CSC321:

Extra Lecture (not on the exam)
Non-linear dimensionality reduction

Geoffrey Hinton

Linear methods of reducing dimensionality

- PCA finds the directions that have the most variance.
  - By representing where each datapoint is along these axes, we minimize the squared reconstruction error.
  - Linear autoencoders are equivalent to PCA
- Multi-Dimensional Scaling arranges the low-dimensional points so as to minimize the discrepancy between the pairwise distances in the original space and the pairwise distances in the low-D space.

high-D distance, low-D distance

$$Cost = \sum_{ij} \left(\| \mathbf{x}_i - \mathbf{x}_j \| - \| \mathbf{y}_i - \mathbf{y}_j \| \right)^2$$

- MDS is equivalent to PCA (if we normalize the data right)

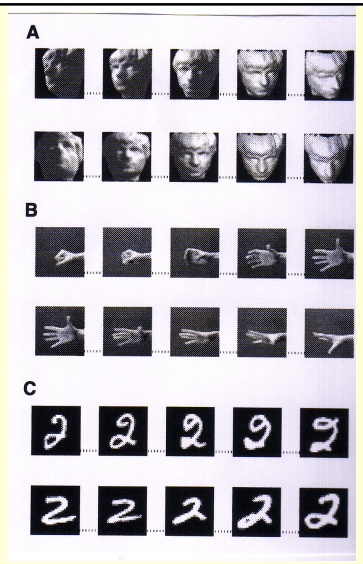
Other non-linear methods of reducing dimensionality

- Non-linear autoencoders with extra layers are much more powerful than PCA but they are slow to optimize and they get different, locally optimal solutions each time.
- Multi-Dimensional Scaling can be made non-linear by putting more importance on the small distances. An extreme version is the Sammon mapping:

high-D distance, low-D distance

$$Cost = \sum_{ij} \left(\frac{\| \mathbf{x}_i - \mathbf{x}_j \| - \| \mathbf{y}_i - \mathbf{y}_j \|}{\| \mathbf{x}_i - \mathbf{x}_j \|} \right)^2$$

- Non-linear MDS is also slow to optimize and also gets stuck in different local optima each time.



Linear methods cannot interpolate properly between the leftmost and rightmost images in each row.

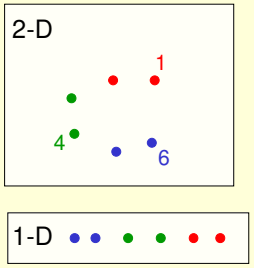
This is because the interpolated images are NOT averages of the images at the two ends.

The method shown here does not interpolate properly either because it can only use examples from the training set. It cannot create new images.

IsoMap: Local MDS without local optima

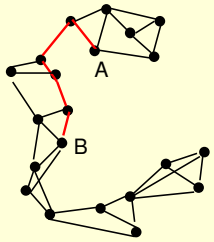
- Instead of only modeling local distances, we can try to measure the distances along the manifold and then model these intrinsic distances.
  - The main problem is to find a robust way of measuring distances along the manifold.
  - If we can measure manifold distances, the global optimisation is easy: It's just global MDS (i.e. PCA)

If we measure distances along the manifold, d(1,6) > d(1,4)



How Isomap measures intrinsic distances

- Connect each datapoint to its K nearest neighbors in the high-dimensional space.
- Put the true Euclidean distance on each of these links.
- Then approximate the manifold distance between any pair of points as the shortest path in this graph.



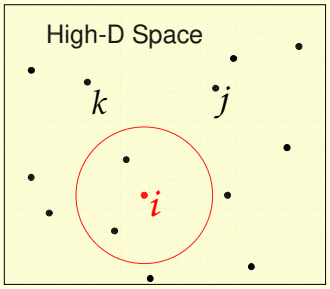
Using Isomap to discover the intrinsic manifold in a set of face images

A probabilistic version of local MDS

- It is more important to get local distances right than non-local ones, but getting infinitessimal distances right is not infinitely important.
  - All the small distances are about equally important to model correctly.
  - Stochastic neighbor embedding has a probabilistic way of deciding if a pairwise distance is "local".

Computing affinities between datapoints

- Each high dimensional point, i, has a conditional probability of picking each other point, j, as its neighbor.
- The conditional distribution over neighbors is based on the high-dimensional pairwise distances.



$$p_{j|i} = \frac{e^{-d_{ij}^2 / 2\sigma_i^2}}{\sum_k e^{-d_{ik}^2 / 2\sigma_i^2}}$$

probability of picking j given that you start at i

Turning conditional probabilities into pairwise probabilities

To get a symmetric affinity between i and j we sum the two conditional probabilities and divide by the number of points (points are not allowed to choose themselves).

joint probability of picking the pair i,j

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

This ensures that all the pairwise affinities sum to 1 so they can be treated as probabilities.

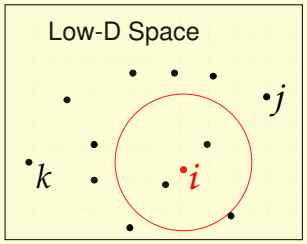
$$\sum_{i<j} p_{ij} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} p_{ij} = 1$$

Throwing away the raw data

- The pairwise affinities are all we are going to use for deciding on the low-dimensional representations of the datapoints.
- There are many possible ways to define the affinity of a pair of high-dimensional data-points.
  - Other types of affinity may work better.
  - Temporal adjacency is especially easy to use.

Evaluating an arrangement of the data in a low-dimensional space

- Give each data-point a location in the low-dimensional space.
  - Define low-dimensional probabilities symmetrically.
  - Evaluate the representation by seeing how well the low-D probabilities model the high-D affinities.



$$q_{ij} = \frac{e^{-d_{ij}^2}}{\sum_{k<l} e^{-d_{kl}^2}}$$

The cost function for a low-dimensional representation

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

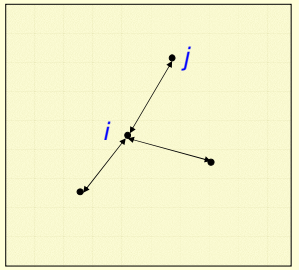
- For points where p_{ij} is large and q_{ij} is small we lose a lot.
  - Nearby points in high-D really want to be nearby in low-D
- For points where q_{ij} is large and p_{ij} is small we lose a little because we waste some of the probability mass in the Q distribution.
  - Widely separated points in high-D have a mild preference for not being too close in low-D.
  - But it doesn't cost much to make the manifold bend back on itself.

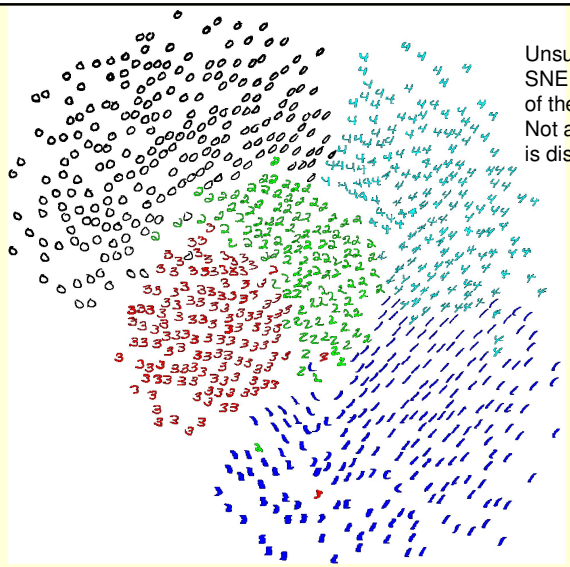
The forces acting on the low-dimensional points

$$\frac{\partial KL(P \| Q)}{\partial \mathbf{y}_i} = 2 \sum_j (\mathbf{y}_i - \mathbf{y}_j)(p_{ij} - q_{ij})$$

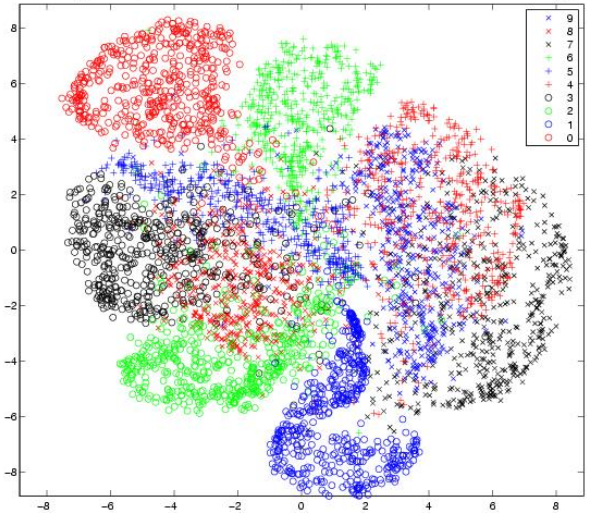
extension: $(\mathbf{y}_i - \mathbf{y}_j)$ stiffness: $(p_{ij} - q_{ij})$

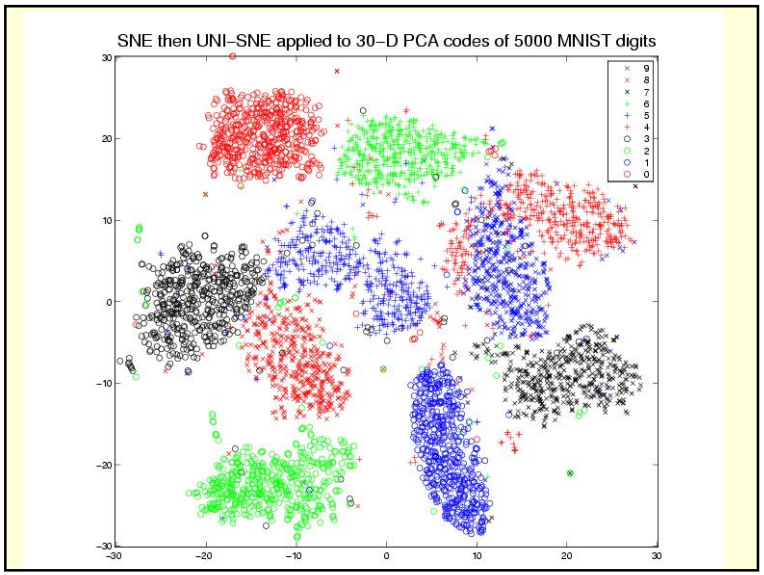
- Points are pulled towards each other if the p's are bigger than the q's and repelled if the q's are bigger than the p's
  - Its equivalent to having springs whose stiffnesses are set dynamically.

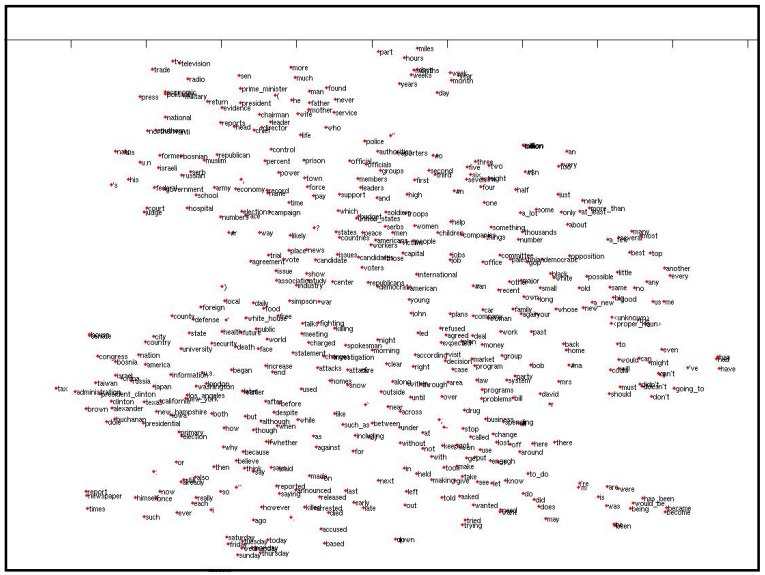


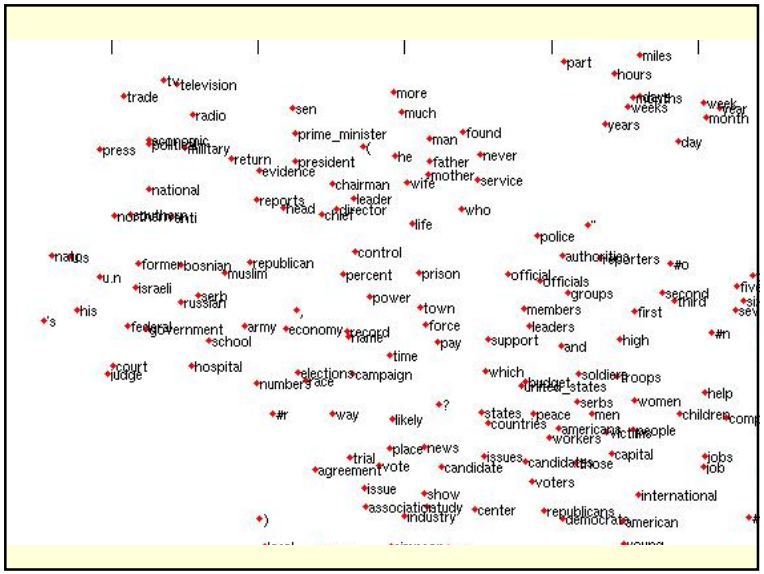


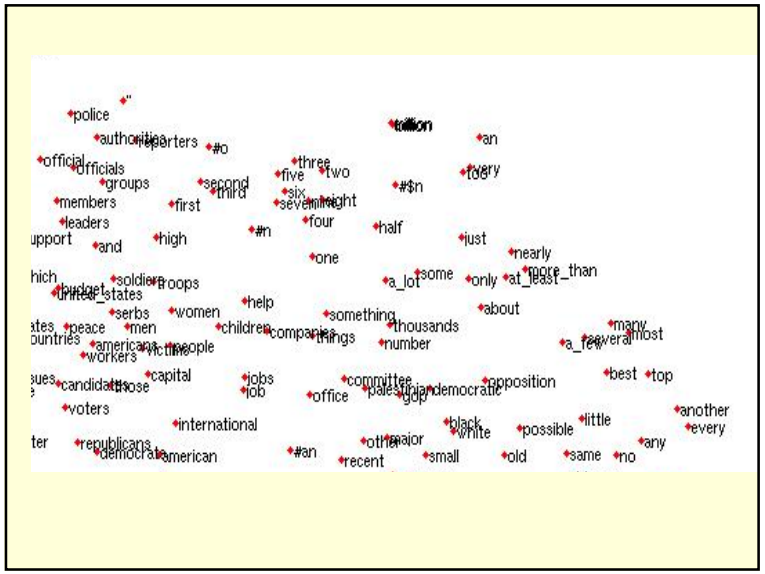
Unsupervised SNE embedding of the digits 0-4. Not all the data is displayed

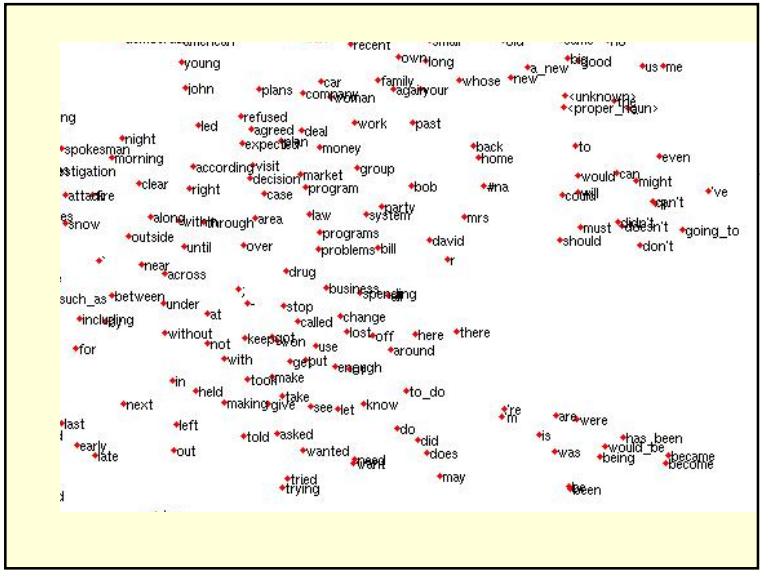


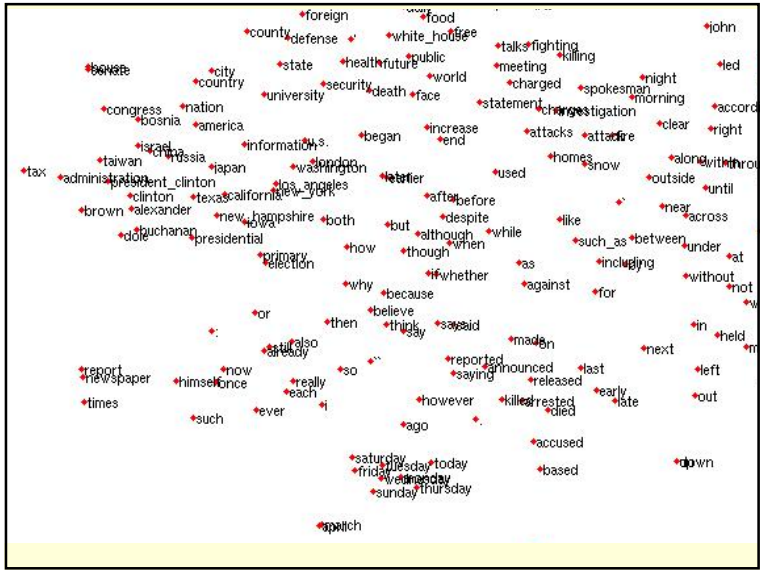

SNE then UNI–SNE applied to 30–D PCA codes of 5000 MNIST digits
9
8
7
6
5
4
3
2
1
0
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30








recent
young
own
long
a_new
big
good
us
me
john
plans
car
family
whose
new
<unknown>
<proper_noun>
refused
led
agreed
deal
work
past
night
spokesman
money
back
home
to
even
morning
according
visit
decision
market
group
would
can
might
clear
right
case
program
bob
#na
could
can't
've
snow
area
law
system
party
mrs
must
doesn't
going_to
outside
until
over
programs
problems
bill
david
should
don't
near
across
drug
business
spending
such_as
between
under
at
stop
called
change
including
without
not
keep
lost
off
here
there
for
with
use
around
get
put
enough
in
held
took
make
to_do
next
making
give
take
see
let
know
're
'm
are
were
last
left
told
asked
do
did
is
has_been
would_be
early
late
out
wanted
does
was
being
became
become
tried
trying
may
been


foreign
food
john
county
defense
white_house
free
talks
fighting
killing
state
health
future
public
meeting
led
city
country
security
world
charged
spokesman
night
university
death
face
statement
morning
congress
nation
bosnia
america
increase
attacks
clear
right
began
end
information
homes
snow
along
taiwan
russia
japan
washington
tax
administration
president_clinton
used
outside
until
clinton
texas
california
los_angeles
new_york
after
before
brown
alexander
new_hampshire
iowa
both
despite
like
near
across
buchanan
dole
presidential
but
although
while
such_as
between
under
how
though
when
primary
election
as
including
at
if
whether
without
why
because
against
for
not
or
believe
then
think
say
said
in
held
also
already
made
on
next
report
newspaper
now
so
reported
announced
last
left
himself
once
really
saying
released
each
early
times
however
killed
arrested
late
out
ever
i
died
such
ago
accused
saturday
tuesday
today
friday
based
down
sunday
thursday
speech