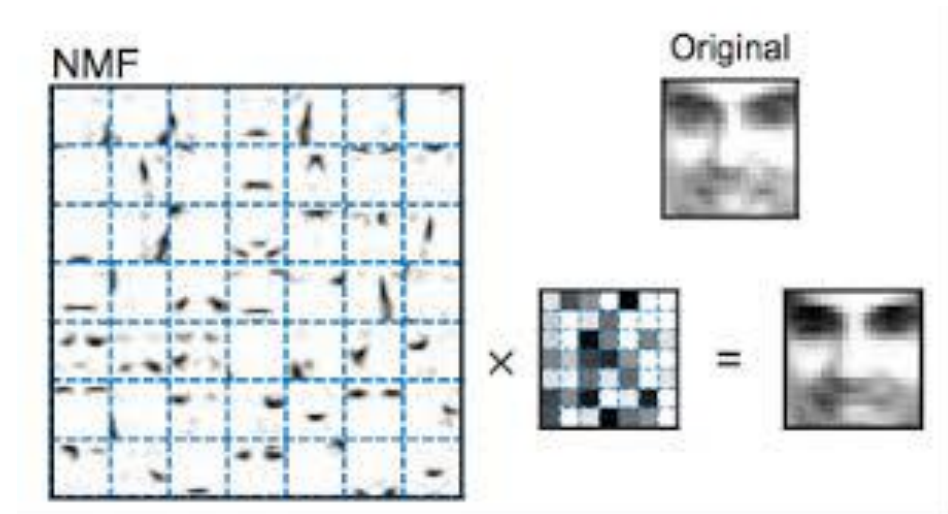
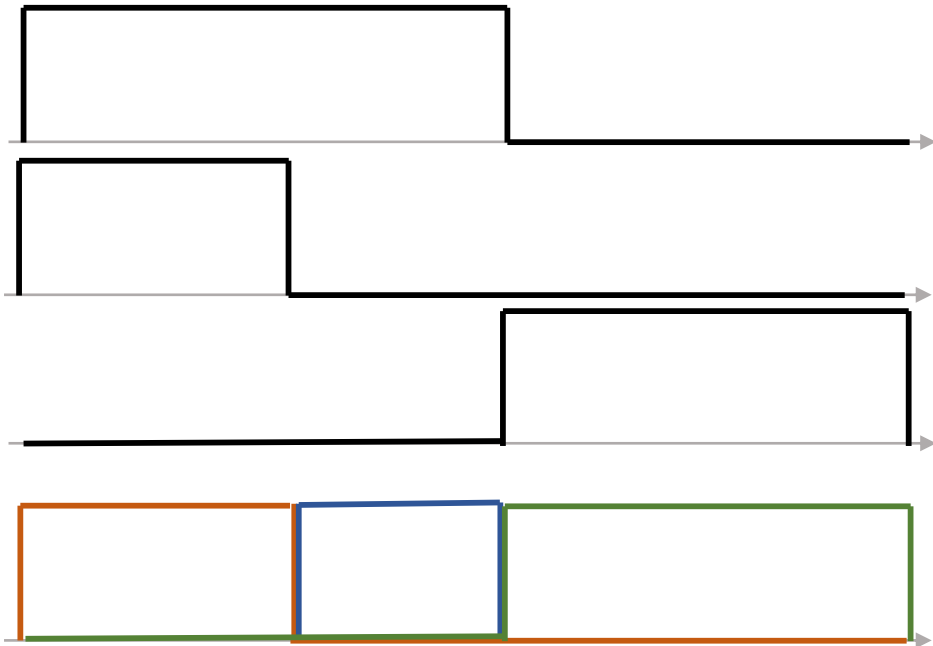
Non-Negative Matrix Factorization

$$\min_w ||X - xw||$$

With constraint: X, x & w are all non-negative

What nNMF giveth, it can't taketh away!

➡ Parts based representation



nNMF Example

BBC

Sign in

News

Sport

Weather

Shop

Earth

Travel

NEWS

HomeVideoWorldUS & CanadaUKBusinessTechScienceStoriesEntertainment

Health

Humans sense 10 basic types of smell, scientists say

By Michelle Roberts
Health editor, BBC News online

19 September 2013

f

🐦

💬


✉

Share

The thousands of aromas humans can smell can be sorted into 10 basic categories, US scientists say.



Prof Jason Castro, of Bates College, and Prof Chakra Chennubhotla, of the



Publish

About

Browse

 OPEN ACCESS

 PEER-REVIEWED

RESEARCH ARTICLE

Categorical Dimensions of Human Odor Descriptor Space Revealed by Non-Negative Matrix Factorization

Jason B. Castro , Arvind Ramanathan, Chakra S. Chennubhotla 

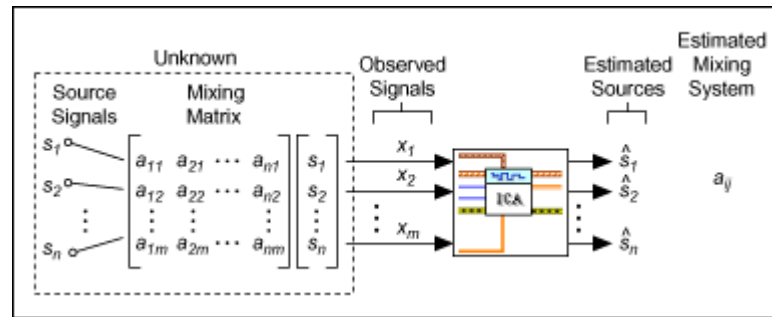
Published: September 18, 2013 • <https://doi.org/10.1371/journal.pone.0073289>

W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
FRAGRANT	WOODY, RESINOUS	FRUITY, OTHER THAN CITRUS	SICKENING	CHEMICAL	MINTY, PEPPERMINT	SWEET	POPCORN	SICKENING	LEMON
FLORAL	MUSTY, EARTHY, MOLDY	SWEET	PUTRID, FOUL, DECAYED	ETHERISH, ANAESTHETIC	COOL, COOLING	VANILLA	BURNT, SMOKY	GARLIC, ONION	FRUITY, CITRUS
PERFUMERY	CEDARWOOD	FRAGRANT	RANCID	MEDICINAL	AROMATIC	FRAGRANT	PEANUT BUTTER	HEAVY	FRAGRANT
SWEET	HERBAL, GREEN, CUT GRASS	AROMATIC	SWEATY	DISINFECTANT, CARBOLIC	ANISE (LICORICE)	AROMATIC	NUTTY (WALNUT ETC)	BURNT, SMOKY	ORANGE
ROSE	FRAGRANT	LIGHT	SOUR, VINEGAR	SHARP, PUNGENT, ACID	FRAGRANT	CHOCOLATE	OILY, FATTY	SULFIDIC	LIGHT
AROMATIC	AROMATIC	PINEAPPLE	SHARP, PUNGENT, ACID	GASOLINE, SOLVENT	MEDICINAL	MALTY	ALMOND	SHARP, PUNGENT, ACID	SWEET
LIGHT	LIGHT	CHERRY (BERRY)	FECAL (LIKE MANURE)	PAINT	SPICY	ALMOND	HEAVY	HOUSEHOLD GAS	COOL, COOLING
COLOGNE	HEAVY	STRAWBERRY	SOUR MILK	CLEANING FLUID	SWEET	CARAMEL	WARM	PUTRID, FOUL, DECAYED	AROMATIC
HERBAL, GREEN, CUT GRASS	SPICY	PERFUMERY	MUSTY, EARTHY, MOLDY	ALCOHOLIC	EUCALIPTUS	LIGHT	MUSTY, EARTHY, MOLDY	SEWER	HERBAL, GREEN, CUT GRASS
VIOLETS	BURNT, SMOKY	BANANA	HEAVY	TURPENTINE (PINE OIL)	CAMPHOR	WARM	WOODY, RESINOUS	BURNT RUBBER	SHARP, PUNGENT, ACID

Independent Component Analysis:

Where the basis vectors are the primary interest

Cocktail party problem:



Similar to previous problem:

- We want signals to be independent
- Orthogonal vectors (PCA) are related through negative information
- Find vectors that have zero mutual information or
- Maximize non-gaussianity

Distance Based Methods

Variable Based

Samples	Measurement 1	Measurement 2	Measurement 3	Measurement 4
1	123	0.2	21	0.01
2	324	0.3	30	0.01
3	3131	0.1	12	0.02
4	125	0.19	25	0.015

Correlated Variables
Form Single Dimension

Examples:

1. Principal Components Analysis
2. Non-negative Matrix Factorization

Distance Based

Samples	Measurement 1	Measurement 2	Measurement 3	Measurement 4
1	123	0.2	21	0.01
2	324	0.3	30	0.01
3	3131	0.1	12	0.02
4	125	0.19	25	0.015

Similar Original Data



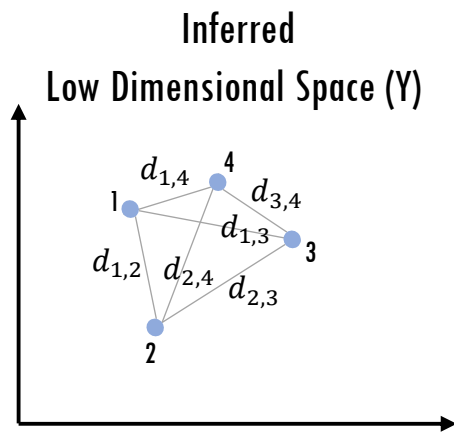
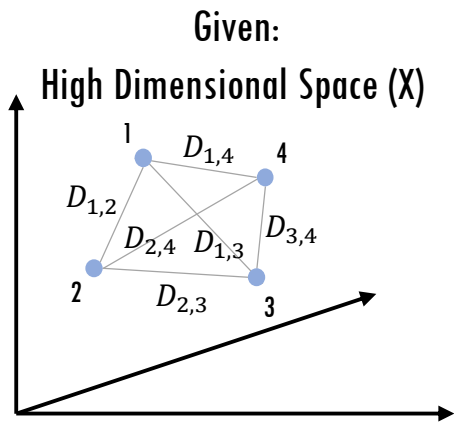
Similar After Reduction

Examples:

1. Multidimensional Scaling
2. t-Distributed Stochastic Neighbor Embedding

Intuitive Idea:

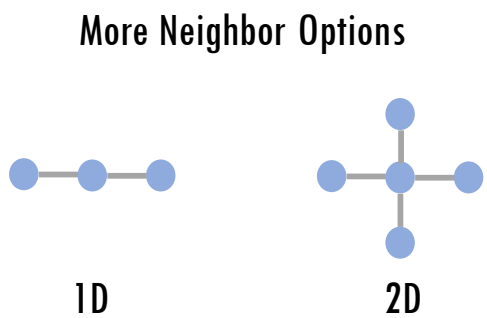
Get distances to “agree” low and high dimensions



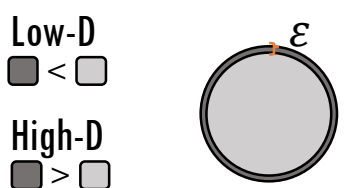
Naïve Implementation: Get $D = d$ numerically

$$\min_Y \sum_{i,j} \|D_{i,j} - d_{i,j}\|$$

Challenge: Points are distributed fundamentally different in higher dimensions!



Most Volume in Shell in High-Dimensions

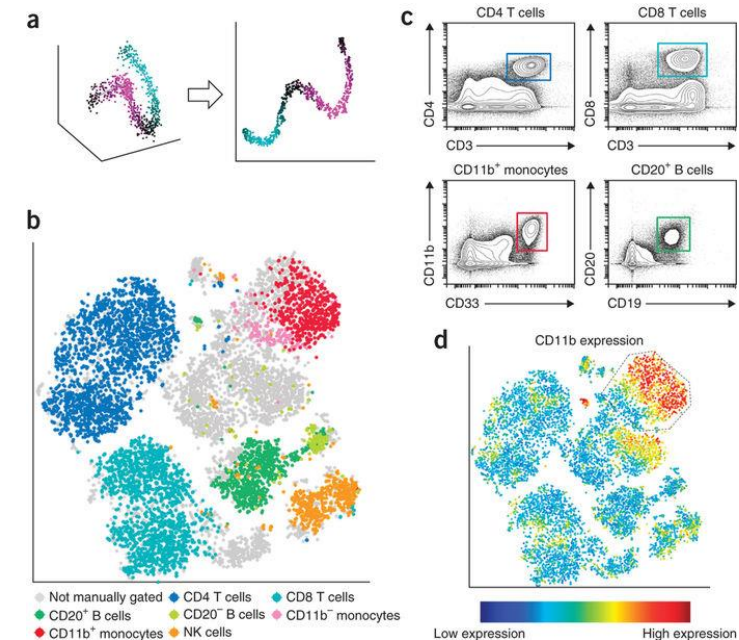


Non-metric Multidimensional Scaling

- Don't require Distance to have same values, but satisfy ordering inequalities
- If $D_{a,b} < D_{m,n}$ then $d_{a,b} < d_{m,n}$
- Effectively try to get rankings $D_{i,j}$ and $d_{i,j}$ to agree
- Computationally intensive for large number of points
- Considers both small and large scale structure

t-Distributed Stochastic Neighbor Embedding (tSNE)

- It's the new hotness! Especially in single cell profiling
- Seems to work well for identifying clusters of similar cells
- Intuitive idea
 - Neighbors in Original Space \Leftrightarrow Neighbors In Embedded Space
 - Maximize agreement in probability of being neighbors
 - Define probabilities differently in original & embedded
 - Longer tail in embedded (low) dimension to account for smearing
- Better at preserving short range than long range information
- Has free parameter related to neighborhood size
- Faster for large data sets than MDS
- Is stochastic (different runs produce differing results)
- Fine for visualization, don't treat results as true, or for pre-processing



tSNE: Math

High Dimensional Space (X)

Probability of being neighbors

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)},$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

Point specific sigma:

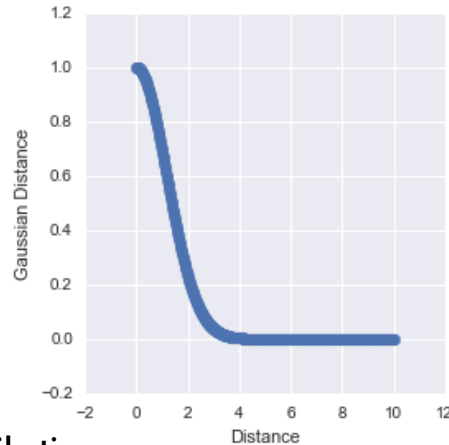
Chosen so points have similar neighbor distributions



σ_i chosen so that input parameter is perplexity

$$2^{-\sum_j p_{i,j} \log_2(p_{i,j})}$$

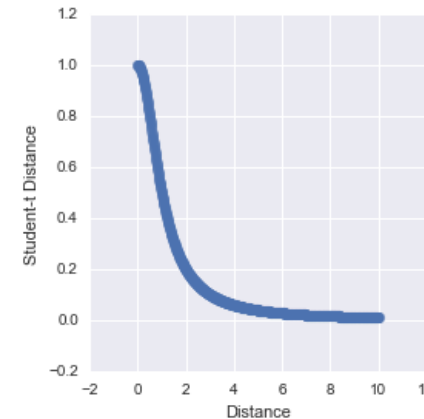
Is equal to specified value



Low Dimensional Space (Y)

Probability of being neighbors

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|\mathbf{y}_i - \mathbf{y}_k\|^2)^{-1}}$$

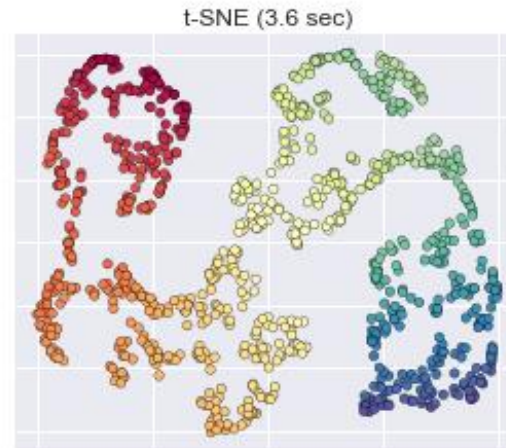
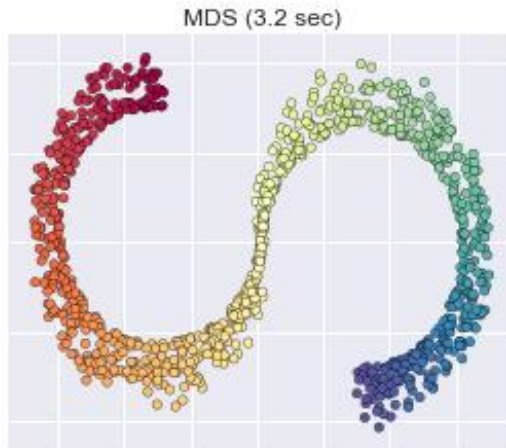
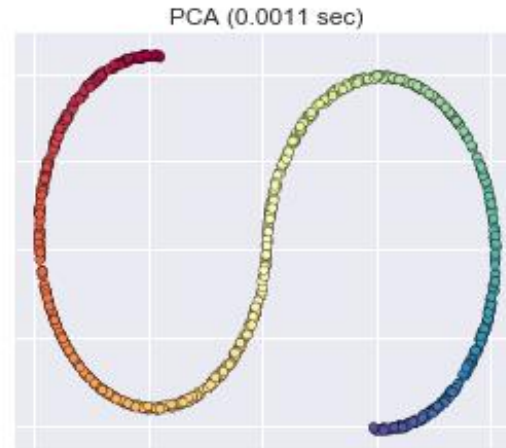
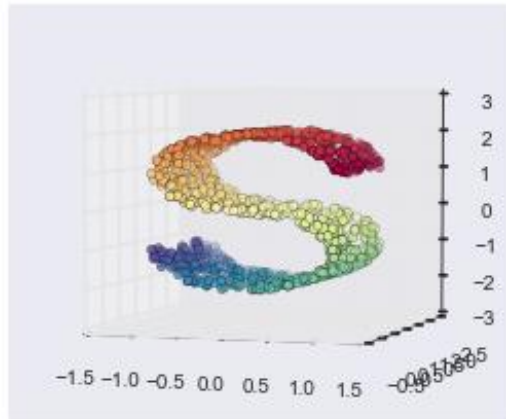


Agreement in Prob
Of Being
Neighbors

Minimize $KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$

Comparison

Manifold Learning with 1000 points, 20 neighbors



- **Linear Variable Models Like PCA:**
 - Faster for large number of points
 - Co-ordinates are interpretable
 - Less dimensional reduction ability
- **Non-linear distance based:**
 - Slower
 - More powerful dimensional reduction
 - Co-ordinates not interpretable