Statistical Machine Learning: Clustering and Dimension Reduction

Daniel J. McDonald and Darren Homrighausen

Indiana University, Bloomington
and Colorado State University

mypage.iu.edu/∼dajmcdon

stat.colostate.edu/∼darrenho

April 2-3, 2013

What is Clustering?

All the previous applications presumed that there is a response Y

However, in some cases, there is no response at all.

- A large curpus of emails sent at the Enron main office right before it collapsed
- Everyone's cell phone behavior and location in a particular city (these data sets do exist)
- The relationship between all stocks on the S&P 500
- A cancer researcher might assay gene expression levels in a group of patients with different cancers

What is Clustering?

All the previous applications presumed that there is a response Y

However, in some cases, there is no response at all.

- A large curpus of emails sent at the Enron main office right before it collapsed
- Everyone's cell phone behavior and location in a particular city (these data sets do exist)
- The relationship between all stocks on the S&P 500
- A cancer researcher might assay gene expression levels in a group of patients with different cancers

An Overview of Clustering

The idea is to find paterns in the data. However, we don't have a supervisor (Y) and hence clustering is sometimes known as unsupervised learning.

Clustering is more difficult than classification/regression because solutions are vague and often unverifiable.

THE SET-UP

Suppose we have observations

$$X_1, \ldots, X_n$$

Here, we want to find a relationship between the X's, commonly by grouping them (putting them into 'clusters').

This is fundamentally different from our previous discussions, as there is no notion of 'prediction' accuracy which we are trying to maximize.

For this talk, clustering will be informal exploratory data analysis via lower dimensional embeddings.

There are many, many other approaches, however.

LOWER DIMESIONAL (METRIC) EMBEDDINGS

Spectral connectivity analysis (SCA)

- Linear and nonlinear
- Dimension reduction or feature creation
- Examples: PCA and Fisher discriminant analysis, Locally linear embeddings, Hessian maps, Laplacian eigenmaps
- Useful tools as inputs to classification, clustering, and regression

Principal component analysis

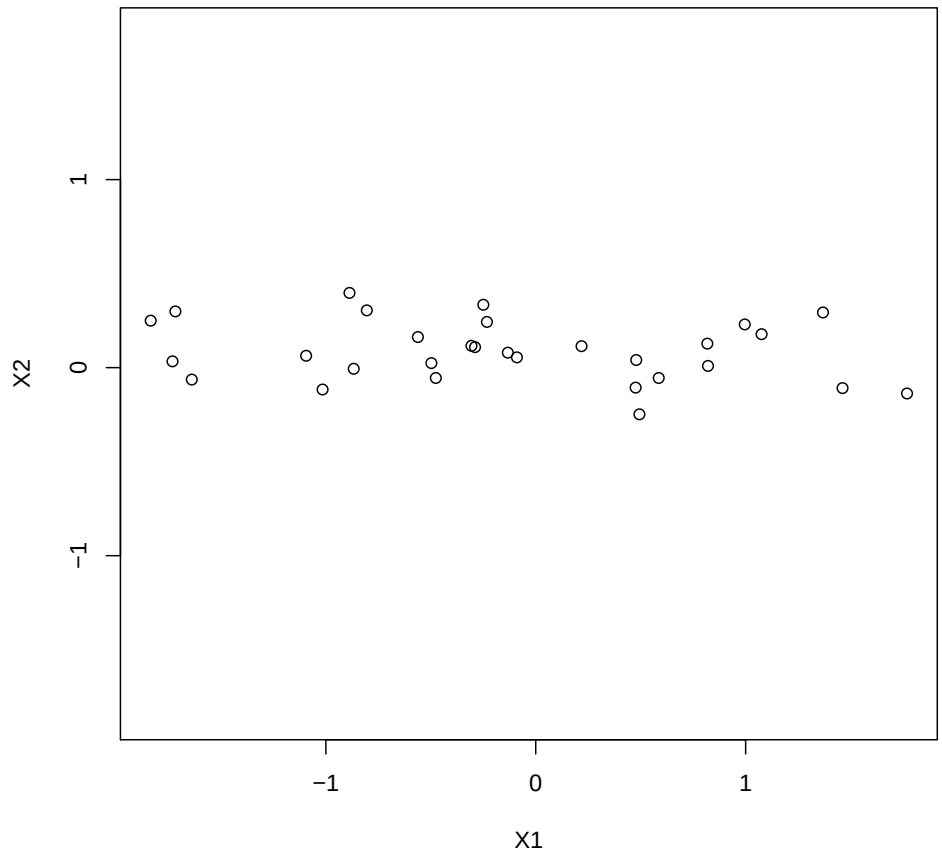
PCA

Collect data: $X_1, \ldots, X_n$ where $X_i \in \mathbb{R}^p$.

1. Center and scale the data matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$
2. Compute the spectral decomposition of $\mathbb{X}^\top \mathbb{X} = VD^2V^\top$
   [Could (and should!) use SVD of $\mathbb{X} = UDV^\top$]
3. The principal component scores are in $\mathbb{R}^n$. These are the coordinates of the observations in each PC. ($UD = \mathbb{X}V$)
4. The principal component loading vectors are in $\mathbb{R}^p$. This is the rotation needed to change the alignment of the original data to the PC axis (V).

LOWER DIMENSIONAL EMBEDDINGS: TOY EXAMPLE

Suppose we have predictors X1 and X2

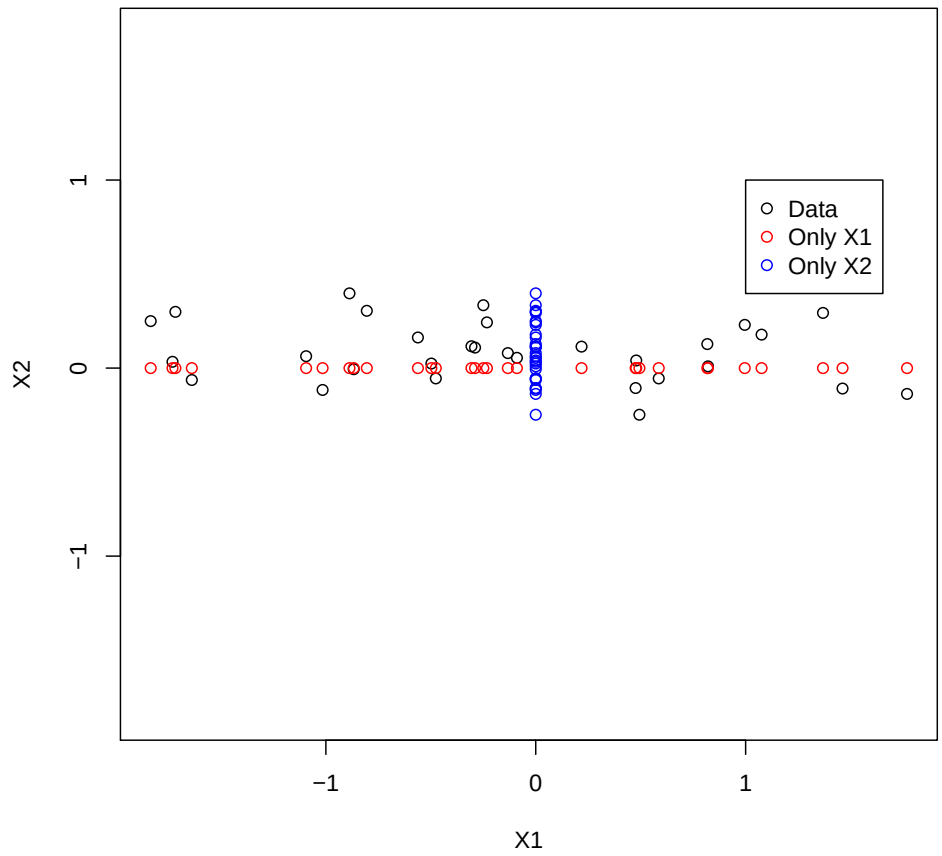


LOWER DIMENSIONAL EMBEDDINGS: TOY EXAMPLE

A lower dimensional embedding is given by

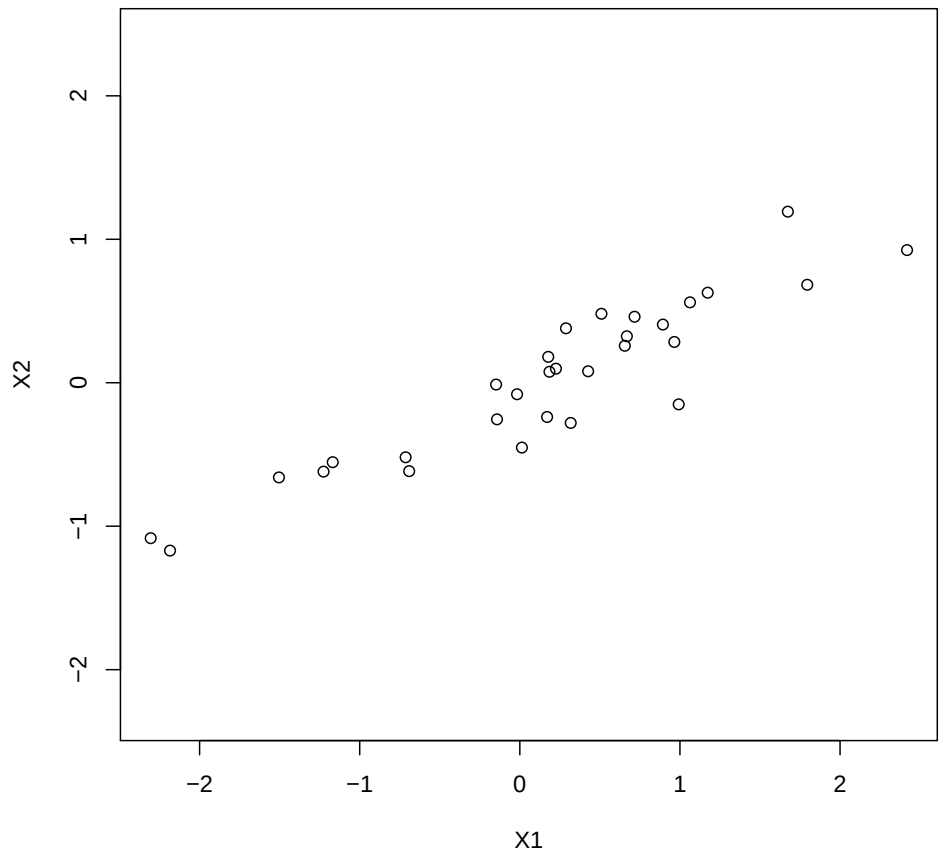
Using the red dots (that is, by setting X2 to zero):

Using the blue dots (that is, by setting X1 to zero):



LOWER DIMENSIONAL EMBEDDINGS: TOY EXAMPLE

An important feature of the First Example is that X1 and X2 aren't correlated with each other. What if they are correlated?

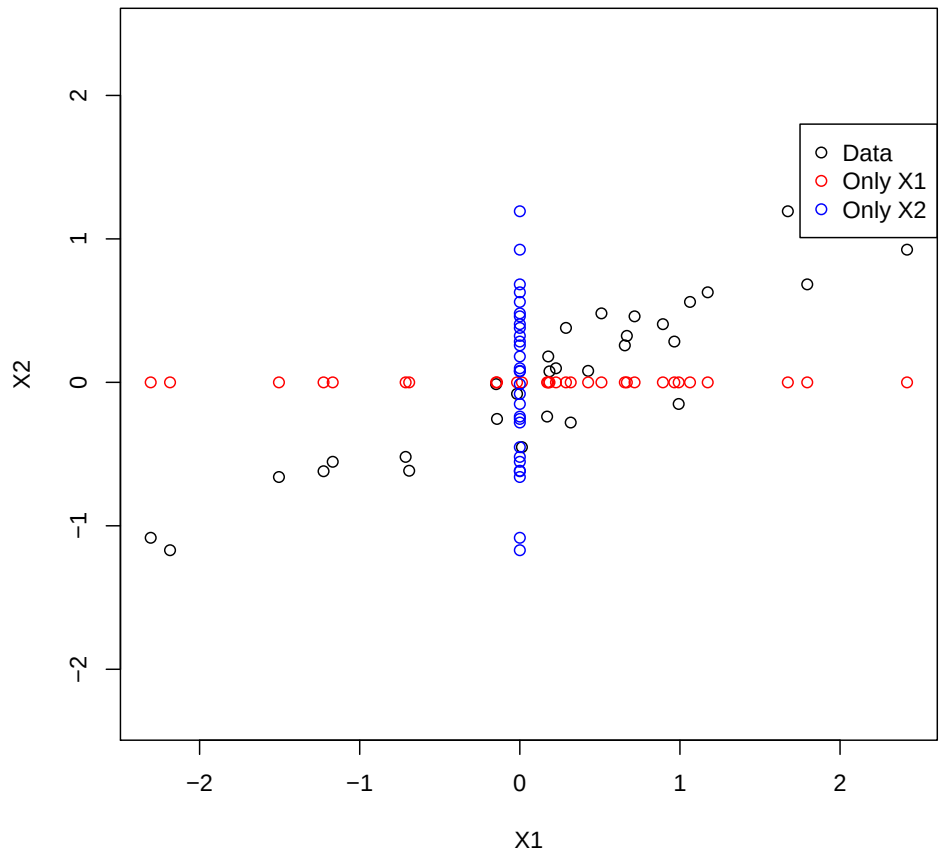


LOWER DIMENSIONAL EMBEDDINGS: TOY EXAMPLE

A lower dimensional embedding is given by

Using the red dots (that is, by setting X2 to zero):

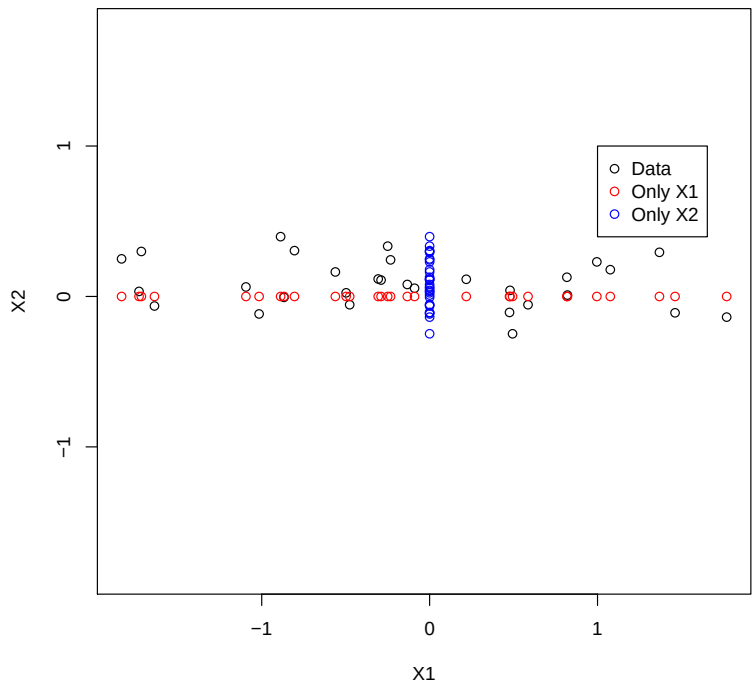
Using the blue dots (that is, by setting X1 to zero):

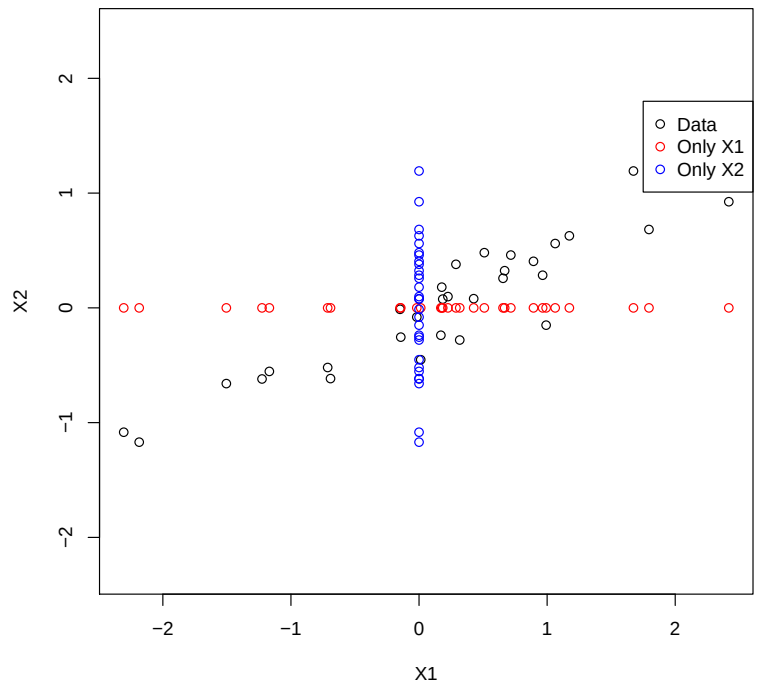


Lower dimensional embeddings: Comparison of examples

The second example loses much more information with this simplistic dimension reduction strategy.

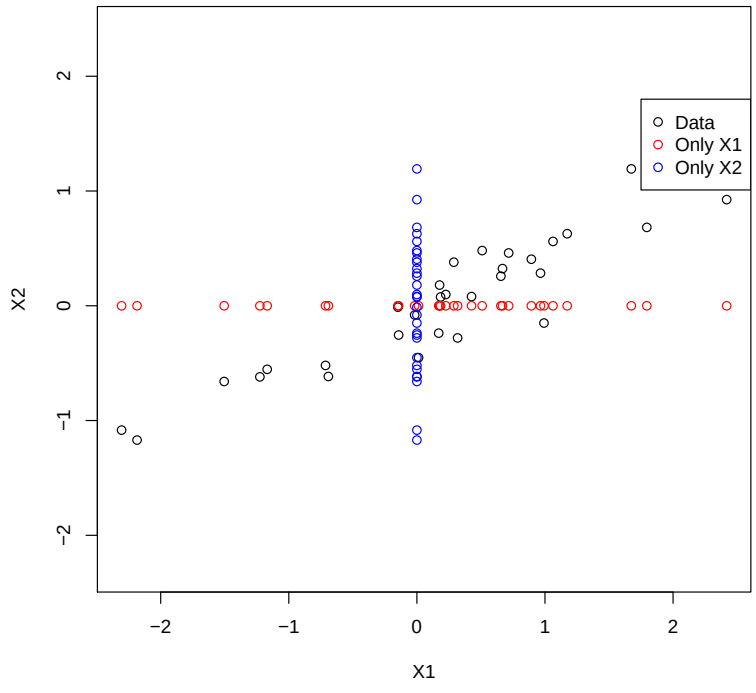
However, if we can find the rotation between the examples, then we can use this simple approach.

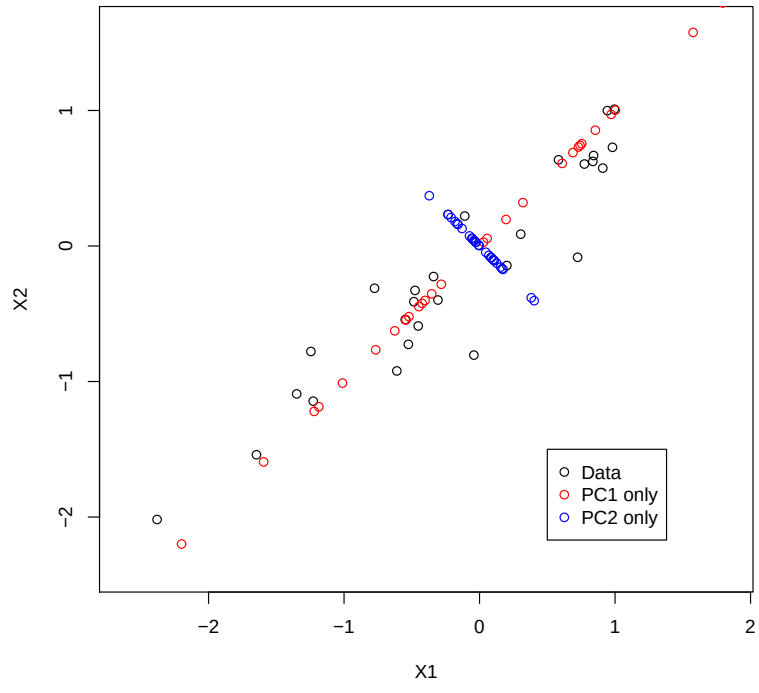




LOWER DIMENSIONAL EMBEDDINGS

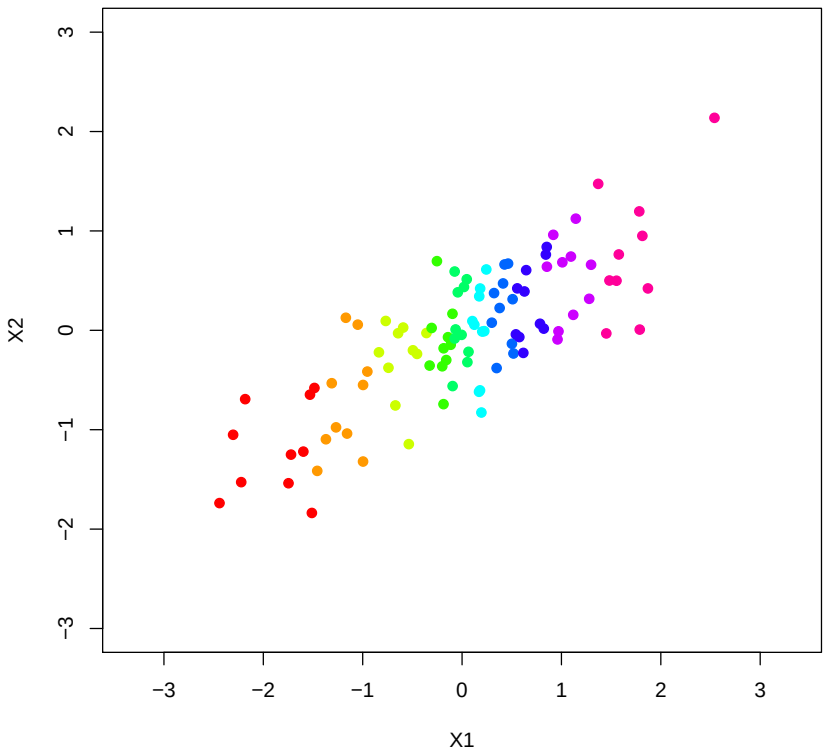
It turns out that Principal Components Analysis (PCA) gives us exactly this rotation (the matrix V).

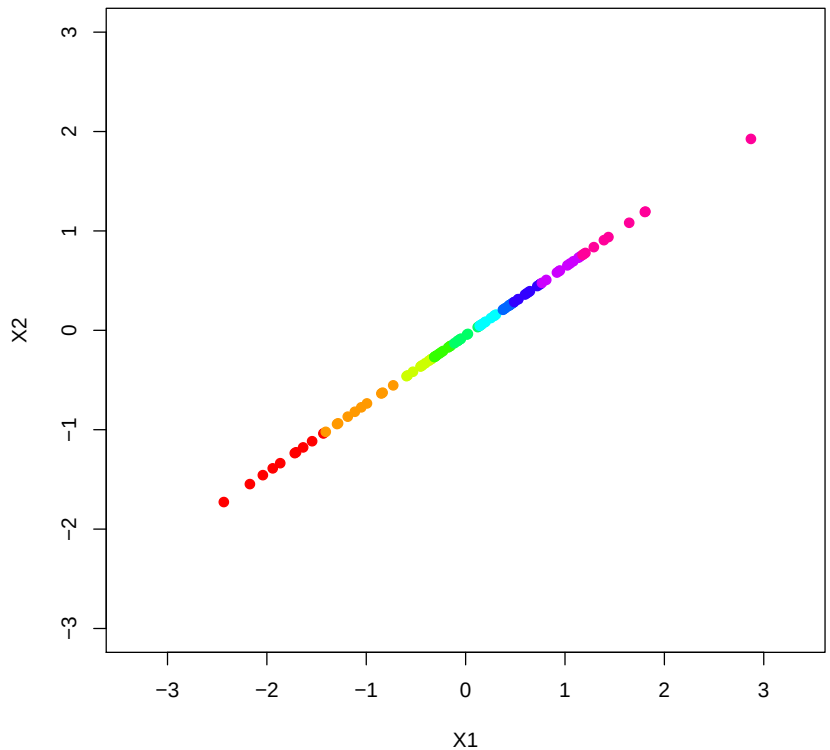




When PCA Works Well

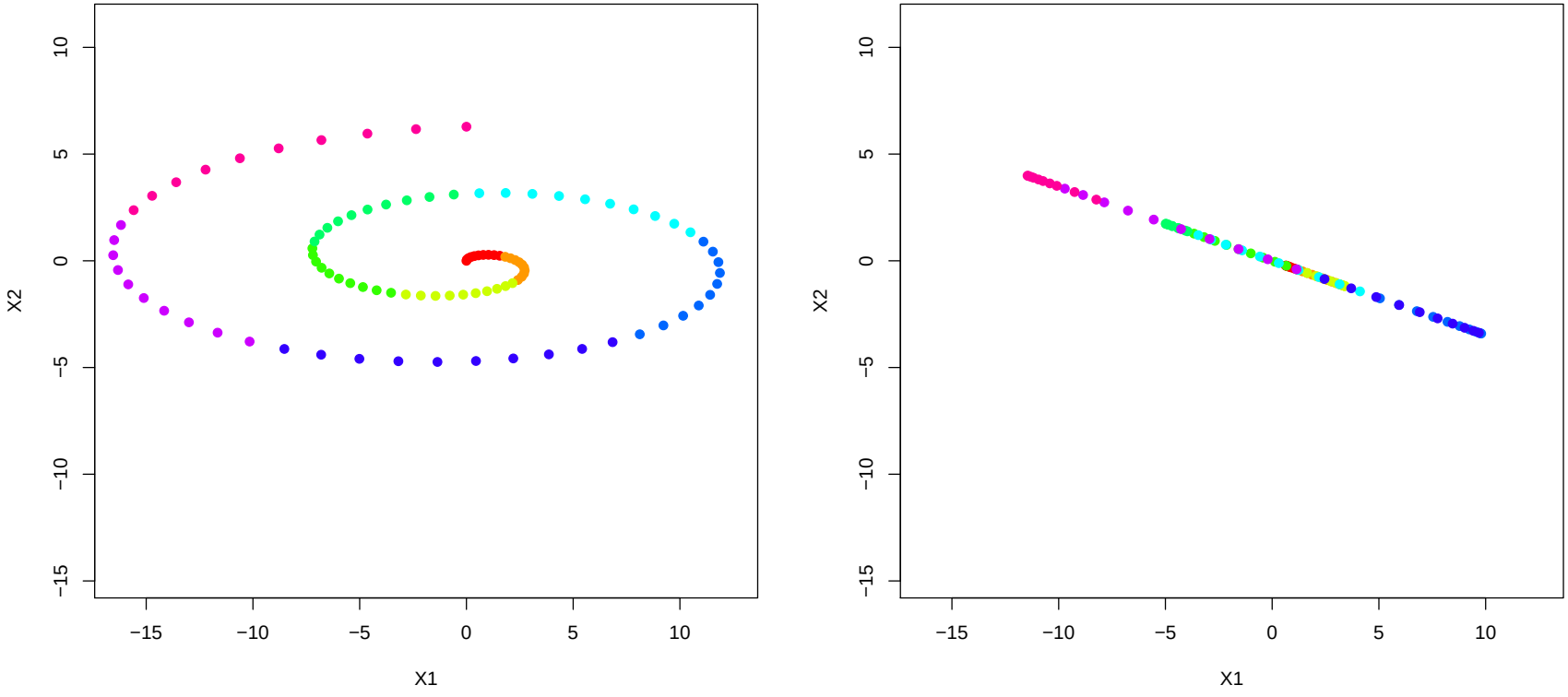
PCA can do effective dimension reduction (that is, explain most of the data with $m < p$ components) as long as the data can be efficiently represented as 'lines' (or planes, or hyperplanes). So, in two dimensions:





When PCA doesn't work well

What about other data structures? Again in two dimensions



Here, we have failed miserably.

EXPLANATION

- PCA wants to minimize distances (equivalently maximize variance). This means it ‘slices’ through the data at the ‘meatiest’ point, and then the next one, and so on. If the data are ‘curved’ this is going to induce artifacts.
- PCA also looks at things as being ‘close’ if they are near each other in a Euclidean sense

  [this is essentially all correlation is].
- On the spiral, our intuition says that things are ‘close’ only if the distance is constrained to go along the curve. In other words, purple and blue are close, blue and red are not.

Nonlinear embeddings

Laplacian Eigenmaps

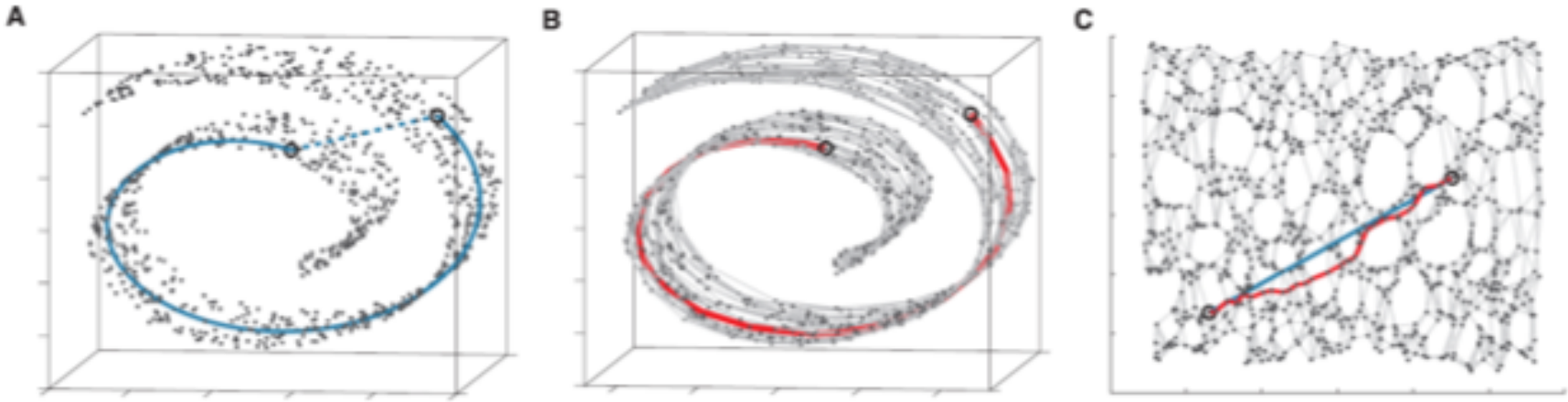
In order to use the intuitive distance, we need to know the geometry of the data. This needs to be estimated.

We can get an estimate of the distance in the unknown geometry that the data come from (known as a manifold) by altering the usual Euclidean distance.

Some notes:

- The name 'Laplacian Eigenmaps' comes from getting the eigenvector decomposition of the Laplacian restricted to the manifold (which is the second derivative version of the gradient).
- If the manifold is smooth, then local Euclidean distance is an approximation to the distance on the manifold.

Local Euclidean distance approximates the geodesic



The red line is the local Euclidean path between the two points, while the blue line is the path along the manifold.

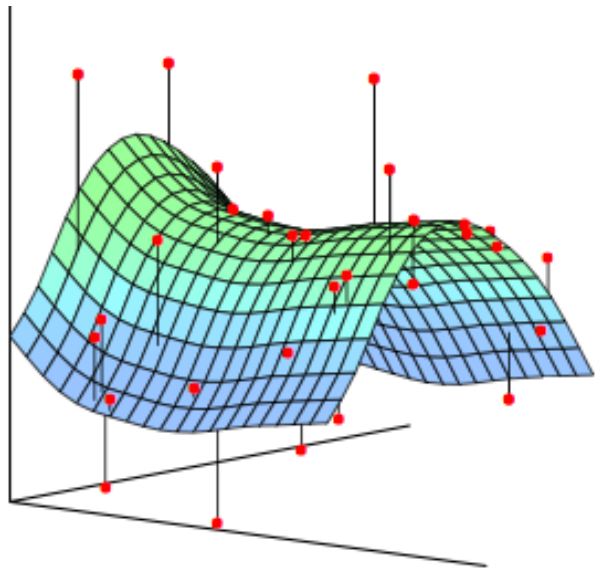
Source: James, Witten, Hastie, Tibshirani (2013)

What is a manifold?

Let's think of a manifold as a lower dimensional structure in our data (that is, $\mathbb{R}^p$).

If that structure is linear, then Euclidean distance is still a fine choice

If that structure is nonlinear, then Euclidean distance isn't applicable:

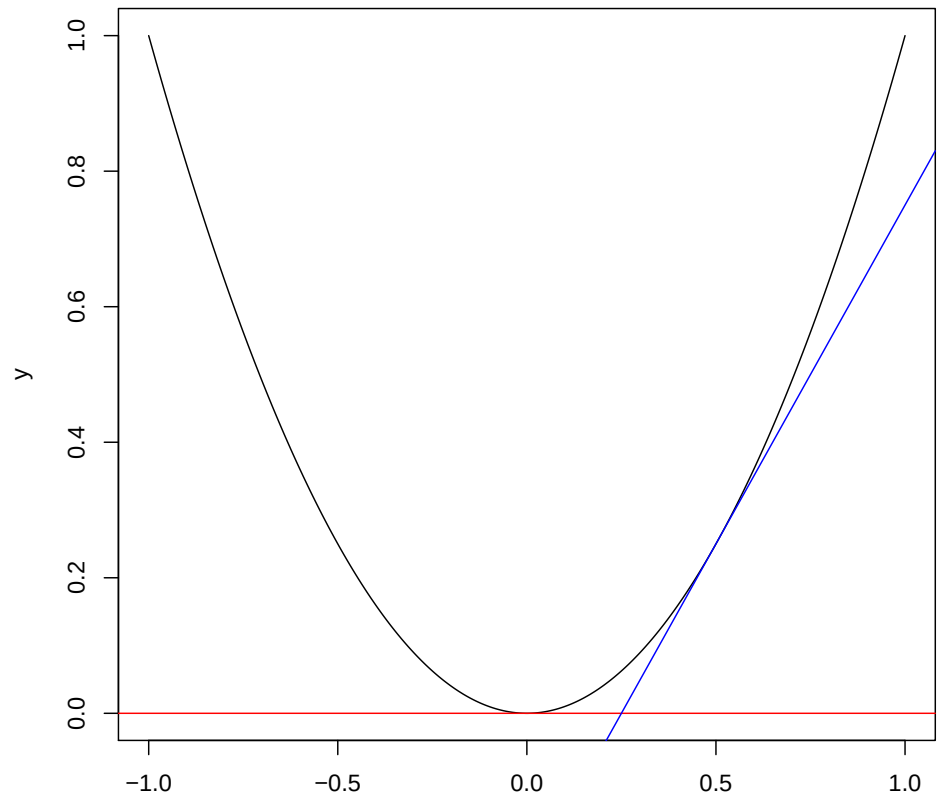


Source: James, Witten, Hastie, Tibshirani (2013)

WHAT IS A MANIFOLD?

How good of an approximation is Euclidean distance? This question is equivalent to how asking: how quickly does the tangent space change? In 1-D, the tangent space is just the first derivative at that point:

$$f(x) = x^2 \Rightarrow f'(x) = 2x.$$



What is a Manifold?

How quickly does the tangent space change? Well, this is the second derivative:

$$f(x) = x^2 \Rightarrow f''(x) = 2$$

Therefore, the quality of the (local) Euclidean distance, depends on the second derivative.

In higher dimensions, the second derivative is known as the Laplacian:

$$\sum_j \frac{\partial^2 f}{\partial x_j^2}$$

Note: This is also known as the divergence of the gradient.

What are Laplacian Eigenmaps, then?

If we think of the Laplacian as an operator mapping a function to the divergence of its gradient, it turn out it looks almost like a matrix operation.

Key Idea: **We can get the Eigenvectors/Eigenvalues of this operator. Analogously to PCA, we can now do inference with these Eigenvectors.**

Laplacian Eigenmaps

Collect data: $X_1, \ldots, X_n$ where $X_i \in \mathbb{R}^p$.

1. Form the distance matrix $\Delta_{ij} = ||X_i - X_j||_2^2$.
2. Compute
$$\mathbb{K} = \exp\left(-\frac{\Delta}{\gamma}\right)$$
3. Form the Laplacian $\mathbb{L} = \mathbb{I} - \mathbb{M}^{-1}\mathbb{K}$,
$$\mathbb{M} = \texttt{diag}(\texttt{rowSums}(\mathbb{K}))$$
4. Compute the spectrum: $\mathbb{L} = U\Sigma U^\top$.
5. Return U_d, where U_d corresponds to the smallest d (nontrivial) eigenvalues of $\mathbb{L}$

   (Note that the eigenvectors of $\mathbb{L}$ and $\mathbb{M}^{-1}\mathbb{K}$ are the same, but $\Sigma(\mathbb{L}) = \mathbb{I} - \Sigma(\mathbb{M}^{-1}\mathbb{K})$)

DEEPER INVESTIGATION

1. Form the distance matrix Δ.

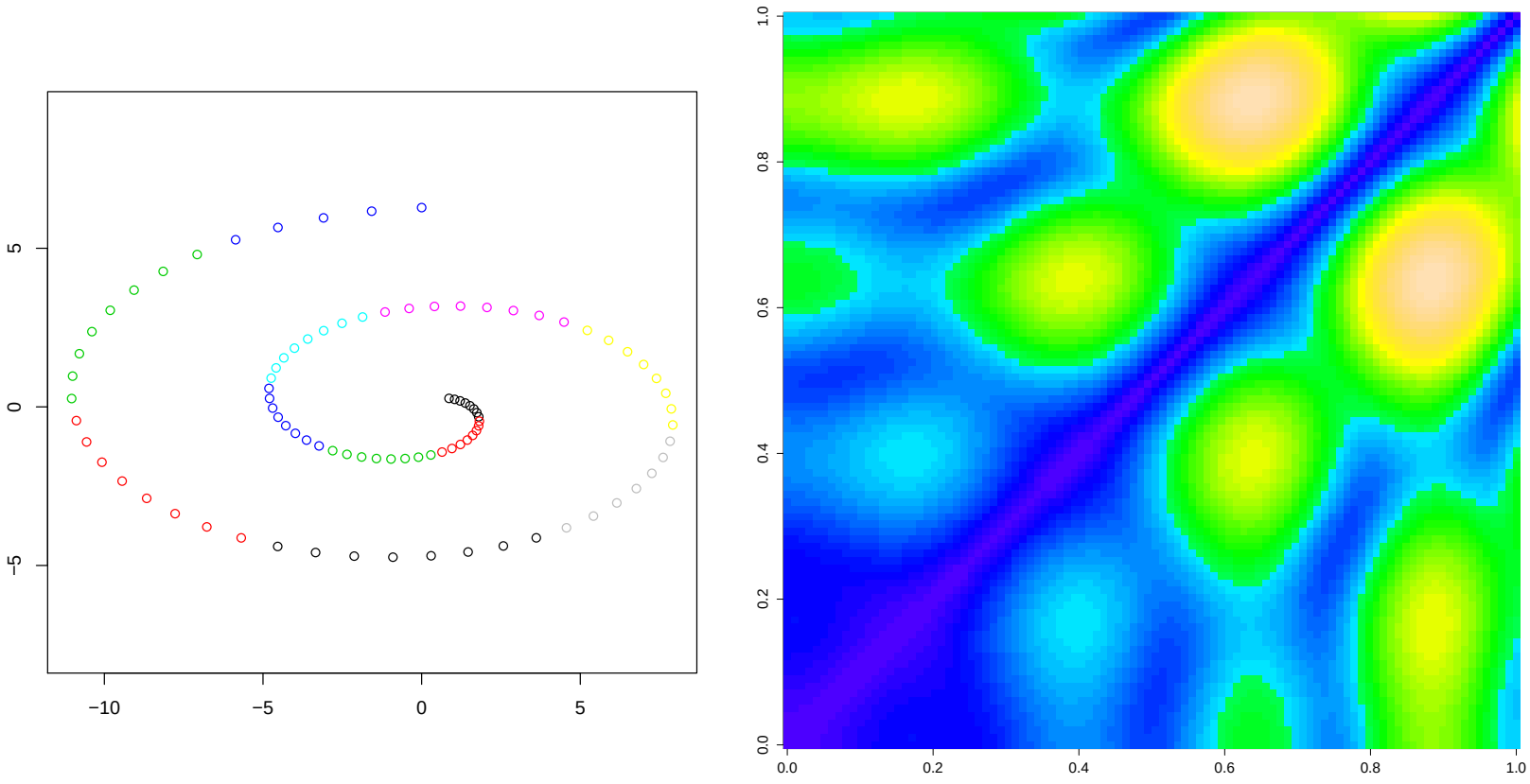


FIGURE : If we think about the center as 0 and the last blue circle as 1, then each entry the plot on the Right is the Euclidean distance between each data point on the plot on the Left (that is, Δ). The color on the Right plot goes from purple (small distance) to beige/pink (large distance).

DEEPER INVESTIGATION

2. Exponentiate Δ to form $\mathbb{K}$ for some fixed γ.

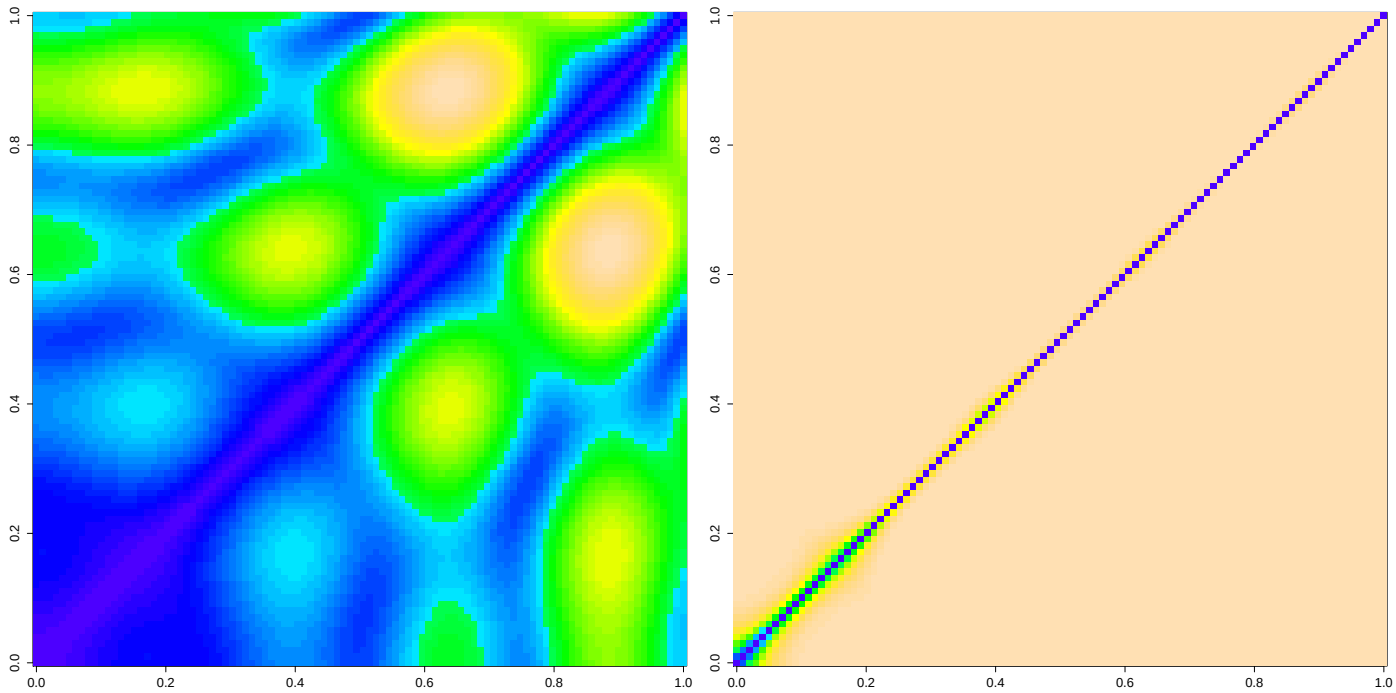
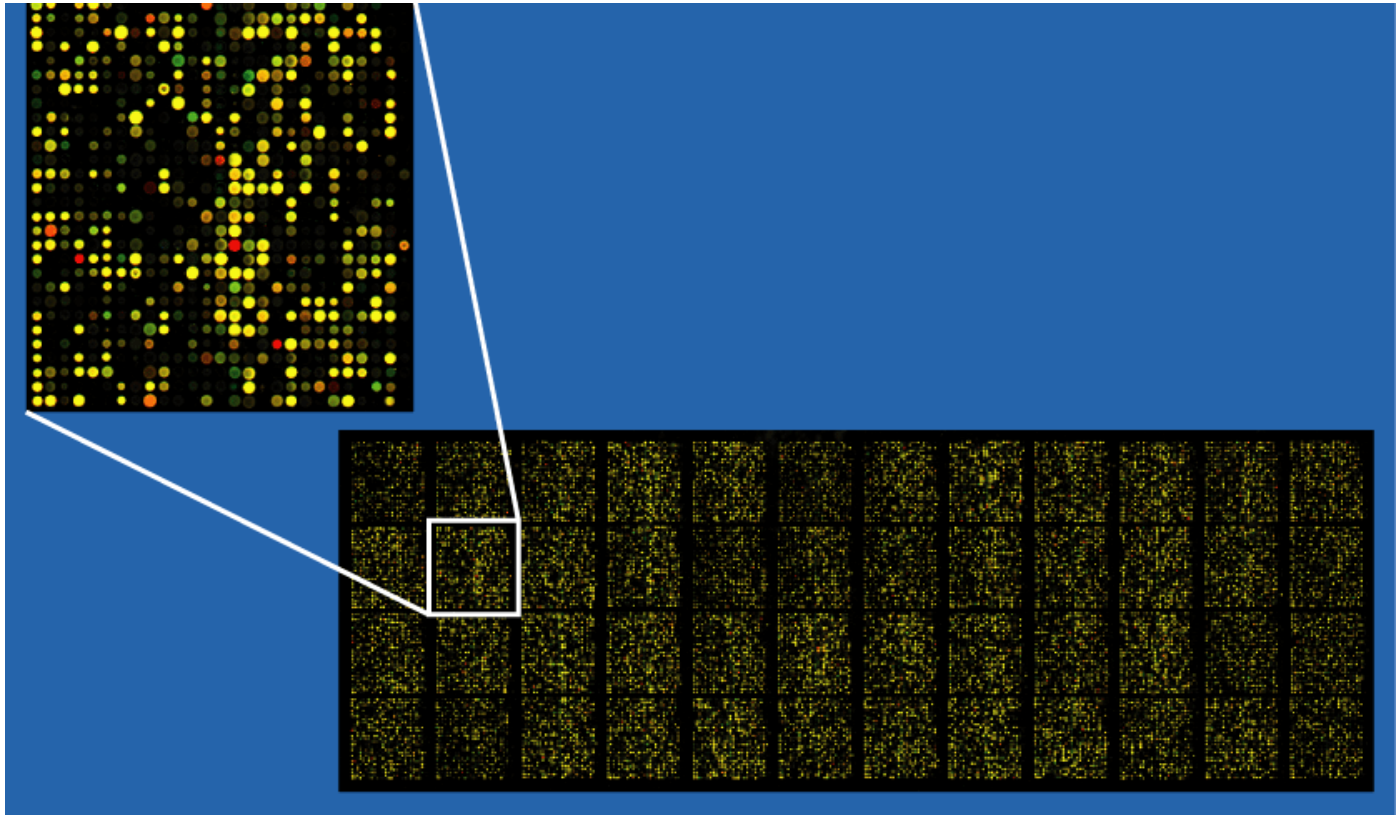


FIGURE : The Left plot is Δ and the Right plot is $\mathbb{K}$ for $\gamma = 0.95$.

AN EXAMPLE: MICROARRAY



Source: Yoon (2006)

AN EXAMPLE

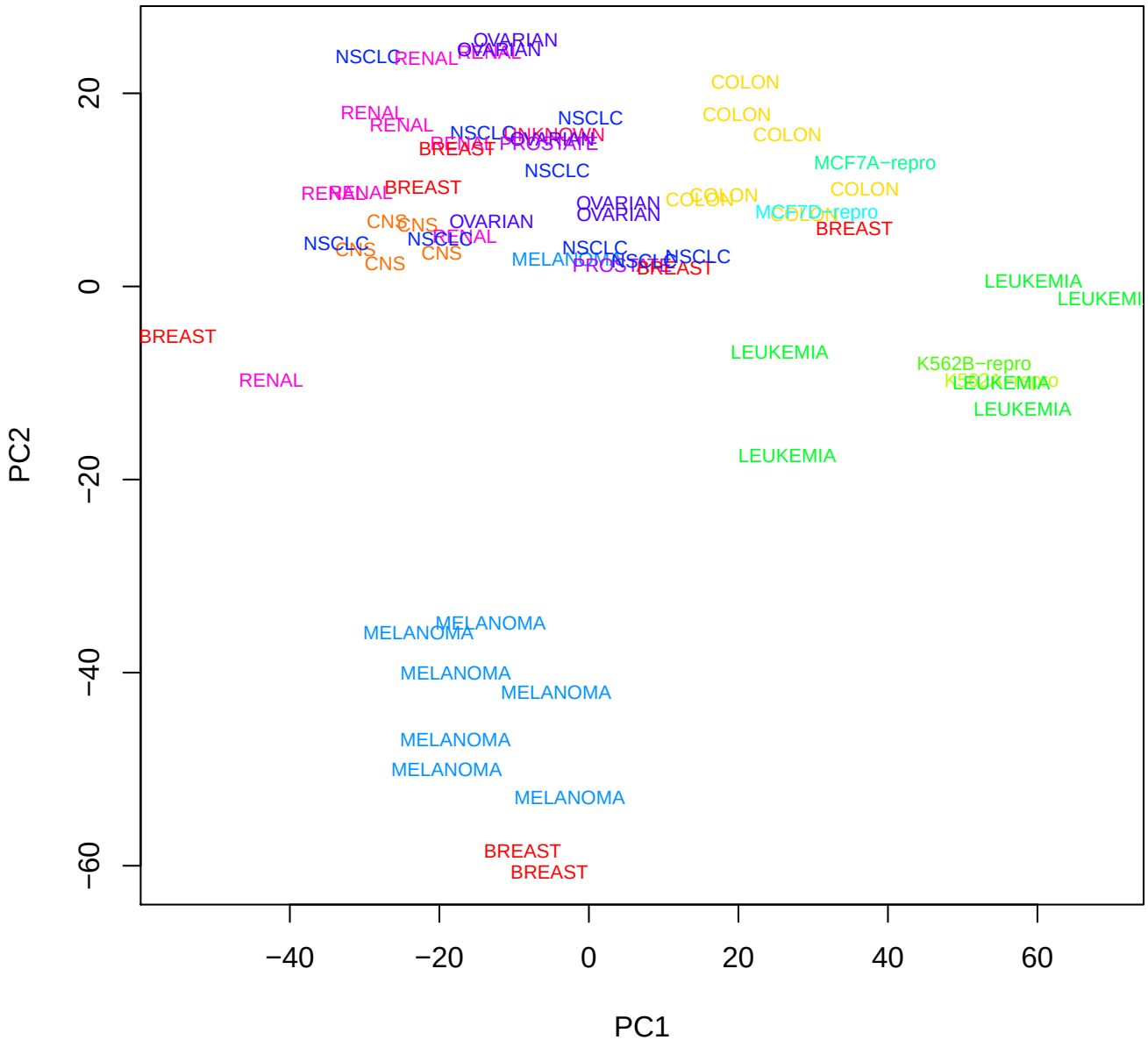
The data are gene expression measurements from cells drawn from 64 different tumors (from 64 different patients) using a device called a microarray.

$\rightarrow$ $n = 64$.

There are 6830 distinct genes in this specific section of the genome, which has been identified as a region of interest.

$\rightarrow$ $p = 6830$.

PCA AS AN EXPLORATORY TECHNIQUE



Laplacian eigenmaps as an exploratory technique

To R $\rightarrow$ for a demonstration!

OTHER METHODS

We have only scratched the surface of dimension reduction, and barely talked about clustering at all.

DIMENSION REDUCTION:
- principal components regression (PCR),
- partial least squares (PLS),
- multi-dimensional scaling (MDS),
- factor analysis, ...

CLUSTERING:
- K-nearest neighbors (KNN),
- K-means,
- hierarchical with various linkages
(ie: single, complete, average, centroid, ...)
- non-negative matrix factorization
- naive Bayes ...

Up next:
Conclusion