Genomic Data Visualization

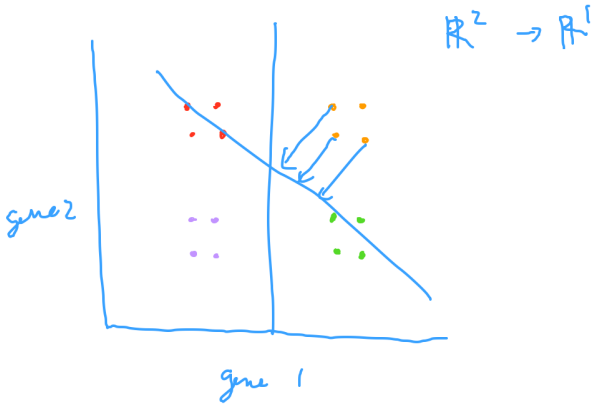
Feb 4, 2022

Return to campus:
- I would like us to prepare to return to in person instruction
-> an immediate return to campus may present some challenges for students who need commute

We will plan to move to in person in 2 weeks time (2/21/22)
-> enough time for those to figure commutes
- enough time for most of educational content to be recorded
-> once we return to inperson, white board notes and slides will still available, however, this type of recording will liekly no longer be available

Continuing with dimensionality reduction
-> last class: linear dimensionality reduction (interpret PCs as linear combinations of genes)
-> this class: non-linear dimensionality reduction -> t stochastic neighbor embedding (tSNE)

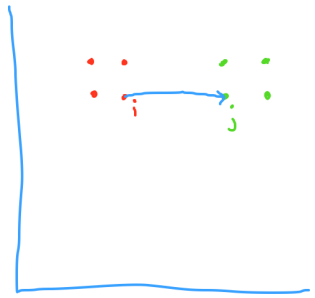
Motivate



what I get with linear dimensionality reduction

what I want

tSNE ex. $\mathbb{R}^2 \to \mathbb{R}^1$



1) compute a scaled similarity between every pair of points in high dimensional space

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

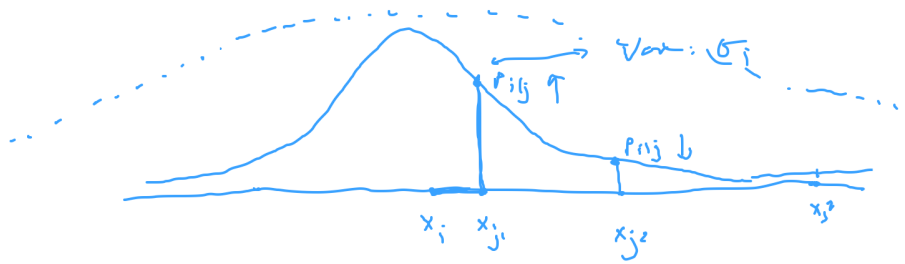
$$p_{i|j} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x - x_j\|^2 / 2\sigma_i^2\right)}$$

hyperparam: perplexity

effective number of neighbors

thing to try out: how does perplexity actually influence data visualization

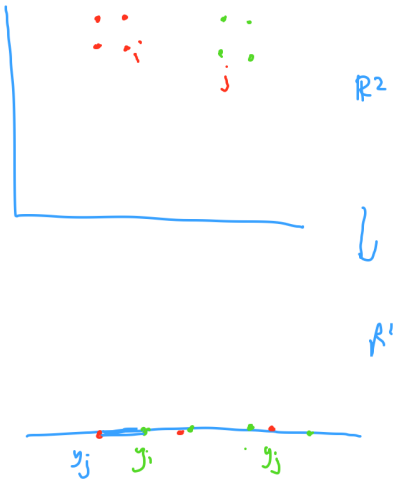
→ gaussian distribution



if x_i is close to x_j → $\|x_i - x_j\|^2$ small

→ $\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)$ big

summarize:
i and j are close, then similarity between is high

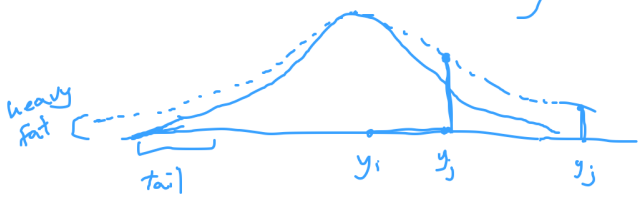


2) randomly squish all our points into a low dimensional space and compute another scaled similarity in this low dimensional space

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_k \sum_{l \neq k} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

↳ t - distribution

(heavy tail)



if y_i and y_j are close:

$\|y_i - y_2\|^2$ small

$\frac{1}{\text{small}} > \frac{1}{\text{big}}$

→ q_{ij} bigger

now we've computed for every pair of i and j a scaled similarity pij in the high dimensional space and qij in the low dimensional space

what tSNE is really trying to do is move points around in this low dimensional space (Q) in order to minimize

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

(p_{ij}: >0)

p_{ij} is fixed

find a Q for which KL is minimized

$\mathbb{R}^N \to \mathbb{R}^M \quad M < N$

$\mathbb{R}^2 \to \mathbb{R}^1$

if p_{ij} is large → i and j are similar/close $\mathbb{R}^N$

if q_{ij} is large → i and j are close $\mathbb{R}^M$

if p_{ij} is small → i and j are far $\mathbb{R}^N$

if q_{ij} is small → i and j are far $\mathbb{R}^M$

if P_{ij}	is large	~~and~~ q_{ij}	is large	→	contribution to KL? ~0
if P_{ij}	is small	q_{ij}	small	→	~0
	large		small	→	KL ↑↑ large
	small		large	→	KL ↑↑ small

If two cells i and j are close together in high dimensional space (pij is large)
-> find a low dimensional embedding where cell i and j can still be close to each other (qij is also large)
If two cells i and j are far apart from each in high dimensional space (pij is small)
-> ideally we want to find a low dimensional embedding where cell i and j can still be far apart (qij is also large)

$\neq$

$$KL(P \| Q) = \sum_{i \neq j} P_{ij} \log\left(\frac{P_{ij}}{q_{ij}}\right)$$

Most likely we can't find a lower dimensional embedding that perfectly maintains all cell cell similarities
-> we have to make some sacrifices

some cells that are close together in high dimensional may need to be farther apart in lower dimensional space

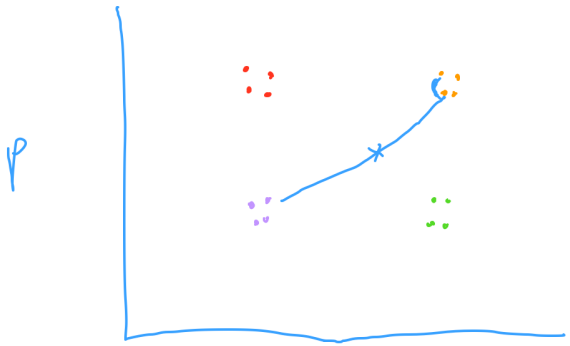
↑ KL (1)

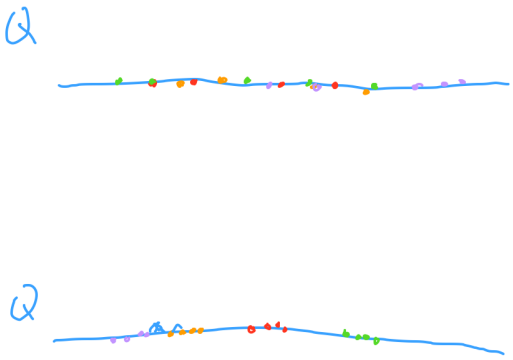
some cells that are far apart in high dimensional space may need to be closer together in lower dimensional space

↓ KL (2)

Since we're trying to minimize the KL divergence…

tSNE tries to keep cells that are close together in high dimensional space close together in low dimensional space
-> but cells that are far apart in high dimensional space, it's not as important that we maintain them as being far apart in low dimensional space





but to say purple cells are more similar to orange cells than red cells would be incorrect