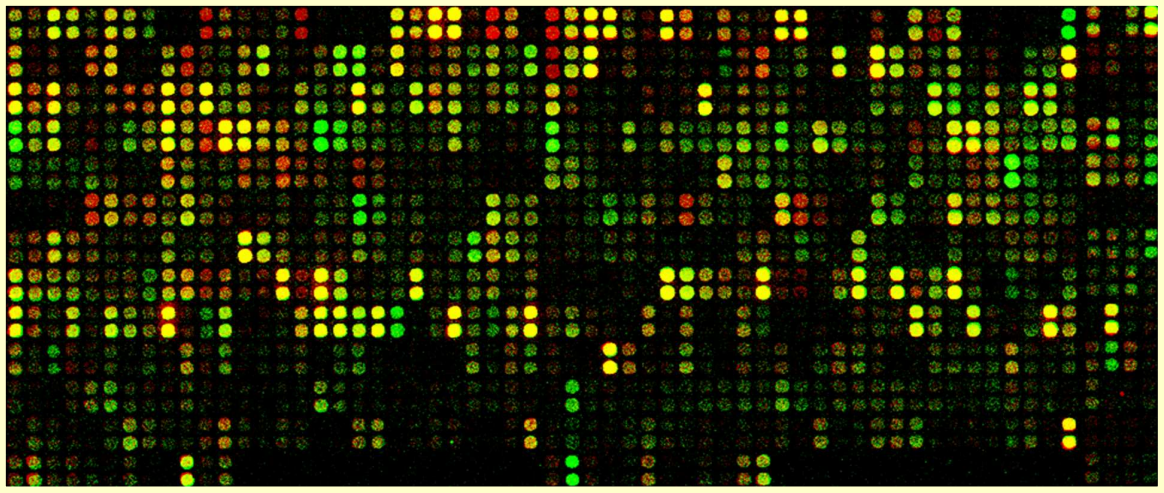
Dimension reduction of microarray data

Christopher Bowman, Richard Baumgartner

August 14, 2006



Outline

- Motivation
- Introduction to microarray data
- The Problem
- Overview of Existing Methods
  - Linear projections
    - PCA, rotated PCA
    - Correlated Feature Extraction
  - Nonlinear embeddings
    - Relative Distance Plane (RDP)
    - Local Linear Embedding (LLE)
    - Isomap
    - Whitney Reduction Network
- Conclusion, and open problems

Inspiration

- We are drowning in information and starved for knowledge. -- Anonymous
- The ability to simplify means to eliminate the unnecessary so that the necessary may speak.--Hans Hofmann
- Everything should be made as simple as possible, but not simpler.-- Albert Einstein

Motivation – basic problem

- Microarrays are a relatively new high throughput data gathering technique.
- Generate thousands of pieces of information about a single biological sample. These data are the expression levels of individual genes in the sample.
- Data can get in the way of “information”.
- Number of obvious features per sample typically much larger than number of intrinsic degrees of freedom representing significant variation in data.
- Gene expression levels are not independent.
- Usually number of samples << than number of features, meaning only a low dimensional subset of feature space is being covered – the rest is redundant.

Molecular Biology 101

- Organisms are made of cells – human's have trillions, yeast has 1.
- Cells are chemical factories that make proteins, which are the building blocks of … well almost everything.
- Each cell contains a complete copy of the genome, encoded in DNA.
- Millions of types of proteins, each has a job (or several jobs).
- Proteins are big, complex molecules.

Molecular Biology 201 (genes)

- Genes are pieces of the genome that are responsible for making a certain protein (or sometimes family of proteins).
- Gene's aren't always working, when they are making protein they're said to be *expressed*.
- Human DNA codes about 30-35 000 genes.
- Humbling thought: rice genome codes about 60 000 genes (but they're smaller).

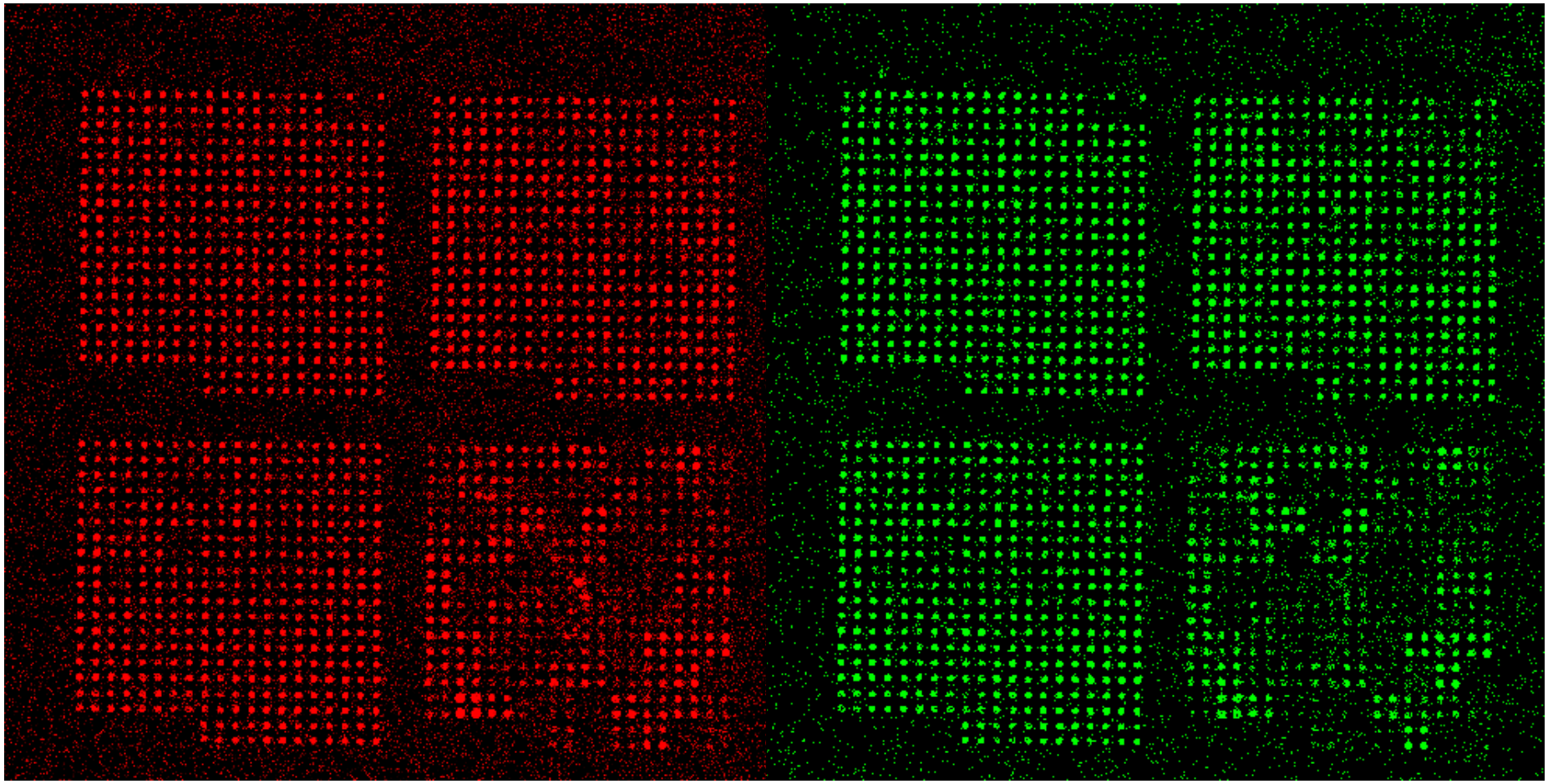
Gene Expression

- Cells are different due to differential gene expression.
- Genes are expressed due to environmental factors.
- When a gene is expressed the DNA of the gene is *transcribed* into messenger RNA (mRNA).
- The mRNA is a copy of the gene that then gets carried to a ribosome where it's *translated* into protein.
- Levels of mRNA give a "snapshot" of what genes are active (expressed) at a given time.

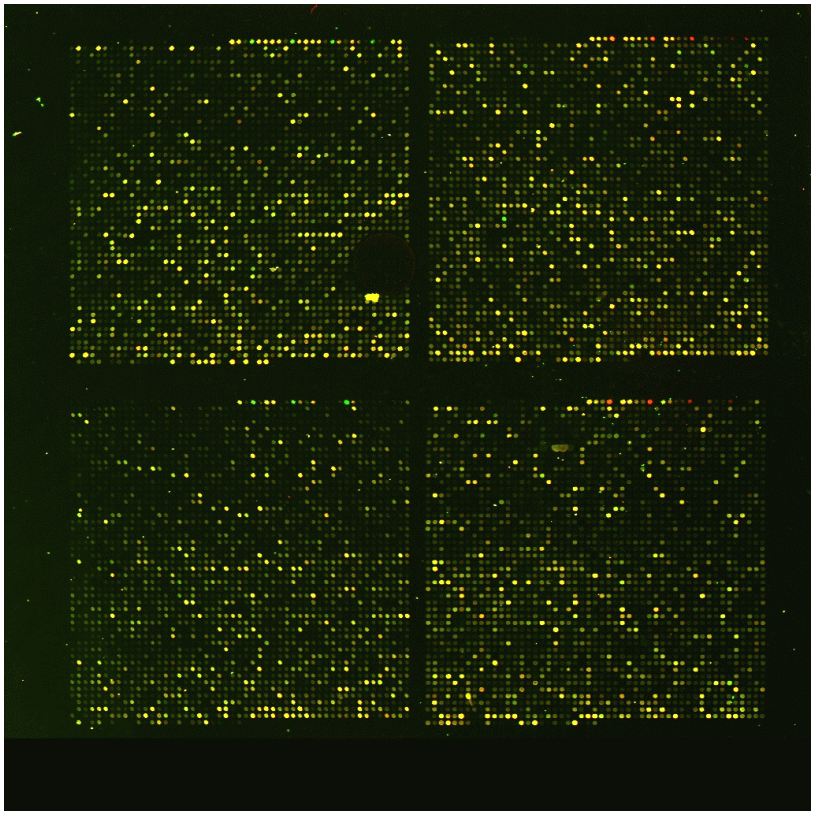
Microarrays

- Allow observation of relative gene expression of all (or most) genes in 2 samples by measuring relative concentrations of mRNA.
- Consist of many (~10 000) drops of genetic material on glass slide.
- Output of experiment is set of 2 digital images, often superimposed.

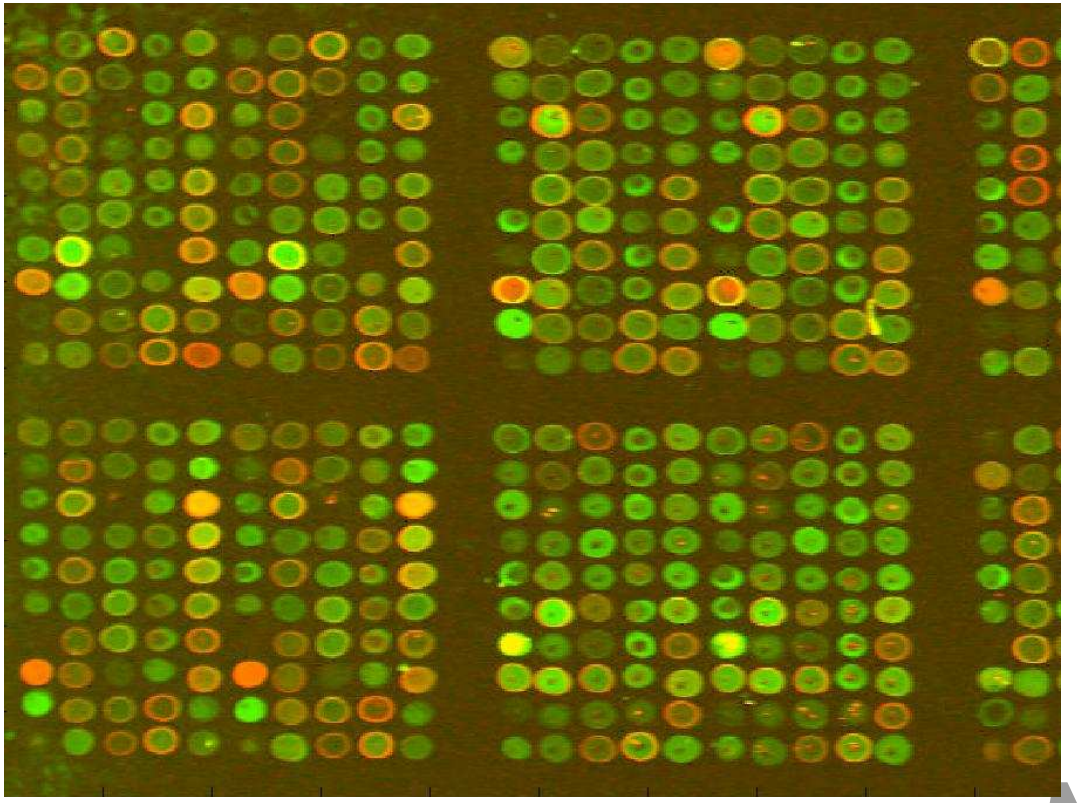
Gene Expression Data: Mouse mRNA From N2a Line, Test of Spotting Buffers.



Gene Expression Data: Public Data, From Yeast



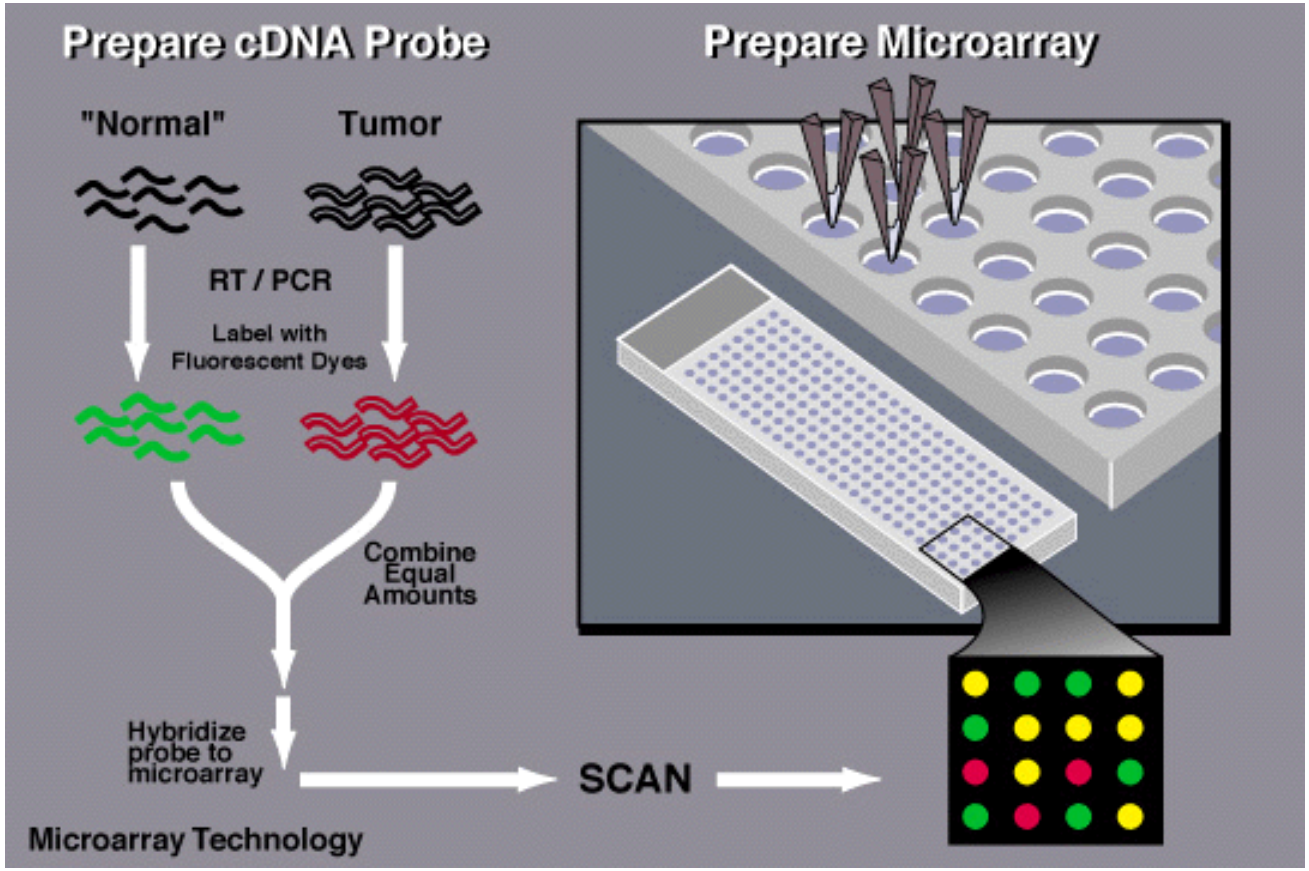
Gene Expression Data: Publicly Available Data



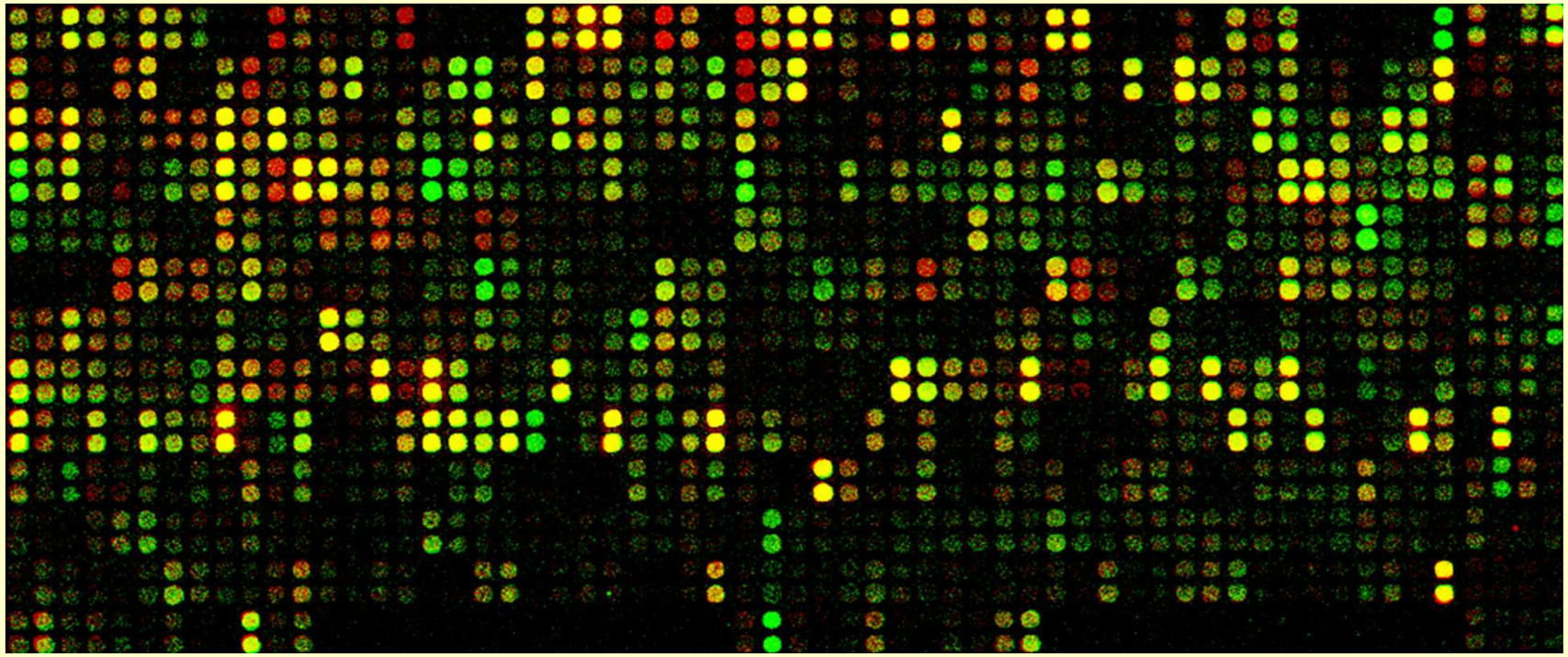
Microarray Experiments

- I'm not a molecular biologist
- Steps to a micro-array experiment:
  1. Choose cell populations
  2. mRNA extraction and reverse transcription to cDNA
  3. Fluorescent labeling of cDNA
  4. Hybridization onto a slide covered in spots which contain DNA from various genes of interest.

Microarray Technology



Microarray Example



Portion of 11K BMAP array hybridized with mouse reference RNA v. mouse brain RNA

NRC · CNRC

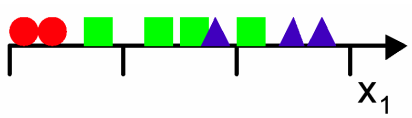
Microarray Data Analysis: Computational Tasks

- Clustering genes: which genes seem to be regulated together?
- Classifying genes: which functional class does a given gene fall into?
- Classifying cell samples*: does this patient have ALL or AML?
- Identifying biomarkers*: does gene X, alone or in combination, predict a disease state?
- Inferring regulatory networks: what is the “circuitry” of the cell?
- Unifying task: Simplify (I.e. reduce the dimensionality of) the data to aid in understanding.

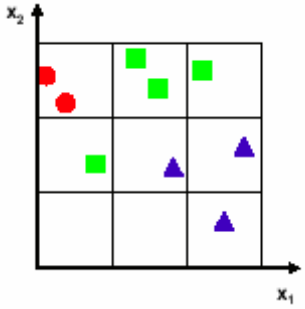
Curse of Dimensionality

- As dimension increases past the familiar 3, strange things happen.
- Volume of sphere in d dimensions is proportional to its radius raised to the d^{th} power.
- Number of points needed for density estimation in d dimensions grows exponentially in d.
- A fixed number of points in a certain volume become more and more sparse.

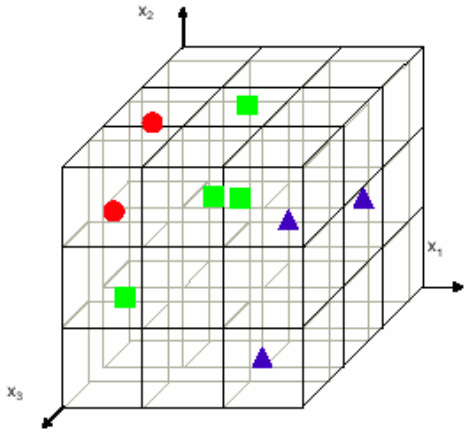
Curse of Dimensionality (example)



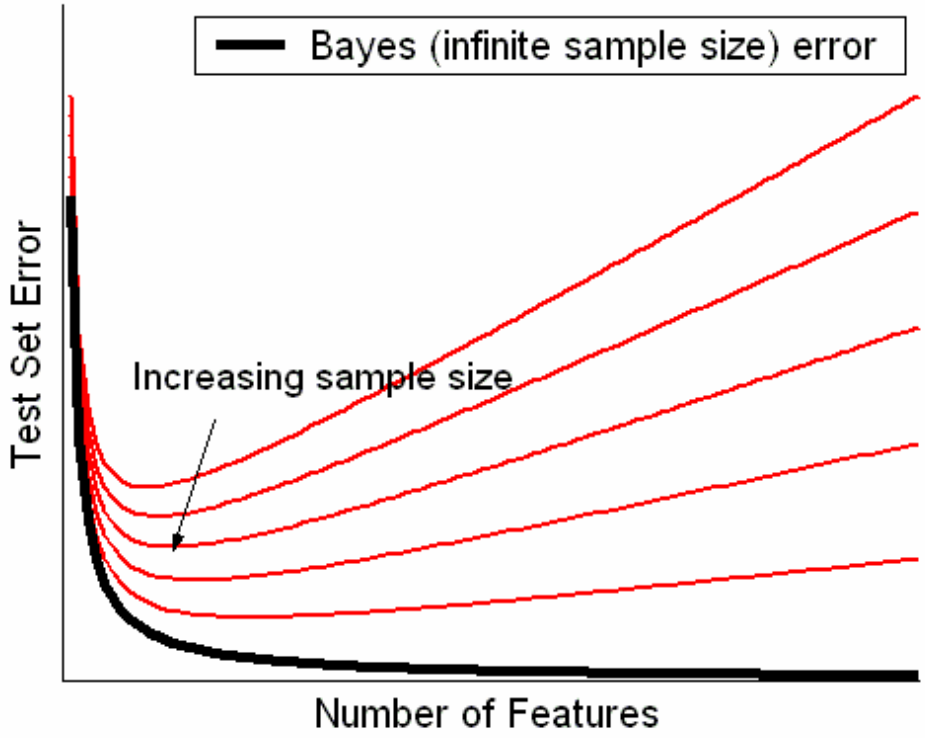
8 points in 3 classes in 1 dimension. Classes overlap, hard to split.



Up to 2 dimensions, easier to separate, points become sparser.



8 points in 3 dimensions, very sparse indeed. To achieve the same density would require 9 times as many points as in 1d.
And it just gets worse…



The curse of dimensionality: The plot shows expected test set error vs. number of features for finite sample sizes. The fewer samples, the fewer features give optimal test set error.

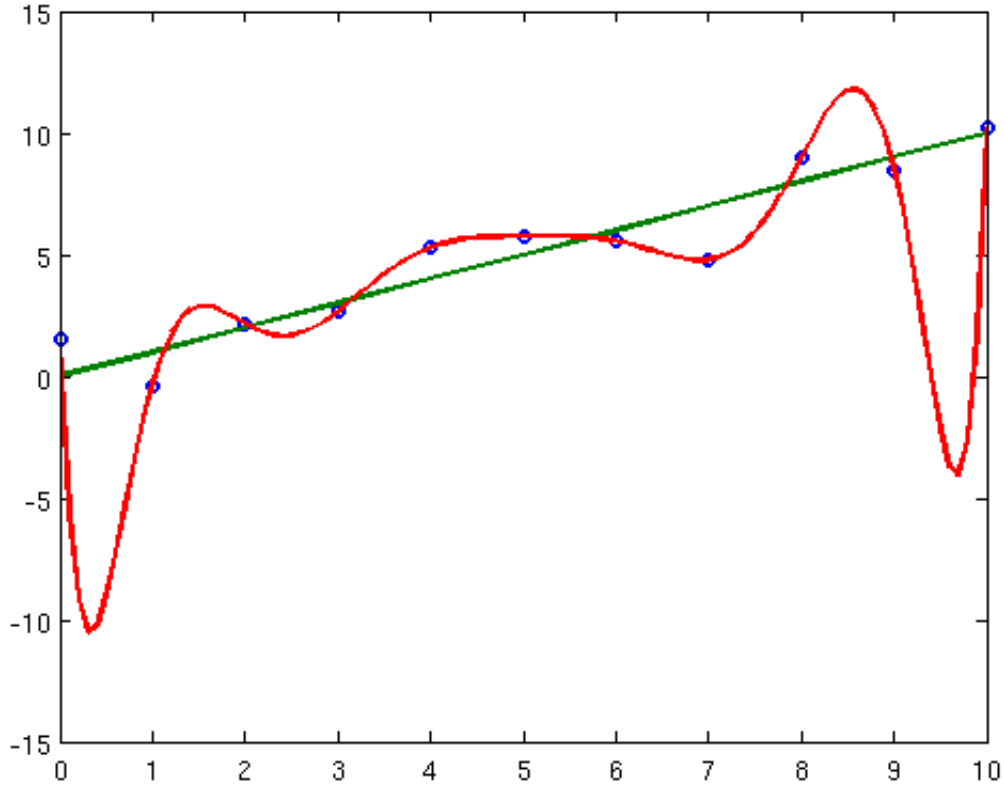
Benefits of Dimension Reduction

- Statistics:
  - Fewer degrees of freedom, harder to over-train.
  - De-correlated (or nearly) features.
  - More data (relative to dimension).
- Visualization:
  - Easier to see what's going on, especially when reduced to 2 or 3 dimensions
- Computation:
  - Smaller problem = faster.

The Problem

- Given a set of m points in an n dimensional linear space, drawn from some unknown distribution V, find a k dimensional (possibly nonlinear, k<<n) manifold that approximates the points “well”.
- “approximates well” is left vague for now,
  - At a minimum it requires generalization – that is, new unseen data points drawn from the distribution V should be well approximated by this manifold.
  - Mean square error, while not universal by any means, is probably a good thing to work with.
- Lets assume that k is prescribed beforehand (although identifying a suitable k is an interesting problem in itself).

Generalization Error

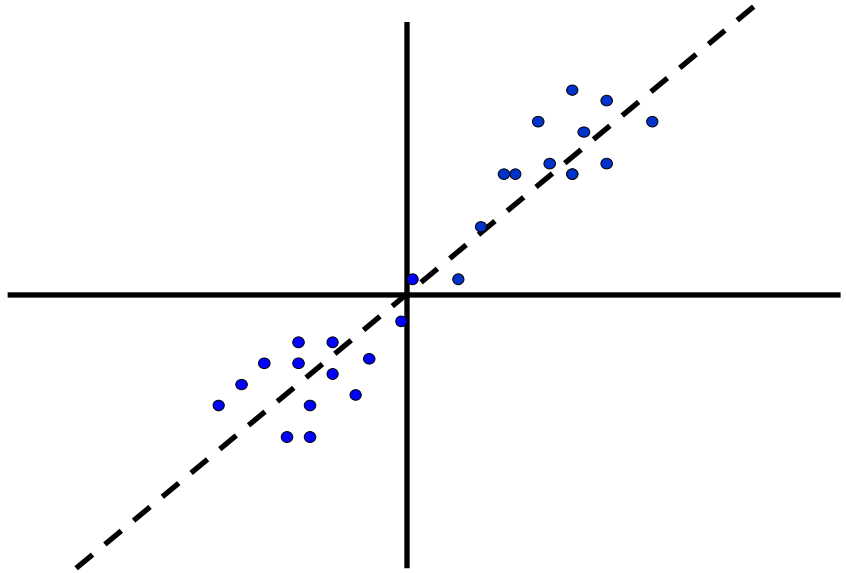


Generalization Error

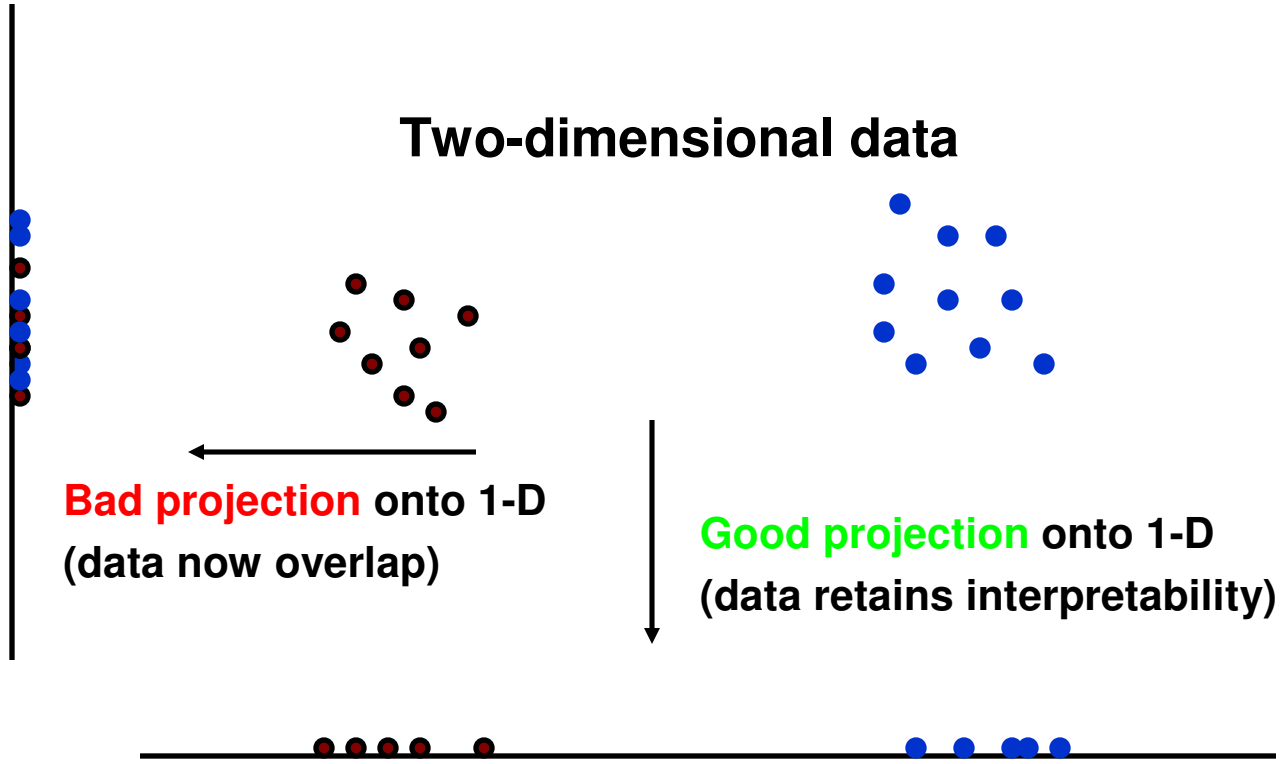
- Generalization error can be accounted for by either using new data to test the projection, or by using cross-validation.
- Lots of ways of doing cross-validation, for example, leave one out, k fold, etc.
- In general though, there needs to be a way to identify where a new, unseen datapoint will project to on the computed low dimensional manifold.

Projection

- General form of dimension reduction.
- Find a subspace (or sub-manifold) in data space that the data "lives" on and only deal with data restricted to that set.

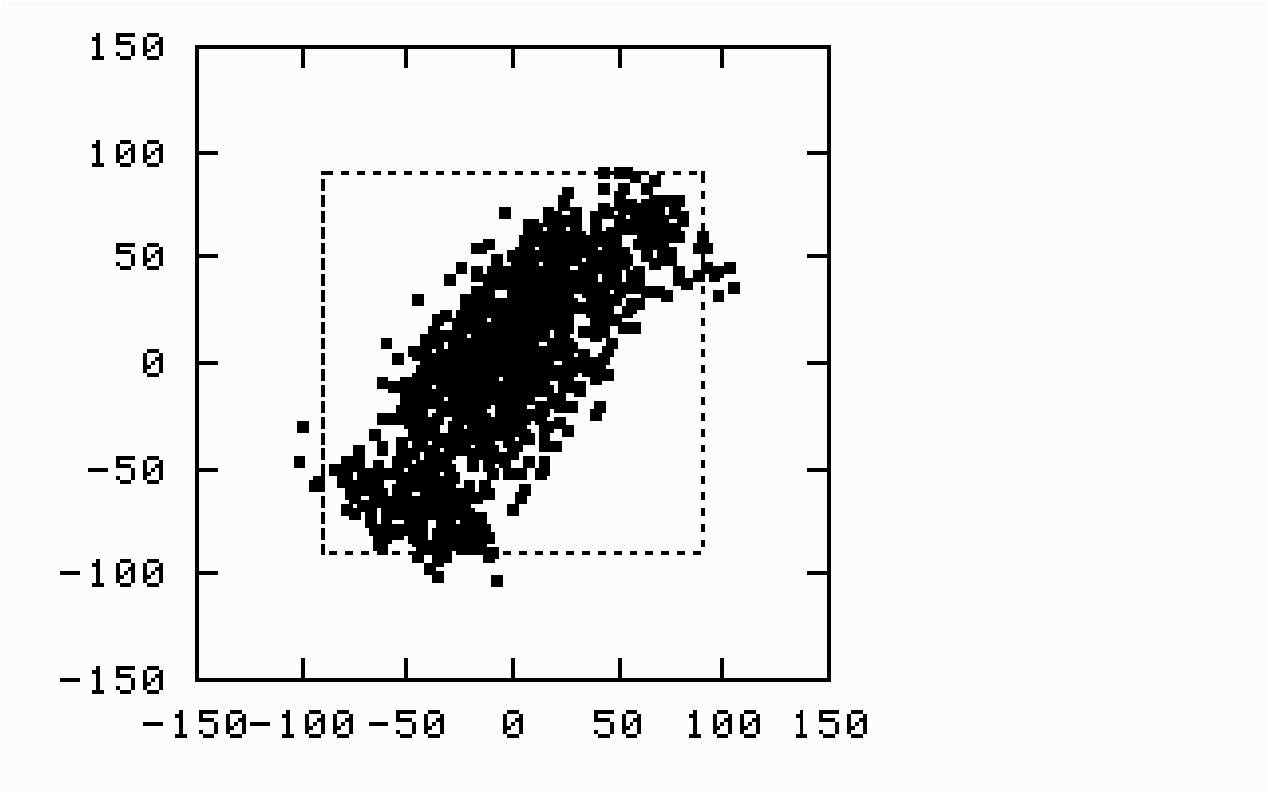


Projections (cont)



Projection

The choice of projection is key:



Correlated Feature Extraction

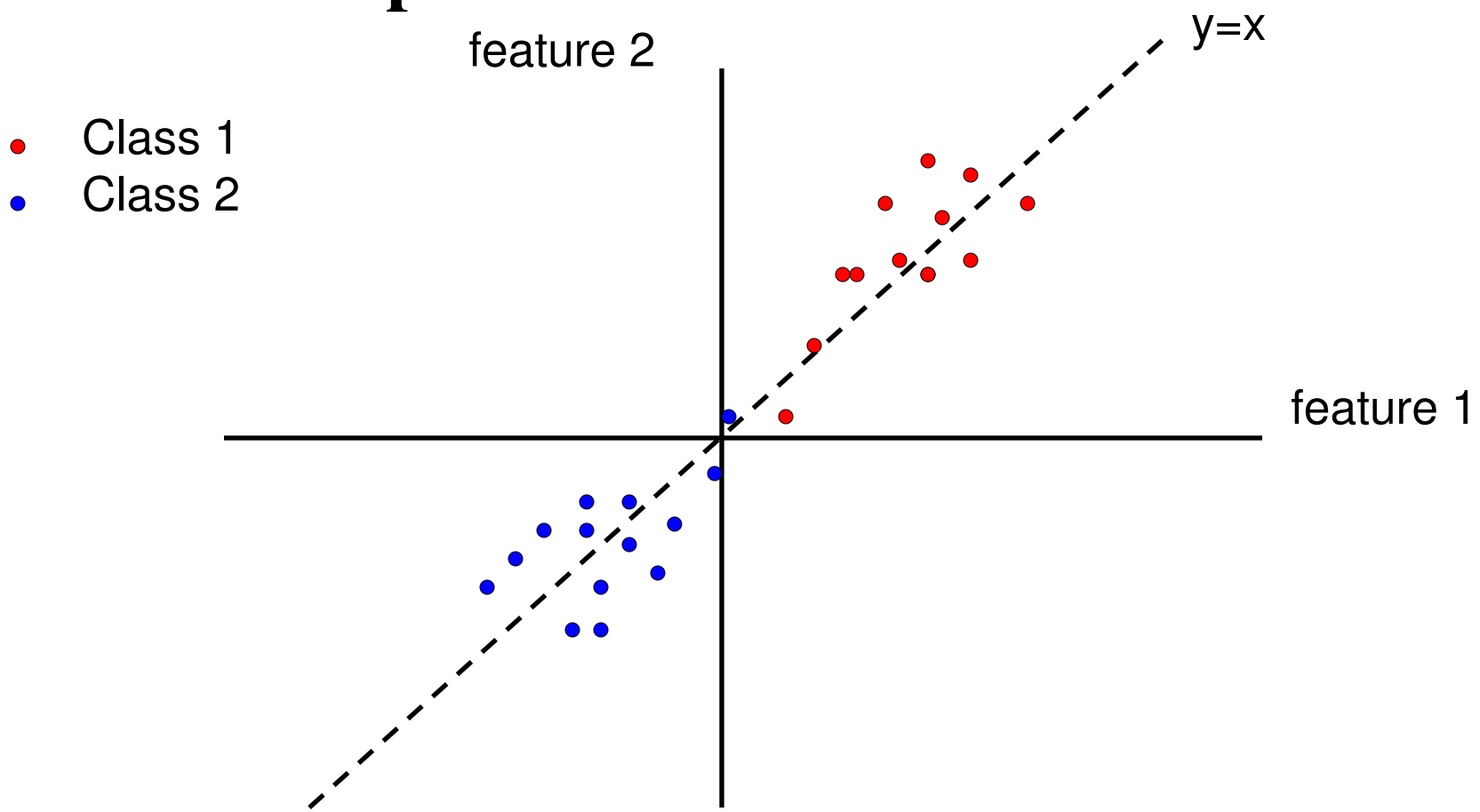
Motivation

- Genes tend to live in families that behave similarly.
- Microarray expression levels are often highly correlated.

The Idea:

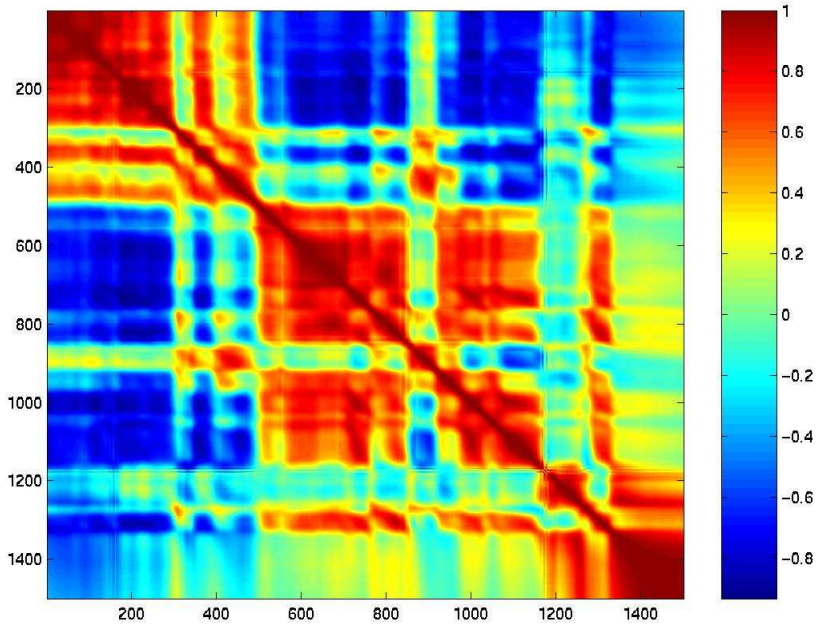
- If 2 features “look” the same, why keep both of them?
- Replace a cluster of correlated features by their mean.
- Doesn’t significantly distort data. Makes some biological sense.
- This is a projection onto a space spanned by indicator functions on the clusters.

CFE example

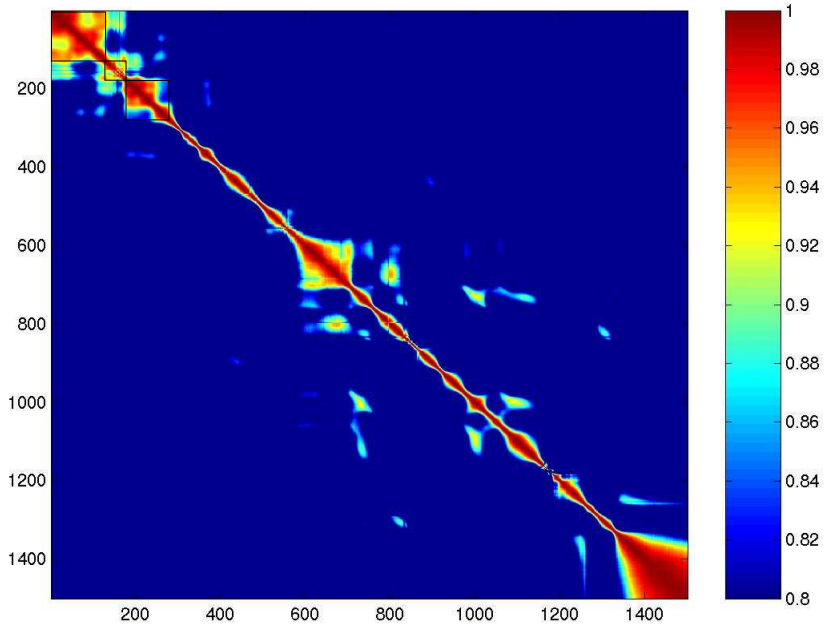


CFE

Correlation between features



Correlations >.8 shown. Boxes show regions to be averaged over



NRC · CNRC

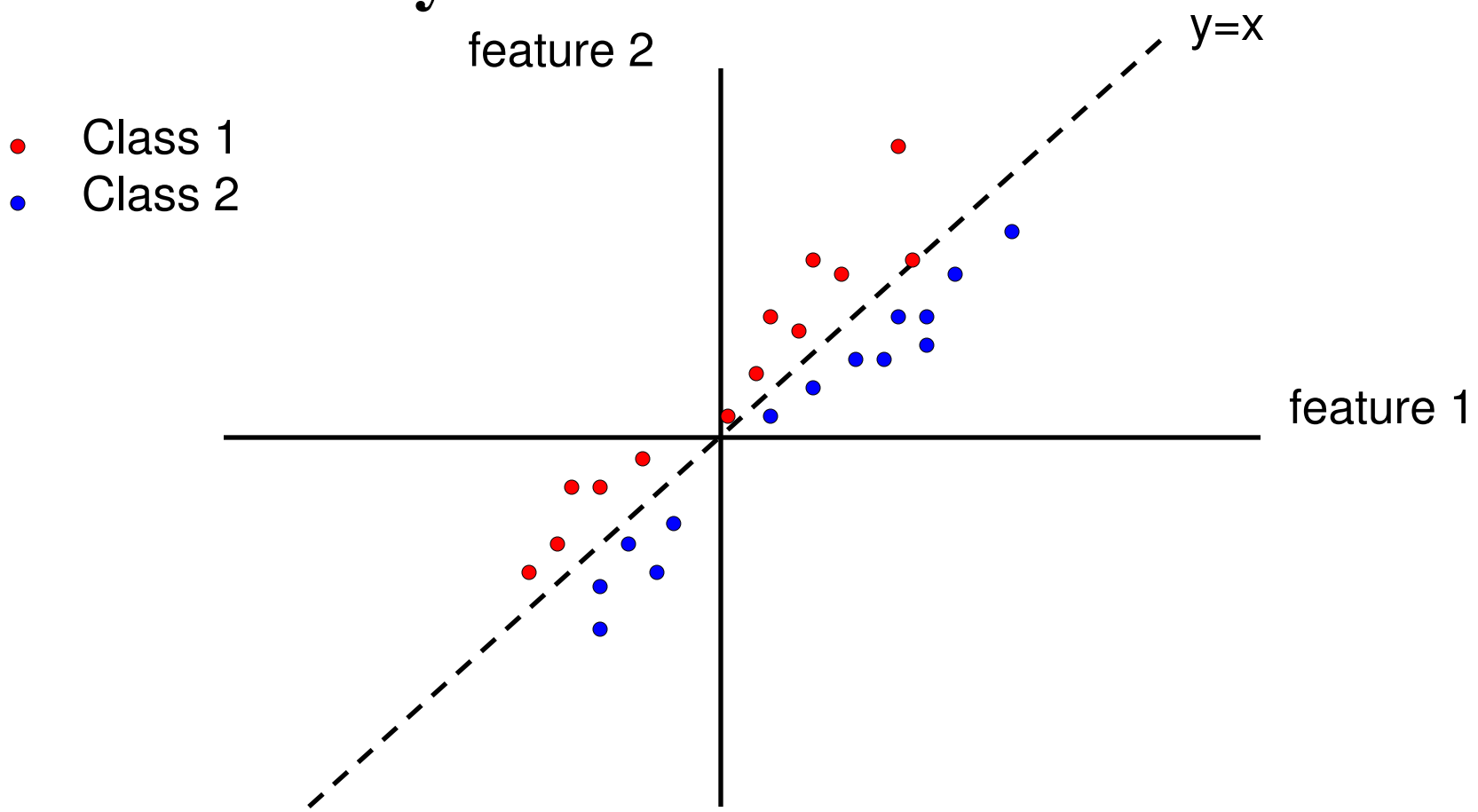
Correlated Feature Extraction (cont)

- Produce a clustering of features that preserves biological relevance.
- That is find all features that lie near each other that correlate higher than some threshold.

CFE drawbacks and possible extensions

- Like all correlation based approaches, CFE is limited to only second order statistics of the data.
- Can in principle be fooled by switching modes – I.e. two features which are highly correlated yet still discriminatory. Can only be addressed by being somewhat conservative in dimension reduction target. The lower the threshold, the more likely to lose a discriminatory feature.
- Doesn't produce a low dimensional projection suitable for visualization, typically only good down to about 20 dimensions or so.

Correlated yet distinct features

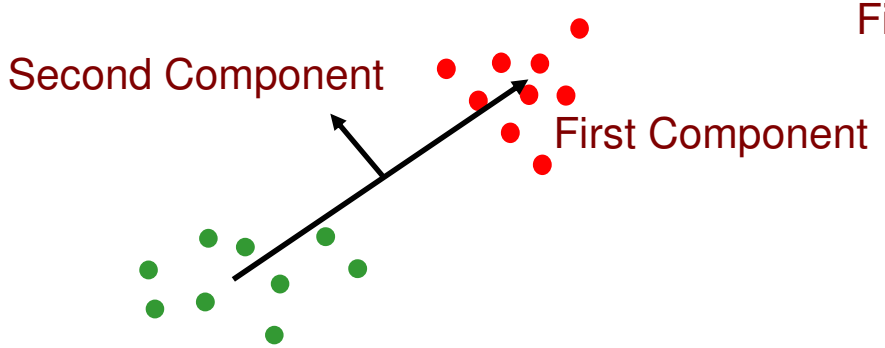


PCA

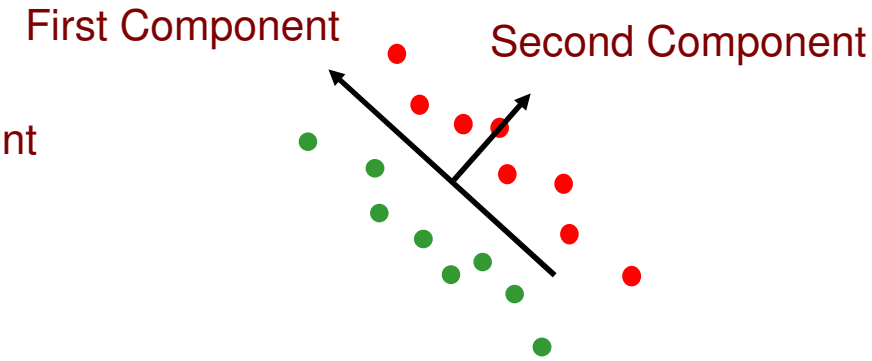
- PCA (Karhunen-Loève) is an old and well understood feature extraction method which is (in some sense) the optimal linear method based on second order statistics.
- Find direction of greatest variance in data, then orthogonal direction with next greatest variance, etc. Dimension reduction is achieved by truncating the expansion after a fixed number of principal components.
- Project data linearly onto the subspace spanned by the principal components that are being retained.
- The subspace spanned by the principal components is what's usually important, not the components themselves (not always true).
- Classical technique – many variants exist.

PCA

- Finds directions of maximum variance in data.
- Maximum variance not always important or discriminatory.



Case I: Higher PC is important for discrimination, projection onto PCs is useful



Case II: Higher PC is NOT discriminatory, projection onto PCs is not helpful

- Case II happens often in practice!

Rotated PCA

- Rotated PCA grew out of the psychology and factor analysis community
- People were looking for a way to increase interpretability of PCA features.
- The idea: PCA gives one choice of a basis for the space spanned by the principal components. Instead rotate (or nearly rotate) this basis so that it still spans the same space, but so that coefficients in new basis are more easily interpretable.
- Many criteria for this, varimax, quartimax, etc.
- Similar drawbacks to PCA, primarily based on the linearity of the method.

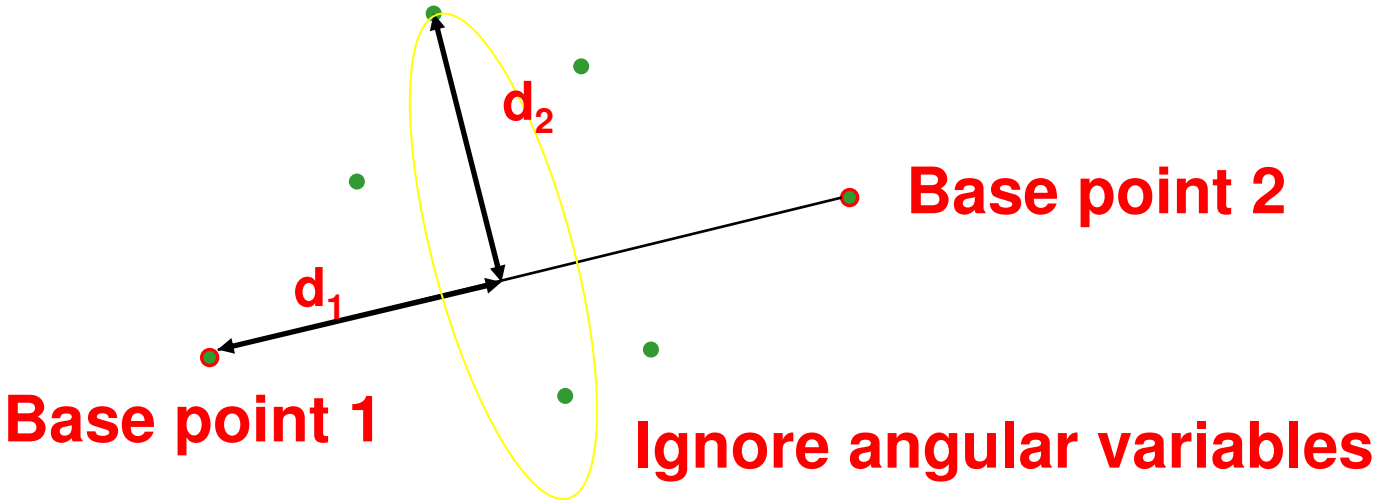
Local PCA

- Cluster data locally and perform PCA on each cluster, the somehow “glue” the PCA subspaces together.
- Crude nonlinear method. Has been applied a lot in practice.
- Difficult to piece together sections of linear representation into a global whole.
- Results depend on parameters.

RDP

- Given 2 anchor points in d dimensions, can find a mapping down to 2 dimensions so that the pair of distances between an arbitrary third point to the 2 anchor points is preserved.
- Generalization of cylindrical polar co-ordinates – Keep *r* and *z*, throw away all angles.
- Given a set of data-points, can perform this mapping using all pairs in the set as anchor points, searching through for “best” pair.
- Useful for visualization.
- “Poor man’s” projection pursuit.

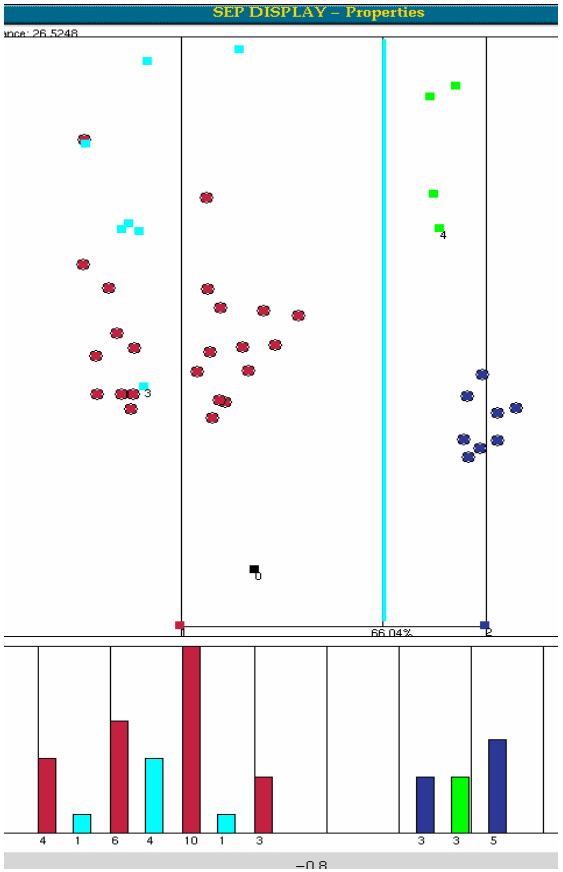
RDP algorithm

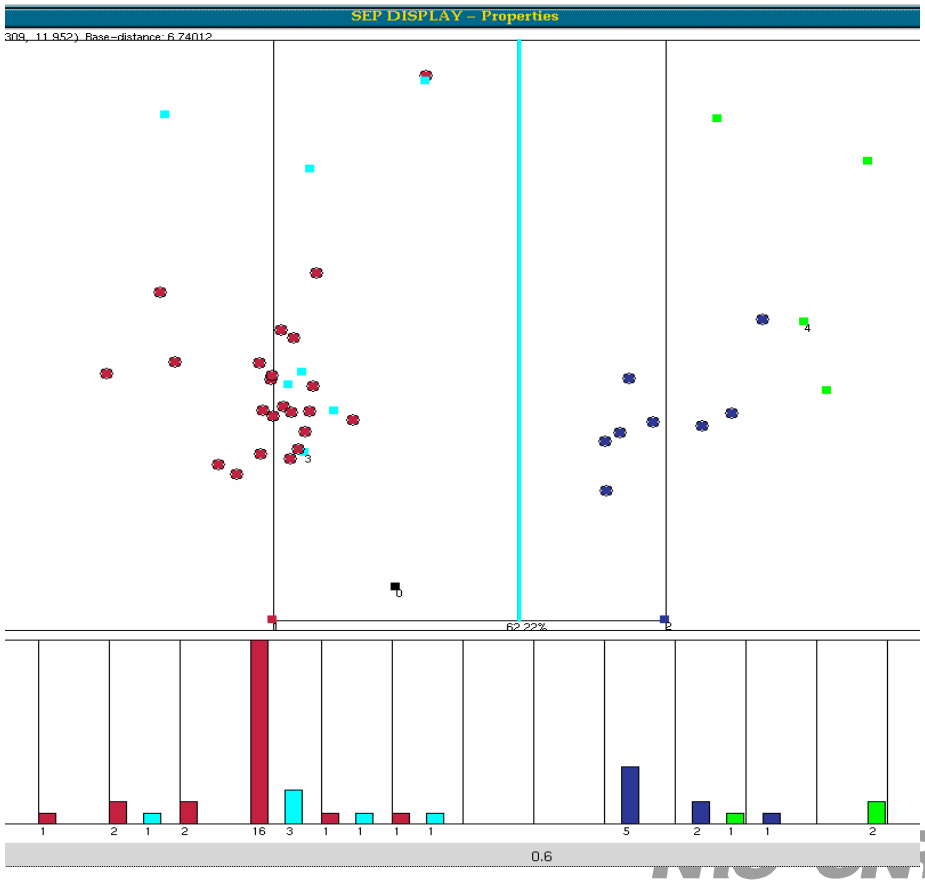


SRBCT Cancers: EWS (23) vs. BL (8):

Visualization by RDP Maps, from

2308 Dimensions 2 Dimensions





RDP Advantages

- Fast, requires little computation.
- Model-free.
- Flexible, allows for exploration.
- Takes class labels into account.
- Poor man's projection pursuit – projection pursuit is unfeasible with the number of features in a typical microarray experiment.
- Unlike PCA isn't based on variances. Won't be fooled when max variance direction isn't discriminatory.

Local Linear Embedding

- Interesting idea, from Roweis and Saul (*Science* v.290 no.5500, Dec.22, 2000. pp.2323--2326).
- Basic plan: If data is clustered around a manifold in high dimensional space, it should be locally flat.
- Local linear information can be used to construct globally nonlinear shapes.
- The idea's been around for a while – local PCA introduced in early 90's.
- The beauty of LLE is that it gives a way to "glue" together the local linear approximations into a smooth global parametrization

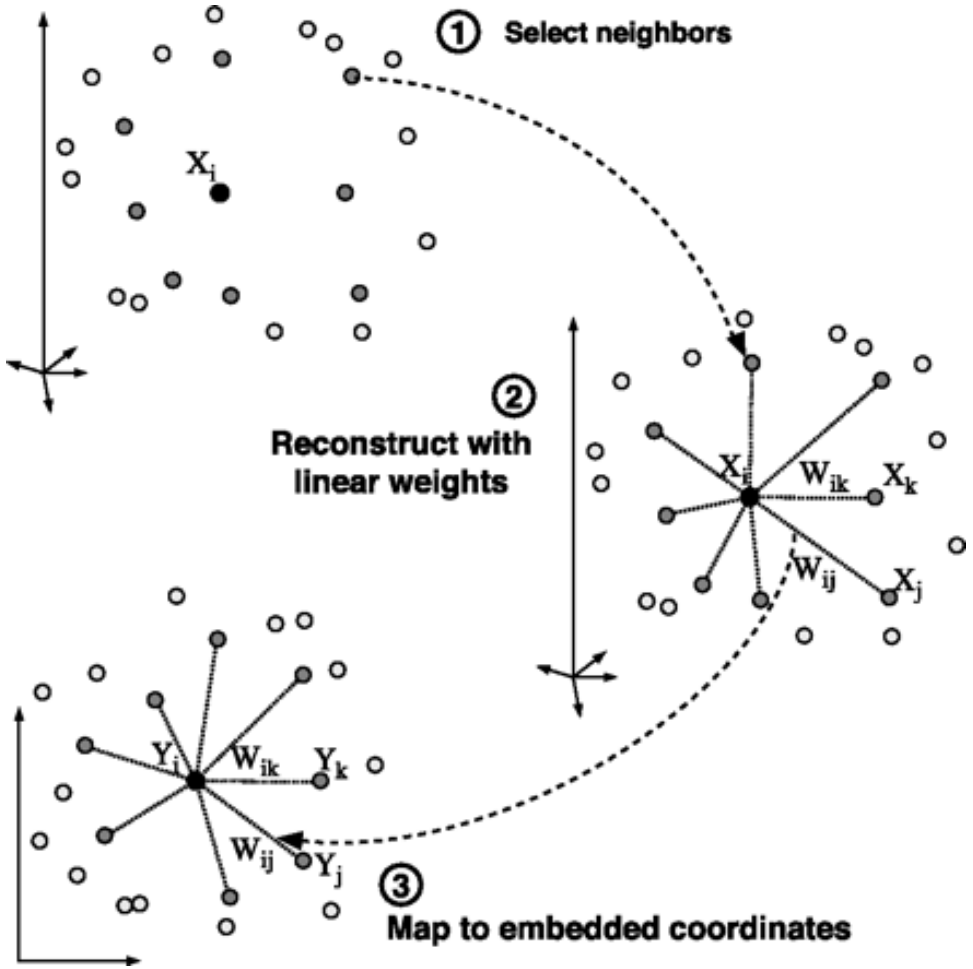
LLE (cont)

- Originally applied to vision.
- Take a scene and fly a camera snapping pictures as you go through it.
- You know that the number of degrees of freedom that the set of pictures contains will be small compared to the (millions of) degrees of freedom represented by the actual images.
- LLE designed to extract this information.
- Application to biomedical data a bit trickier – we don't know how many dimensions our data live on, or even if they cluster around a low dimensional subspace.
- But they had better! Otherwise we aren't collecting NEARLY enough data.
- LLE on microarray data does seem to have some issues with generalization though.

LLE (algorithm)

- Three step process:
  1. Construct distance matrix containing pair-wise distances between all data-points.
  2. Approximate every data-point as a weighted average of its k nearest neighbours. Remember this weight matrix W_{ij}.
  3. Find a set of vectors in a low dimensional space that are also best reconstructed by these W's. These are your reduced co-ordinates.
- All 3 steps can be done fairly quickly – no global optimization required! Comparable in speed to PCA.

LLE (algorithm)

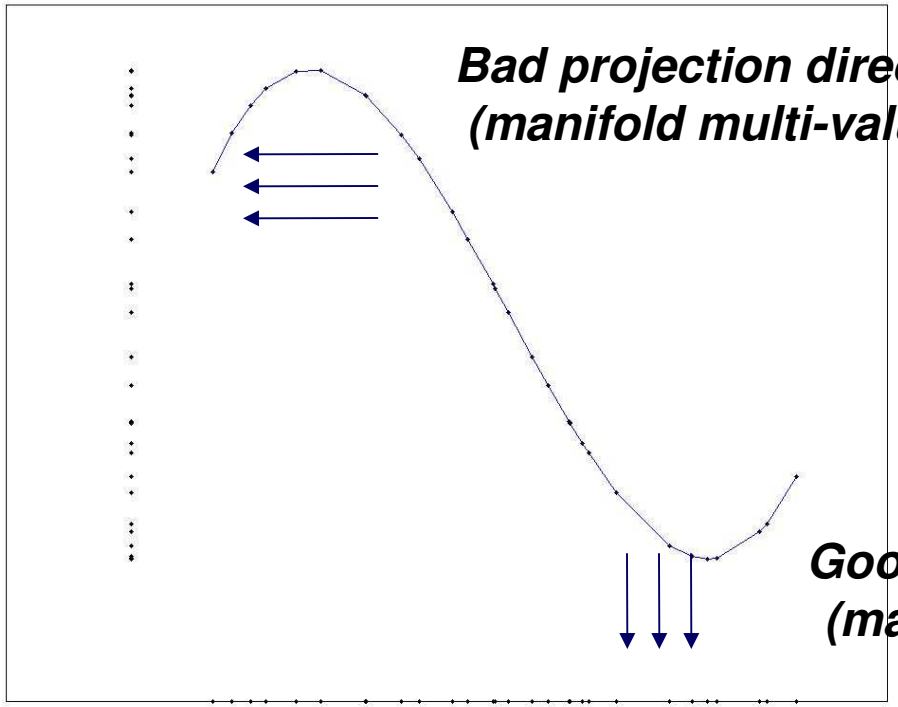


Whitney Reduction Network

- Views the data manifold as the graph of a (non-linear) function over some linear subspace.
- G is the projection of the data onto this subspace, H is the non-linear mapping back.
- Based on Whitney's theorem of differential geometry.
- Proof of Whitney's theorem shows that almost any projection of m-manifold into 2m+1 dimensions is invertible.
- Use this freedom to select a "good" projection – one which is easily invertible.
- What makes a good projection?

Good projections

• Want the data manifold to look single valued when viewed from the reduced space.



Bad projection direction (manifold multi-valued)

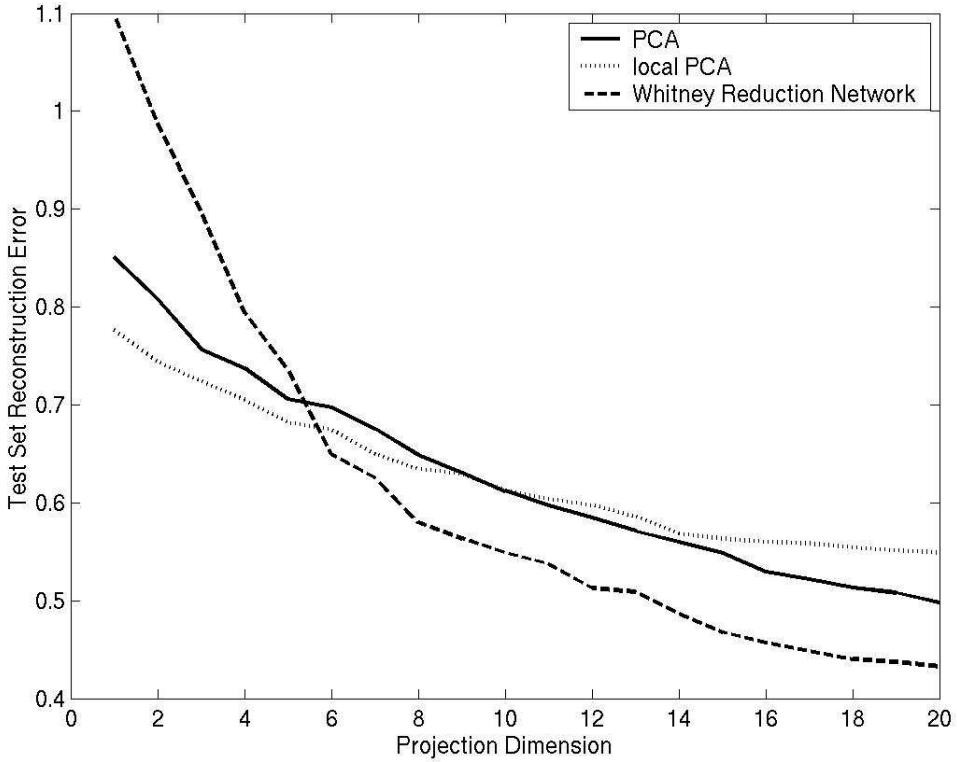
Good projection direction (manifold single-valued)

WRN

- Once projection identified, can locate manifold using function approximation techniques, eg. Radial basis function neural network.
- One big advantage the WRN has is that one can compute reconstruction errors – distances between reduced and original points in the high dimensional space.
- A few drawbacks:
  - Can be slow to identify a good projection.
  - Works with set of all data secants which can get huge for large number of samples.
  - Guaranteed to be sub-optimal – need 2m+1 dimensions for a m dimensional manifold. But close to optimal, so pretty good.
  - Will not work well on very low dimensional reductions.

WRN Results

Cross Validated Reconstruction error:
Mass Spectra



A few references

General Dimension Reduction and Motivation

- ``High dimensional data analysis - the curses and blessings of dimsionality'' Donoho, D. at AMS conference ``Math Challenges of the 21st Century'' Los Angeles, August 6-11, 2000.

Microarrays

- ``Use of a cDNA microarray to analyse gene expression patterns in human cancer.'', DeRisi, J. et. al., Nat Genet. 1996 Dec;14(4):367-70.
- Central nervous system genes involved in the host response to the scrapie agent during preclinical and clinical infection.'', Booth S, et. al. J Gen Virol. 2004;85(Pt 11):3459-71.

More references

- “Unsupervised feature dimension reduction for classification of MR spectra.”, R Baumgartner, R Somorjai, C Bowman, T C Sorrell, C E Mountford, U Himmelreich, Magnetic Resonance Imaging. 2004 Feb ;22:251-6
- “Mapping high-dimensional data onto a relative distance plane: an exact method for visualizing and characterizing high-dimensional patterns”, Journal of Biomedical Informatics Volume 37, Issue 5 (October 2004), Pages: 366 - 379
- “The Whitney Reduction Network: A Method for Computing Autoassociative Graphs” Broomhead, D. S. and Kirby, M. J. Neural Computation Volume 13 , Issue 11 (November 2001) pages: 2595 - 2616
- ``A Global Geometric Framework for Nonlinear Dimensionality Reduction'' Joshua B. Tenenbaum, Vin de Silva, and John C. Langford Science 22 December 2000: 2319-2323.
- ``Nonlinear Dimensionality Reduction by Locally Linear Embedding'' Sam T. Roweis and Lawrence K. Saul, Science 22 December 2000: 2323-2326.

What Remains?

- All methods listed have drawbacks, thus this is an open problem.
- Always more to do. Dimension reduction one of the key research areas in data analysis in 21st century.
- Very amenable to mathematical solution.
- Any good ideas?
- *"The coming century is surely the century of data. A combination of blind faith and serious purpose makes our society invest massively in the collection and processing of data of all kinds, on scales unimaginable until recently."* D. L. Donoho