Multivariate Analysis of Gene Expression Data

a geometrical approach

Jens Nilsson

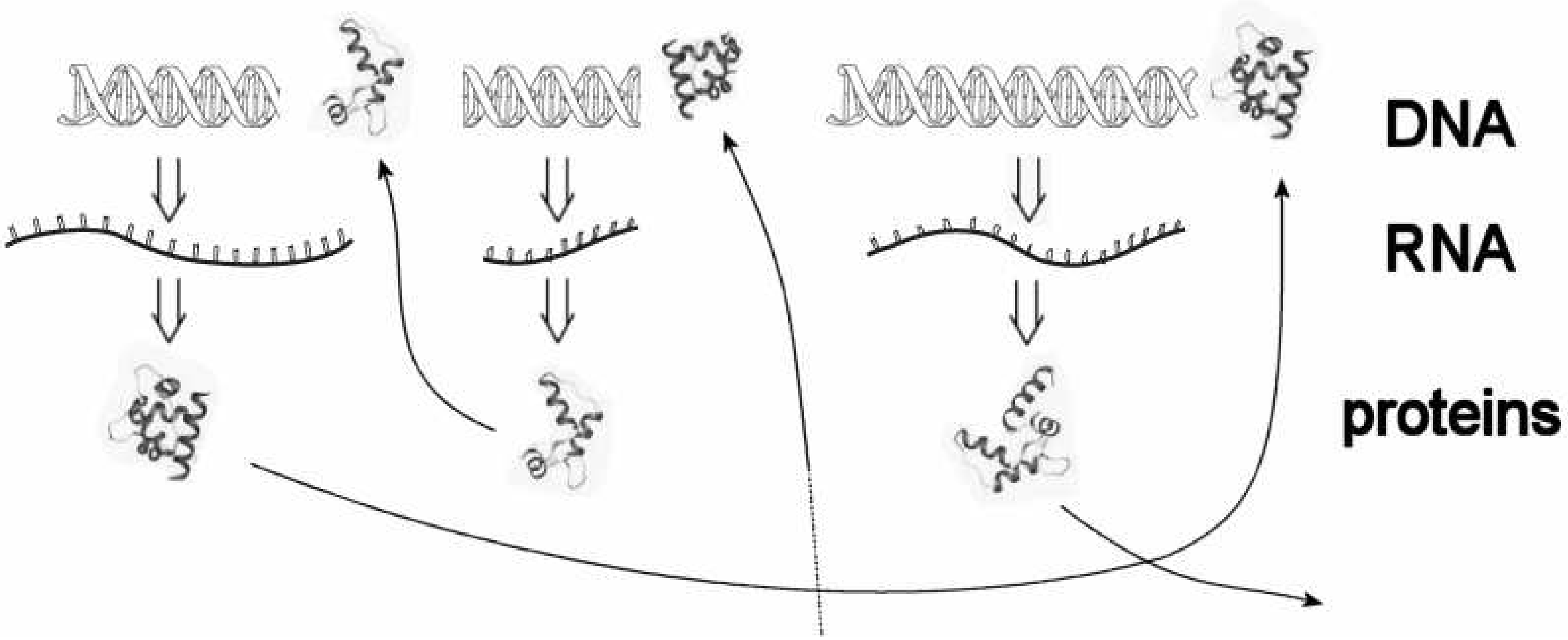
Centre for Mathematical Sciences, Lund University

How can we adress nonlinearities in gene expression data?

Overview

- Why gene expression data may be nonlinear
- Manifold learning
- Exploratory analysis of real data
  - Visualization
  - Variable importance

Genes are functionally related



Regulatory networks

- Some proteins influence the expression of other genes
- Genes interact in a complex dynamical system

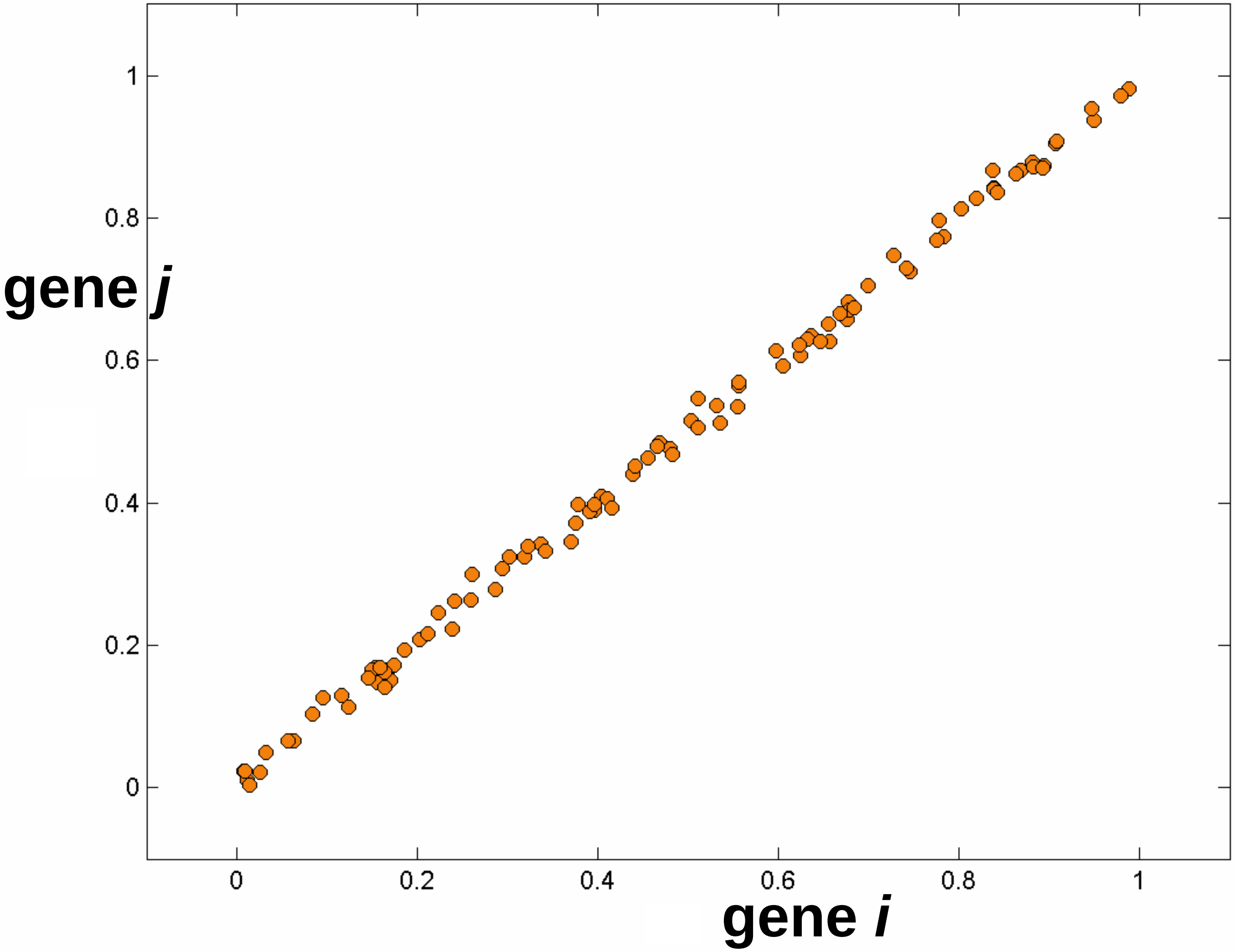
Models of gene regulatory networks

- Boolean networks
- Ordinary differential equations
  - Linear
  - Nonlinear (Michaelis-Menten, etc.)
- Partial differential equations
- Stochastic models
  - Stochastic master equations

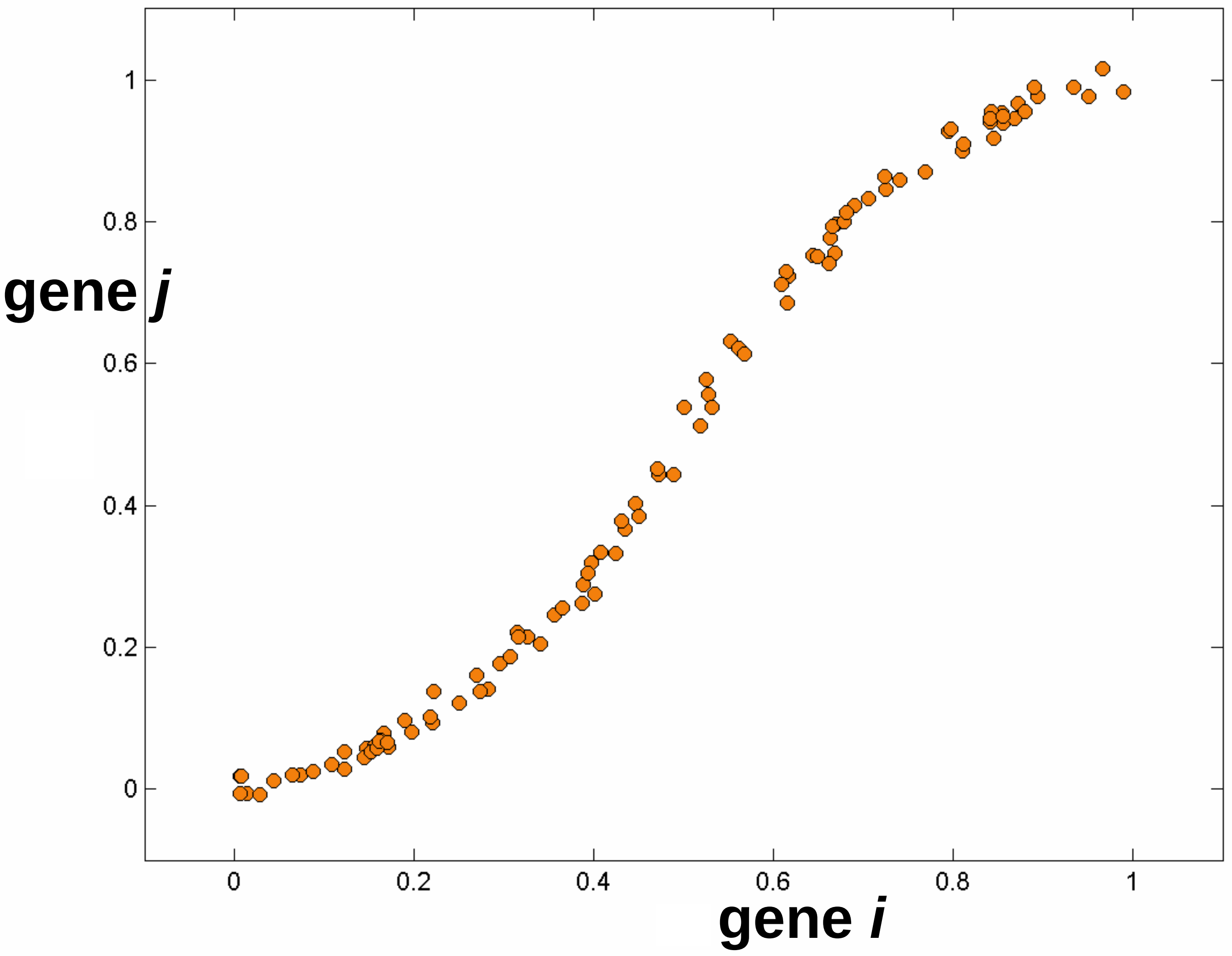
Gene Expression Space

- *m* measurements of *n* genes defines a cloud of *m* points in *n*-dimensional gene expression space
- ... where different domains corresponds to different biological states of the regulatory system
- What is the structure of this point cloud?

Linear relations

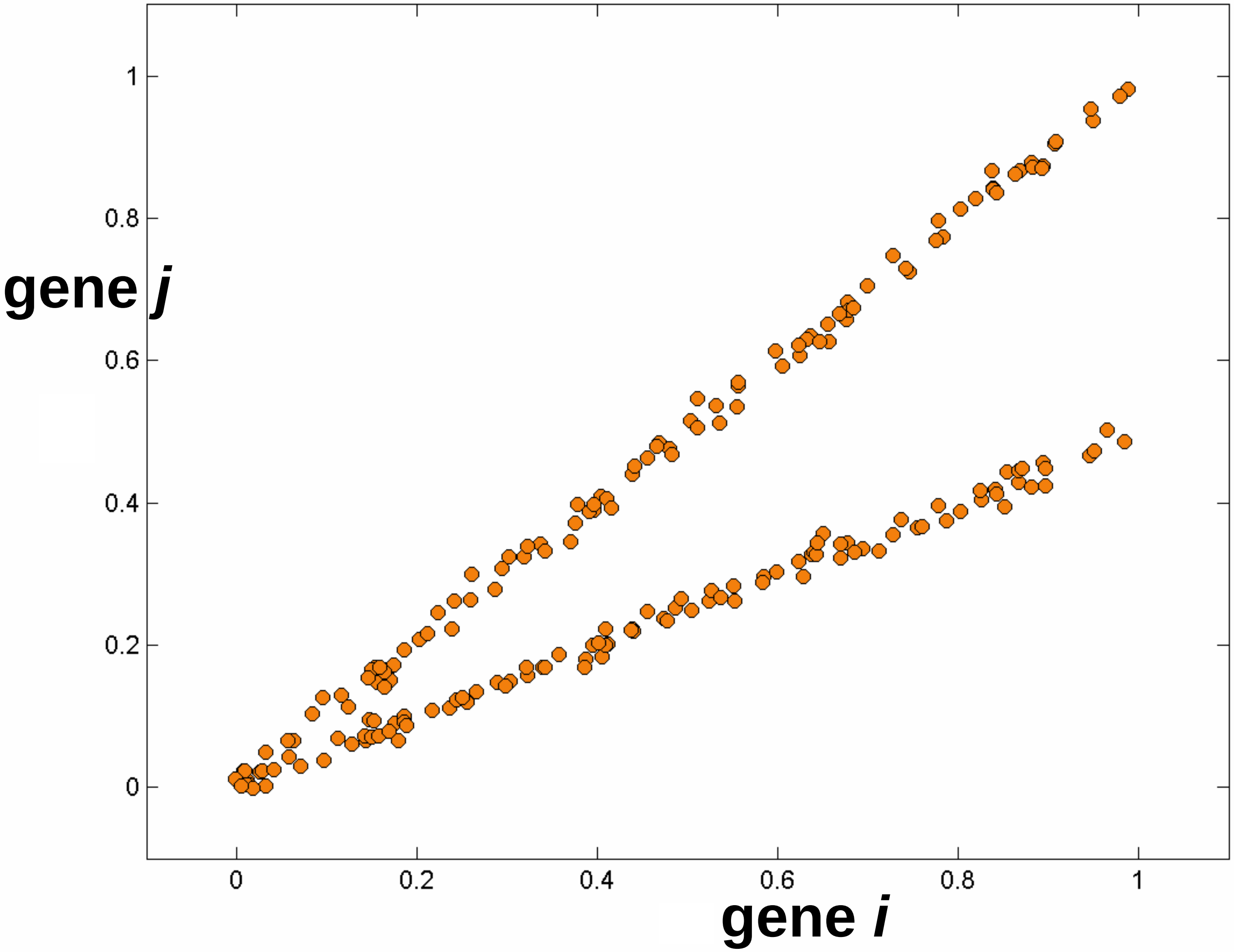


Nonlinear relations



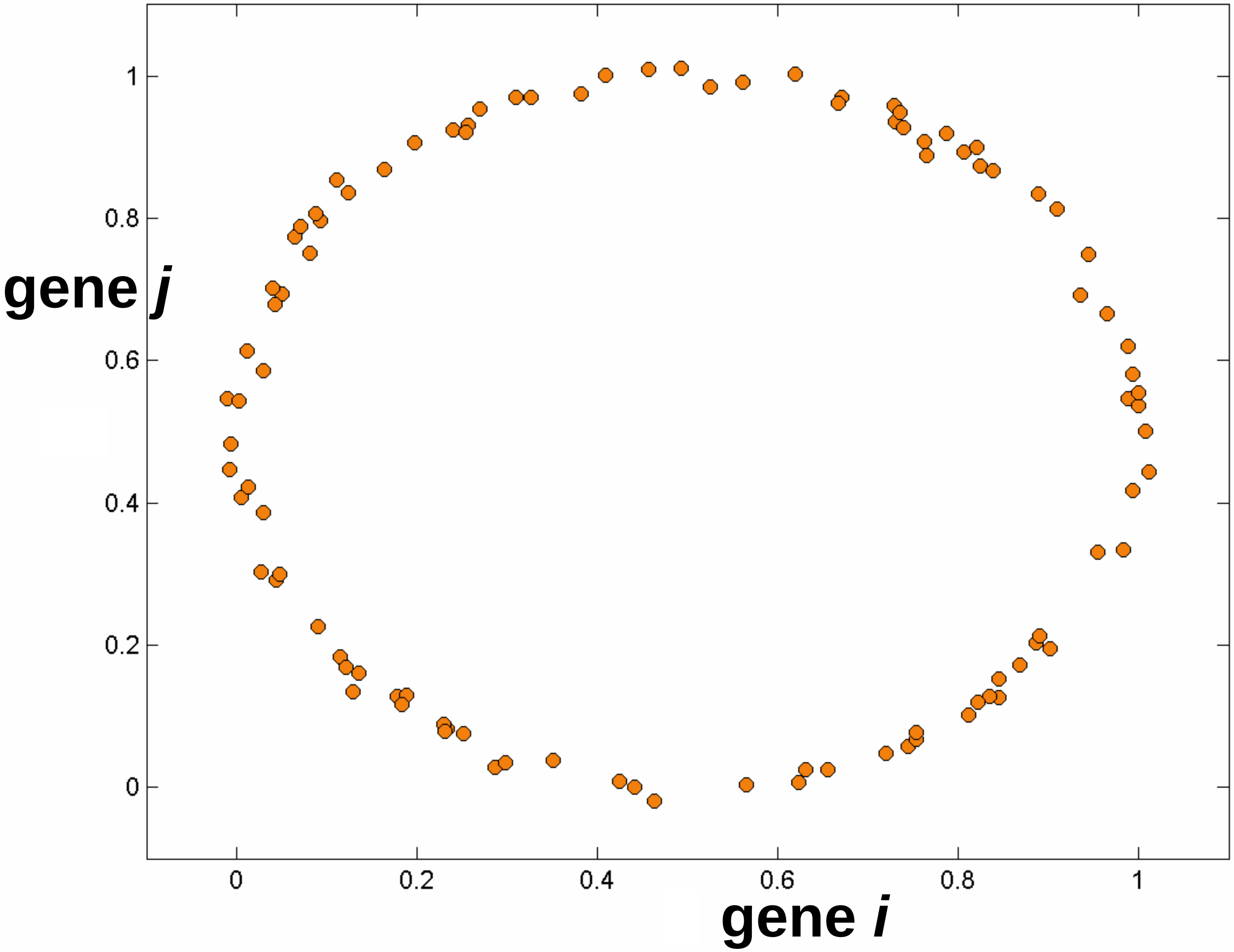
Saturation

Nonlinear relations



Bimodality

Nonlinear relations



Periodicity

Nonlinearities as model consequences

- Spherical geometries in linear ODE’s
  - Spring-mass system
- Torusoidal geometries in nonlinear ODE’s

So far, we conclude…

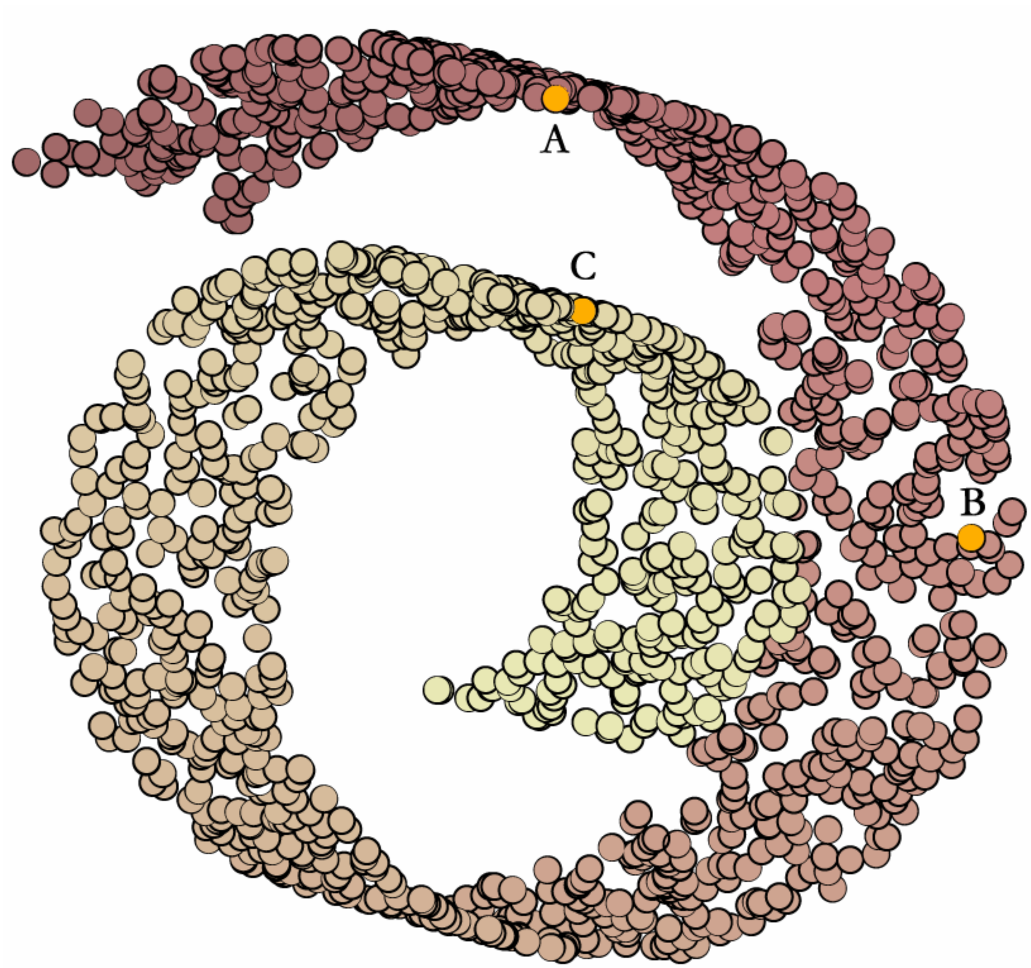
- Nonlinearities may be present in gene expression data

so …

- Data analysis tools should be able to handle this

Manifold Learning

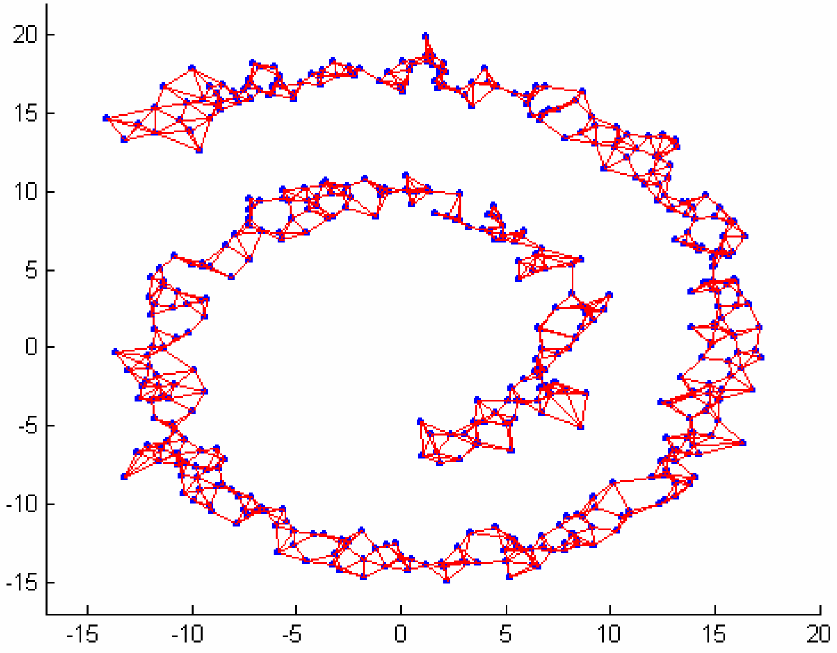
A simple example



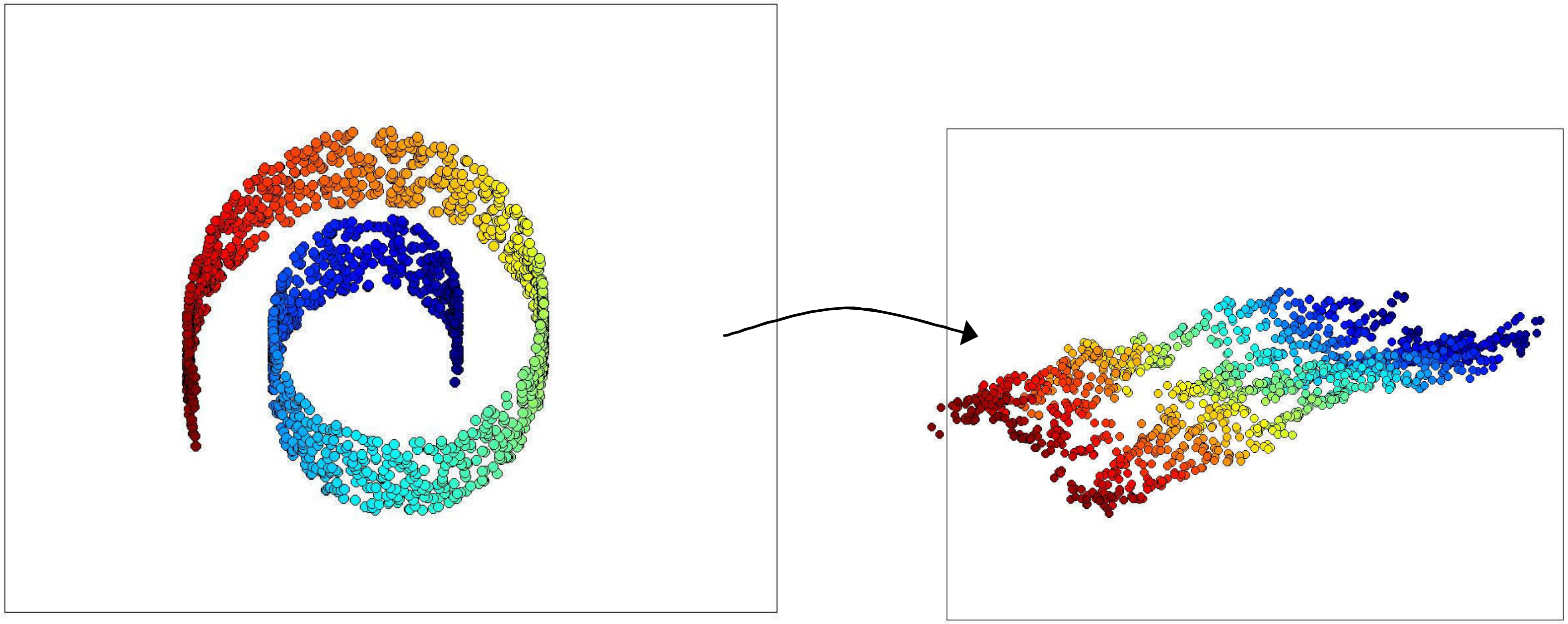
- Geodesic (intrinsic) distance is often more natural than Euclidean (ambient) distance
- Need to infer the manifold structure

Isomap

- **A graph is constructed by connecting each point to its nearest neighbours.**
- **Approximate geodesic distances are calculated by finding the length of the shortest path in the graph between points**
- **Multidimensional scaling yields a lower-dimensional representation**



Isomap recovers intrinsic geometry



observations

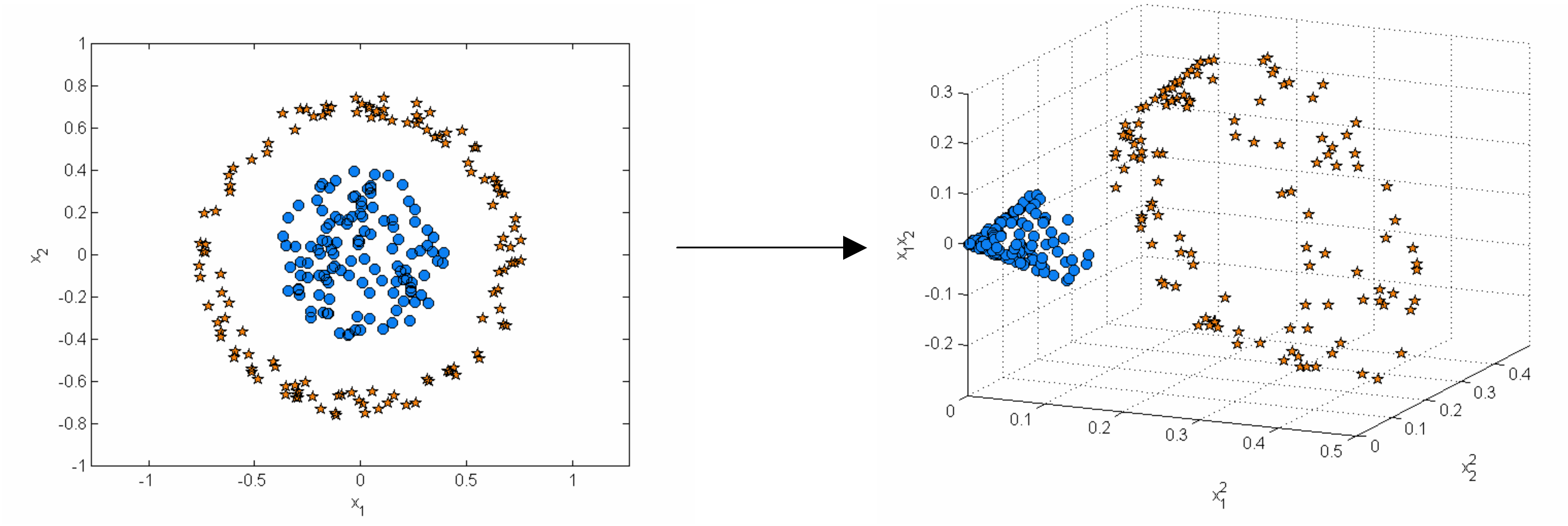
Isomap projection

Some recent methods

- **Isomap (Tenenbaum *et al.* 2000)**
- **Locally Linear Embedding (Roweis & Saul 2000)**
- **Laplacian Eigenmaps (Belkin & Niyogi 2003)**
- **Hessian Eigenmaps (Donoho & Grimes 2003)**
- **Maximum Variance Unfolding (Weinberger *et al.* 2004)**
- **...**

Many of these can be formulated as *kernel methods*

Kernel Methods



- Map data into a *feature space* with desirable properties
  - Linearizing
  - Class separating, etc …
- The mapping is defined *implicitly* via a kernel function – the scalar product in feature space

Kernel view of manifold learning

Nonparametric learning of a kernel whose feature space efficiently parameterizes the underlying manifold

Visualization of gene expression data

Jens Nilsson, Thoas Fioretos, Mattias Höglund, and Magnus Fontes. Approximate geodesic distances reveal biologically relevant structures in microarray data. *Bioinformatics*, 20(6), 2004.

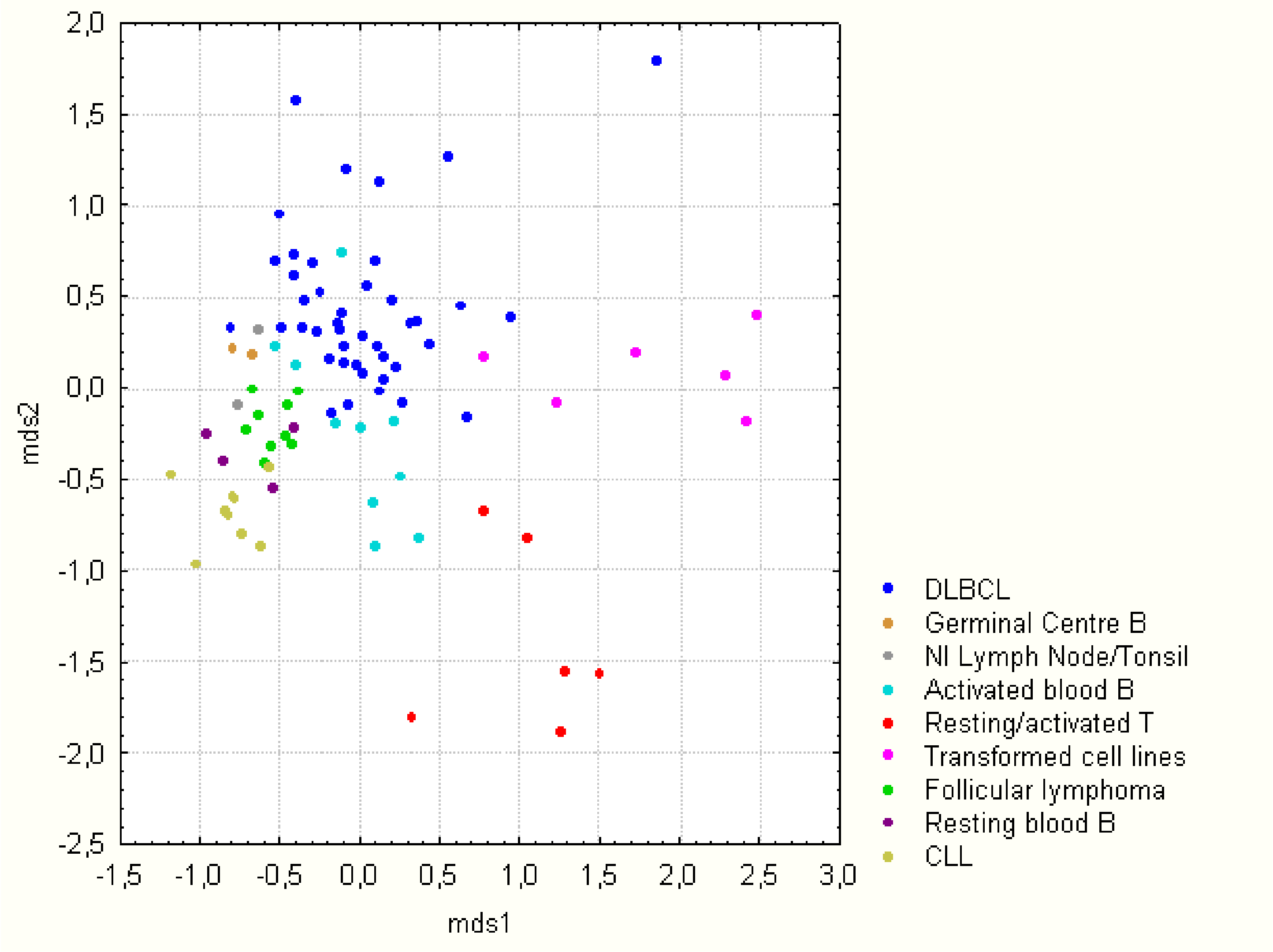
Problem:

- Are approximate geodesic distances more biologically relevant than Euclidean distances?

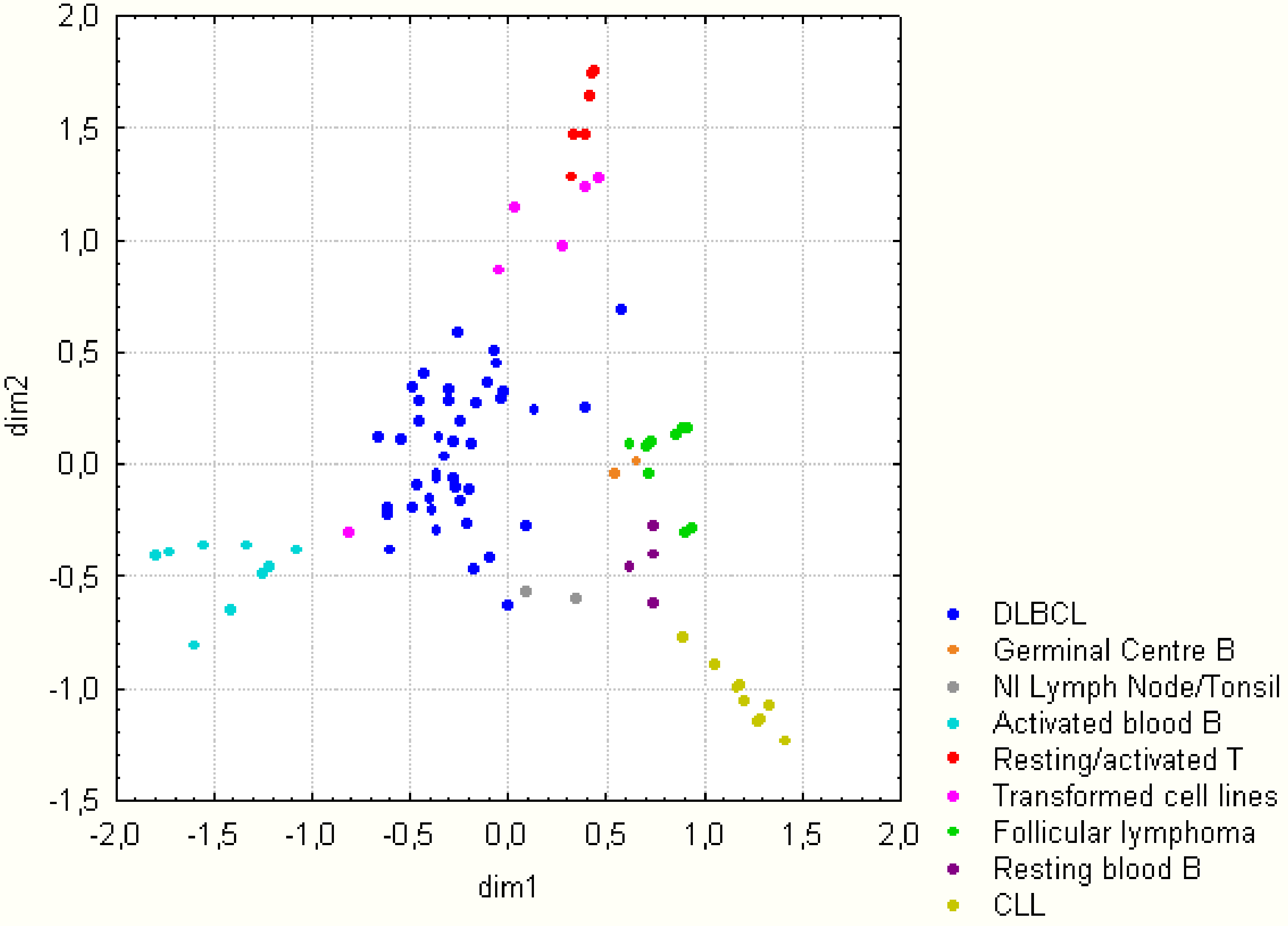
Data:

- 96 lymphoma microarray samples divided into nine diagnostic classes (Alizadeh et al. 2000)

Principal Component Analysis

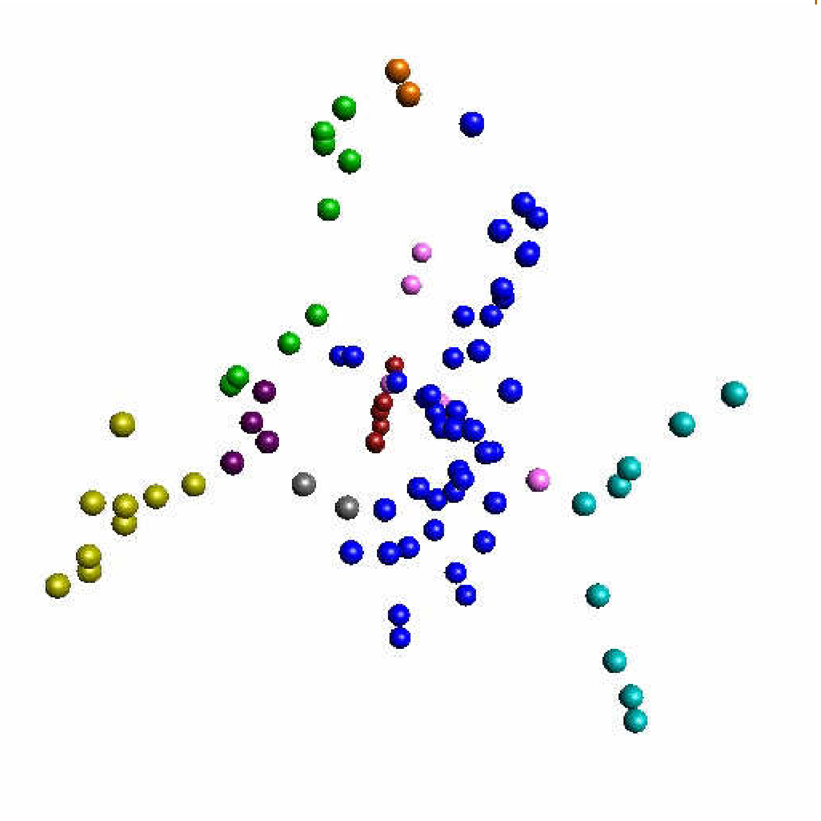


Isomap


2,0
1,5
1,0
0,5
0,0
-0,5
-1,0
-1,5
dim2
-2,0
-1,5
-1,0
-0,5
0,0
0,5
1,0
1,5
2,0
dim1
DLBCL
Germinal Centre B
Nl Lymph Node/Tonsil
Activated blood B
Resting/activated T
Transformed cell lines
Follicular lymphoma
Resting blood B
CLL

Some observations

- Activated blood B class divided into two branches
  - One branch contains all samples stimulated 24-48 h
  - The other contains all samples stimulated 6 h
- Transformed cell lines (TCL) class contains an outlier
  - SUDHL5 is established from a DLBCL tumor
  - Other TCL-samples:
    - 3 based on T-cells
    - 1 of myelomonocytic origin
    - 1 of unknown origin



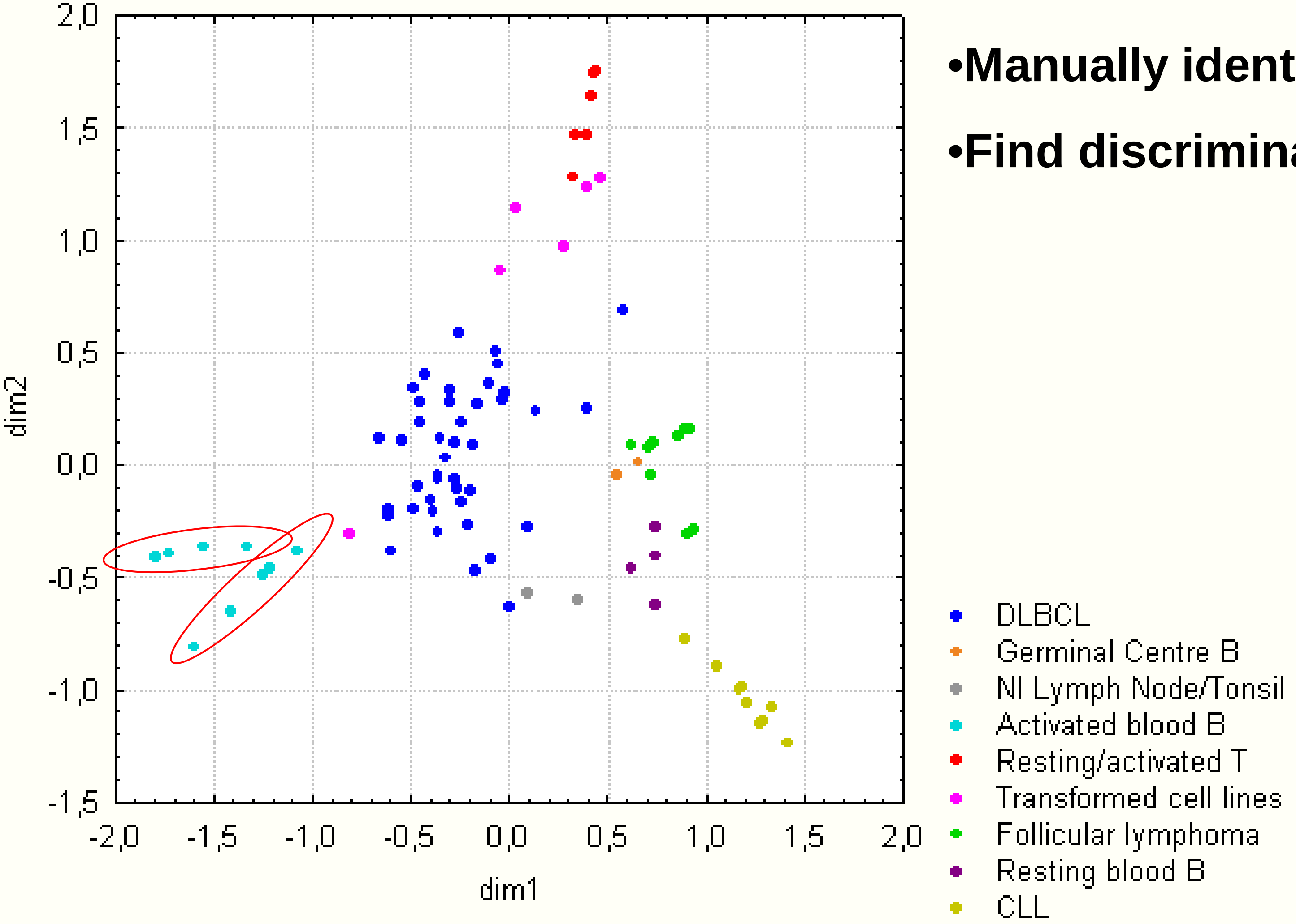
Estimation of variable importance

J. Nilsson, F. Sha, A. Andersson, T. Fioretos, M. Fontes and M. I. Jordan. Variable importance assessment in manifold learning. *Manuscript, 2008*.

Problem

Which variables (genes) are influential for the observed structure?

A possible approach



- **Manually identify subgroups**
- **Find discriminating genes**

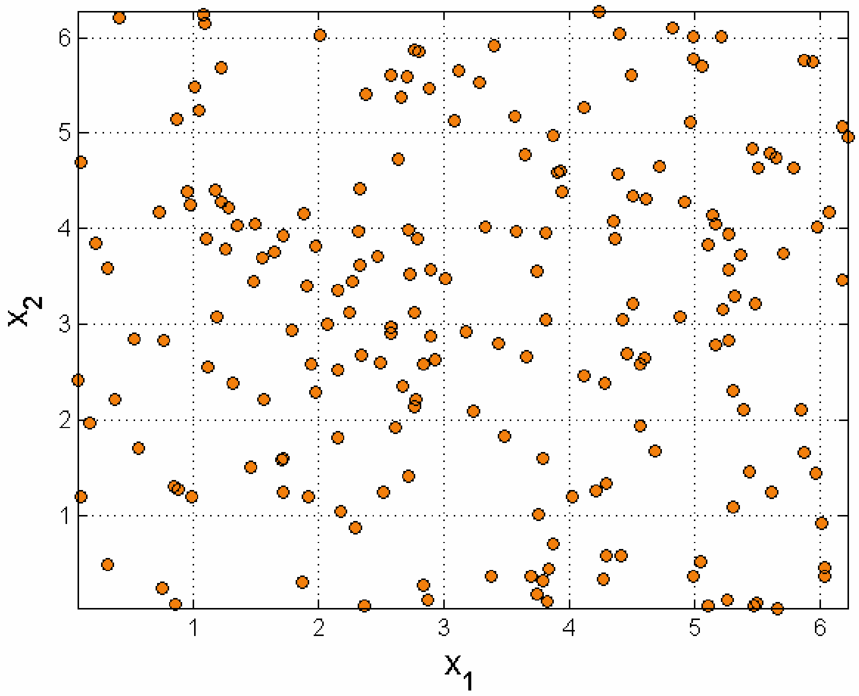
Our approach

- **Manifold learning often involves learning a kernel matrix**
- **… whose feature space describes the intrinsic geometry**
- **Analyze the kernel mapping to assess variable importance**
  - **Feature space – response**
  - **Input space – covariate/explanatory variable**

Sufficient Dimension Reduction (SDR)

- **Li (1991)**
- **Dimensionality reduction in regression**
  - **Covariates *X* and response *Y***
- **Find a linear subspace *Z* of covariate space that is optimally informative w.r.t *Y***
  - ***Central subspace***

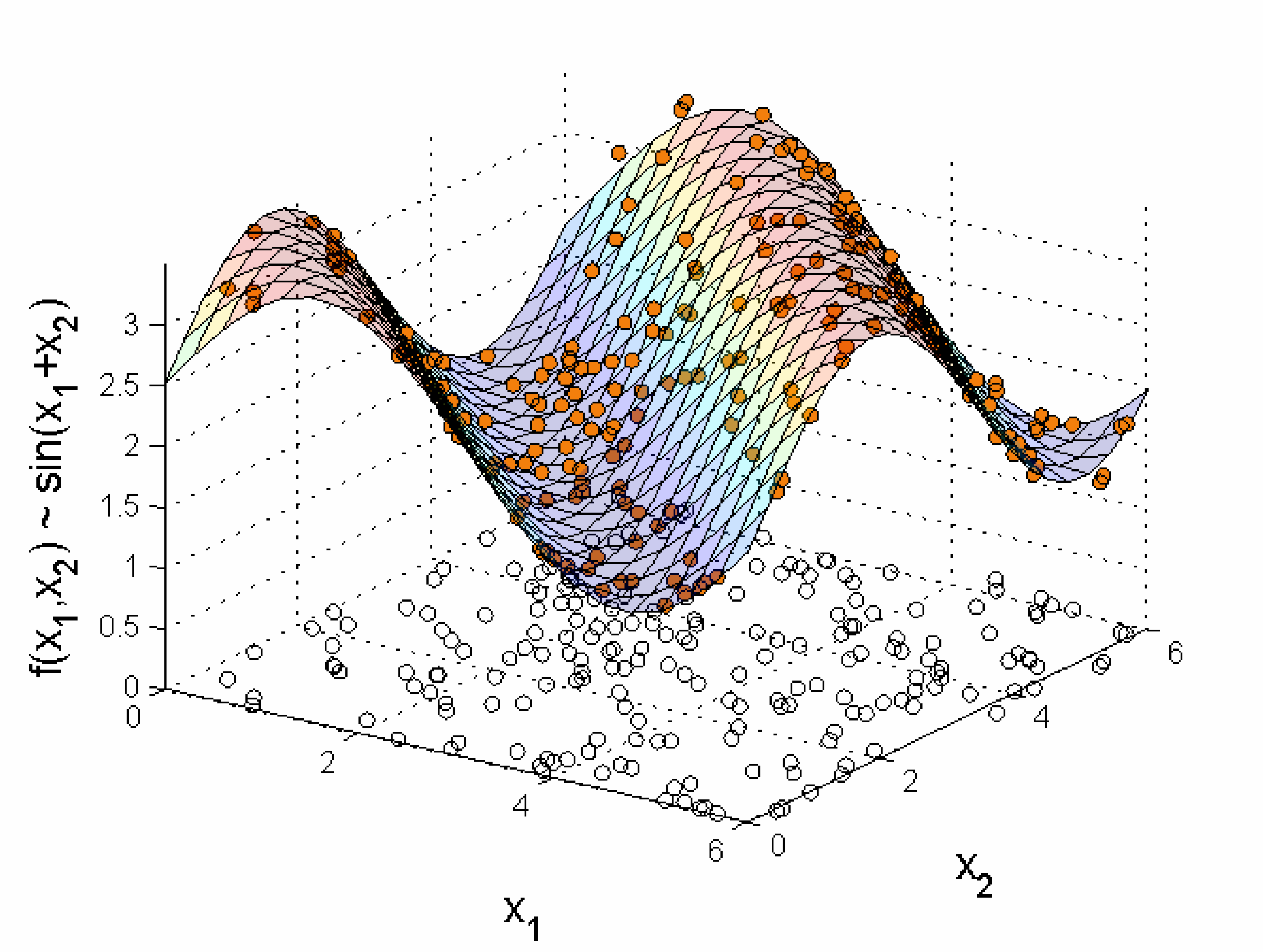
SDR example



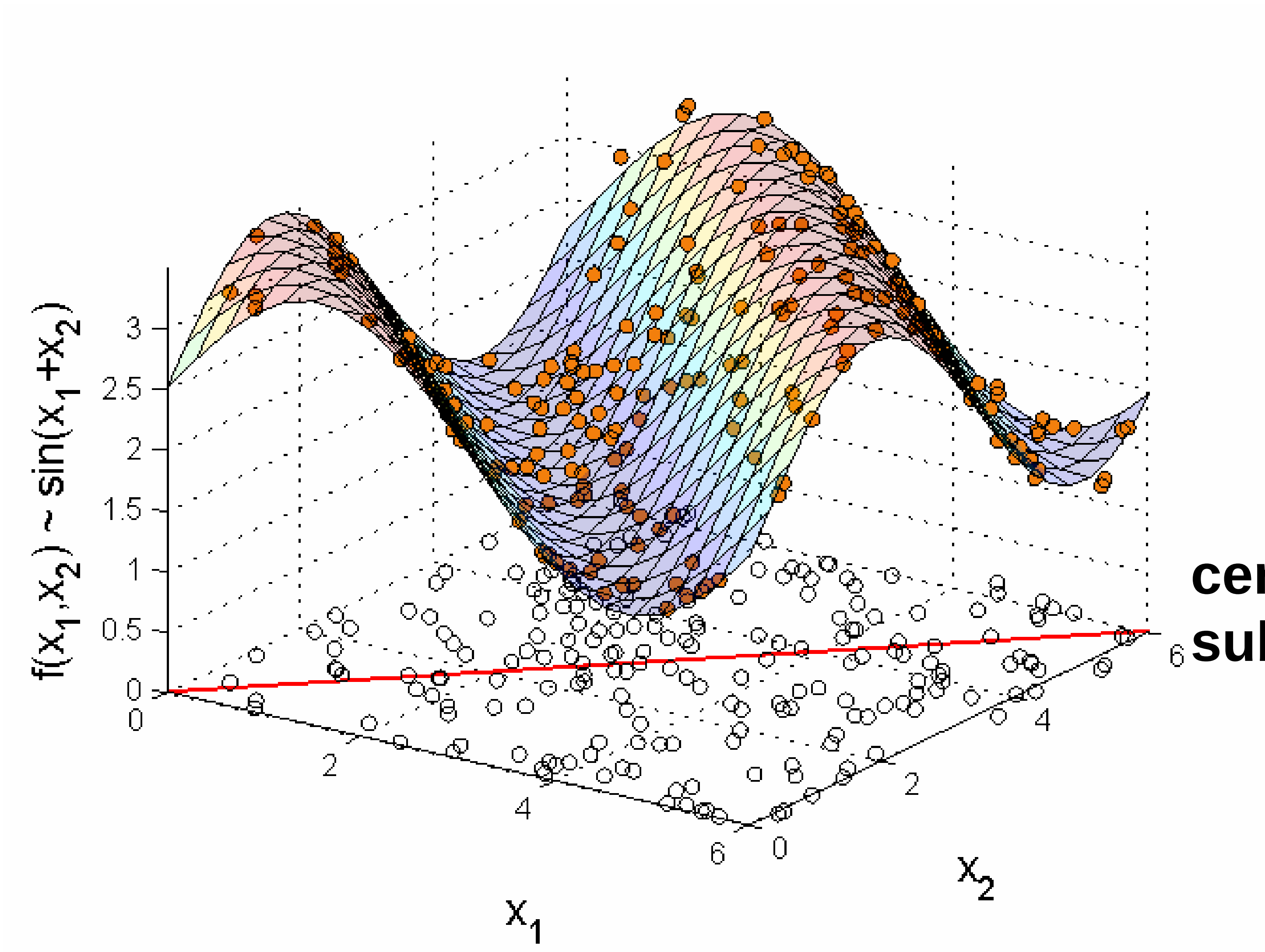
- Consider two covariates x_1, x_2
- Let a response y be defined by

  $y = \sin(x_1+x_2) + \text{const}$

SDR example



SDR example



central subspace

SDR criterion

- Parameterize $\mathcal{Z}$ by $B \in \mathbb{R}^{D \times d}$ where $B^\top B = I$
- Find *B* such that $Y \perp X \mid B^\top X$

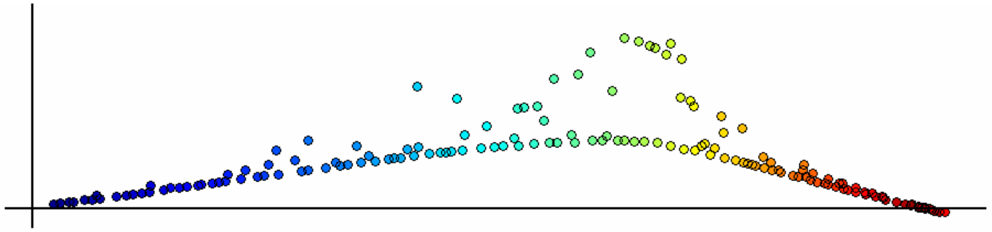
- No assumption on the functional form of the regression relation between *X* and *Y*

Kernel Dimension Reduction (KDR)

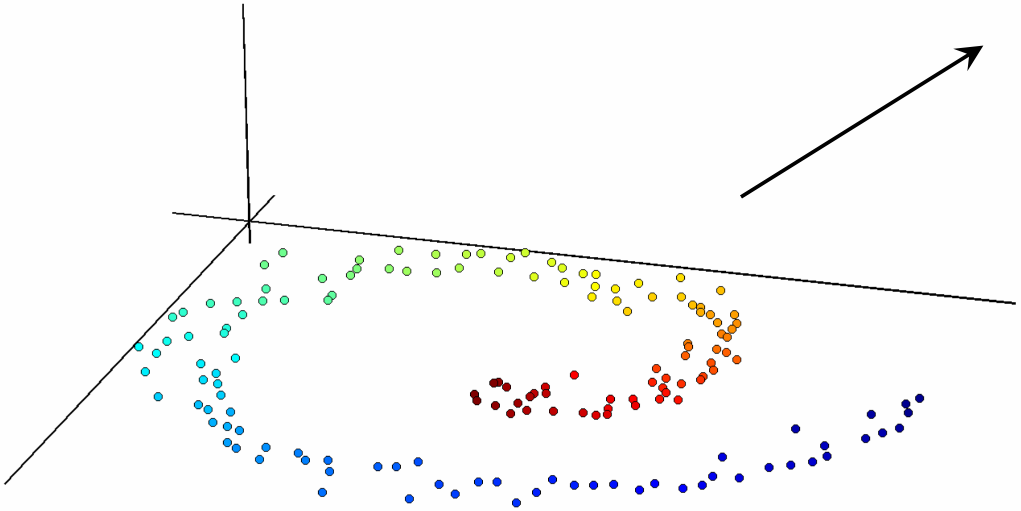
- **Fukumizu, Bach & Jordan (2004, 2006)**
- **Finds the central subspace under weak conditions on *X* and *Y***
- **Kernel formulation**
- **Minimization problem**

$$\begin{array}{ll} \textbf{min} & \operatorname{Tr} ||\boldsymbol{K}_Y^c(\boldsymbol{K}^c_{B^\top X} + N\epsilon \boldsymbol{I})^{-1}|| \\ \textbf{subj. to} & B^\top B = \boldsymbol{I} \end{array}$$

KDR in manifold learning

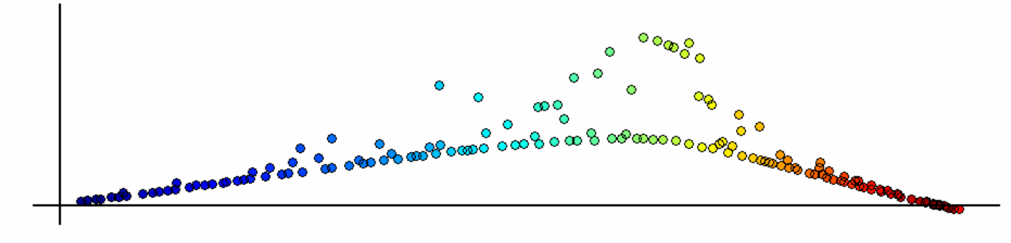


Manifold kernel feature space

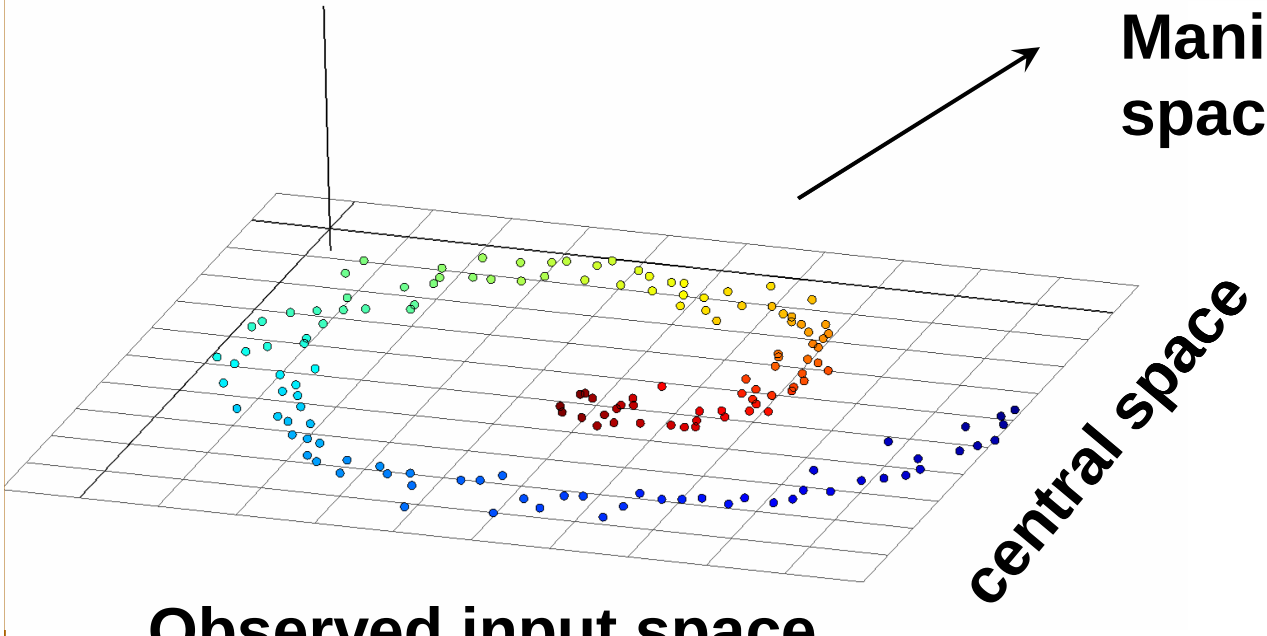


Observed input space

KDR in manifold learning



Manifold kernel feature space



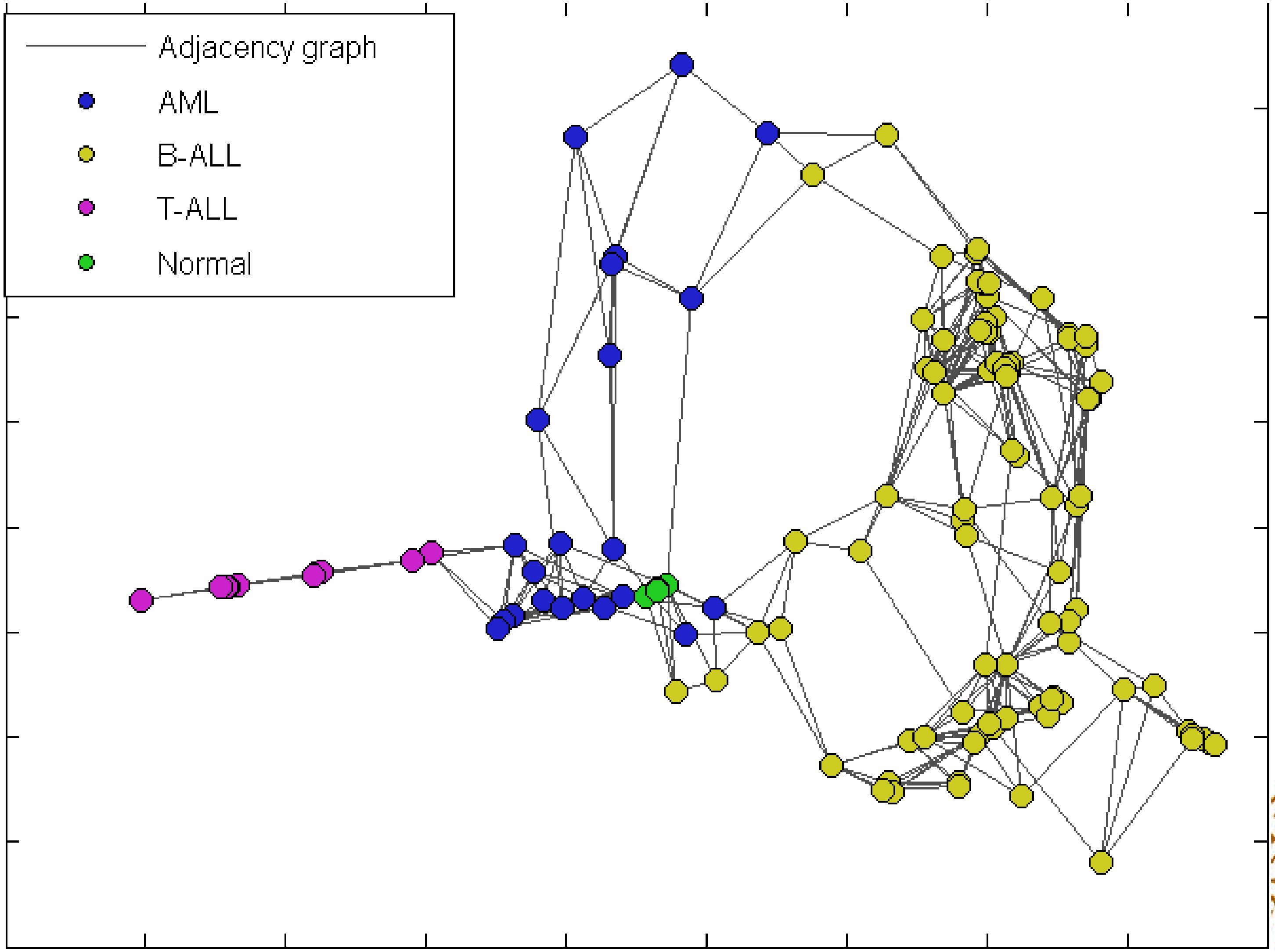
central space

Observed input space

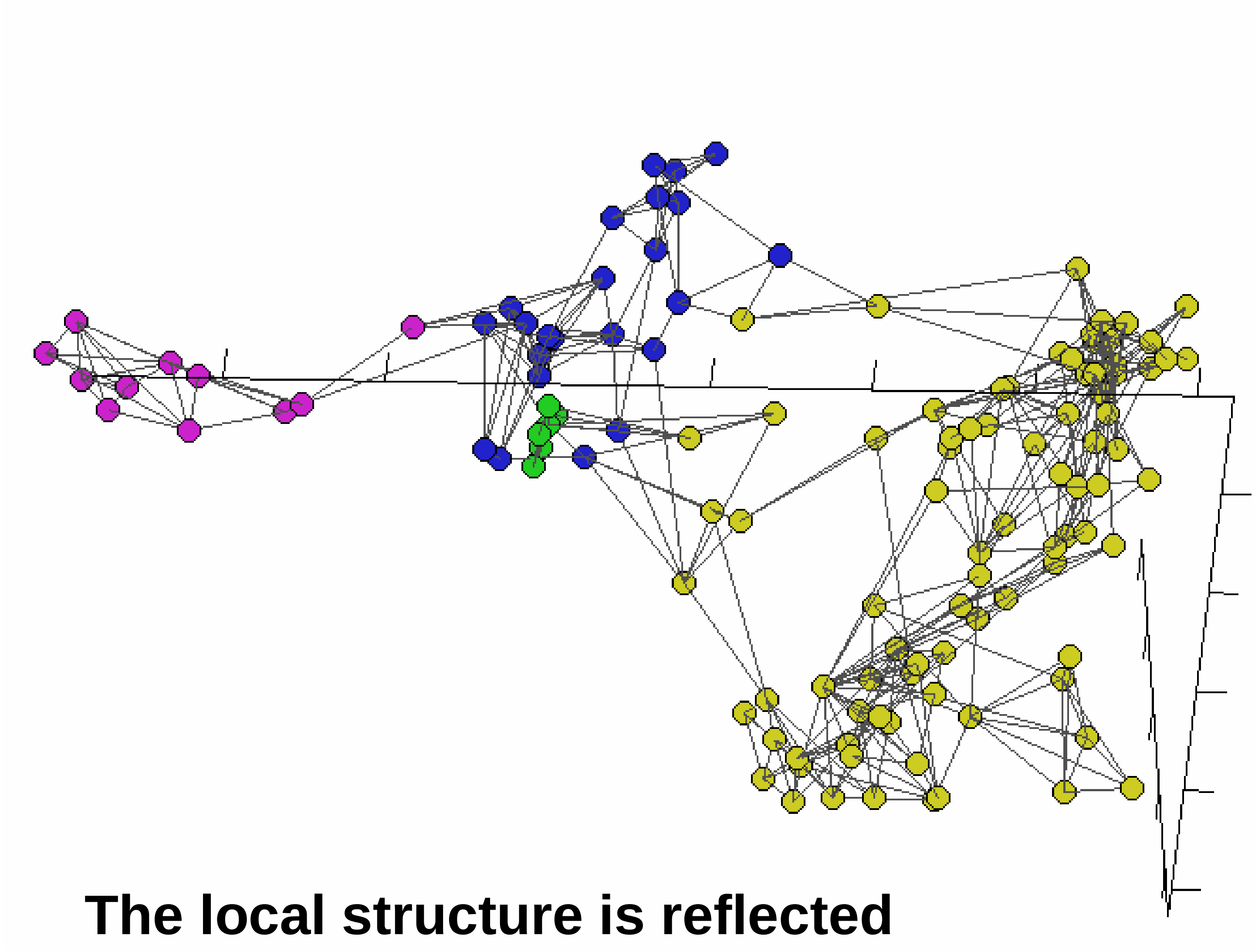
Variable Importance

- Input to KDR
  - K_X : Gaussian kernel on input space
  - K_Y : Manifold kernel
  - d : Dimension of central space
- Output
  - Linear mapping B (n-by-d matrix)
- Row i in B quantifies the influence of variable i
- Variable importance: $\max_{j=1,\ldots,d} |B_{ij}|$

Isomap on Leukemia microarray data



Central subspace w.r.t Isomap kernel



The local structure is reflected

Top influential genes

1. CDNA FLJ39389 fis, clone PLACE6003621
2. B-cell linker
3. Neuritin 1
4. Connective tissue growth factor
5. Aldehyde dehydrogenase 1 family, member A2
6. CD9 molecule
7. RAB32, member RAS oncogene family
8. SH2 domain protein 1A, Duncan’s disease (lymphoproliferative syndrome)
9. Interleukin 8
10. Like-glycosyltransferase
11. N/A
12. Palladin, cytoskeletal associated protein
13. Suppressor of cytokine signaling 2|Transcribed locus
14. Mal, T-cell differentiation protein
15. Integral membrane protein 2A

Many involved in hematopoiesis – formation of cellular blood components

Top influential pathways

Rank known functional groups of genes

- **Mann-Whitney-Wilcoxon test**

1. TCRA PATHWAY 0.0000 14
2. CTLA4 PATHWAY 0.0005 12
3. PROSTAGLANDIN AND LEUKOTRIENE METABOLISM 0.0026 18
4. PEPI PATHWAY 0.0031 5
5. CTL PATHWAY 0.0031 7
6. TCR MOLECULE 0.0048 3
7. EOSINOPHILS PATHWAY 0.0069 4
8. PTDINS PATHWAY 0.0073 10
9. ST GA12 PATHWAY 0.0076 8
10. LYMPHOCYTE PATHWAY 0.0076 10
11. LAIR PATHWAY 0.0084 9
12. PROSTAGLANDIN SYNTHESIS REGULATION 0.0085 12
13. TCR PATHWAY 0.0089 25
14. CSK PATHWAY 0.0101 17
15. AMI PATHWAY 0.0101 17

Many associated to B-cell and T-cell signalling

Conclusions

- Taking nonlinearities into account is important and useful in exploratory analysis of gene expression data
- Dimensionality reduction is not the only application of manifold learning
  - Clustering
  - Classification
  - Regression
  - …

Acknowledgements

Lund University

- **Magnus Fontes**

Lund Univ. Hospital

- **Thoas Fioretos**
- **Mattias Höglund**
- **Anna Andersson**

UC Berkeley

- **Michael I. Jordan**

Yahoo! Research / UC Berkeley

- **Fei Sha**

AstraZeneca

- **Per Broberg**
  (now at Ferring)

Thank You

www.maths.lth.se/matematiklth/personal/jensn/

References

- Alizadeh, A.; Eisen, M.; Davis, E.; Ma, C.; Lossos, I.; A., R.; Boldrick, J.; Sabet, H.; Tran, T.; Yu, X.; Powell, J.; Yang, L.; Marti, G.; Moore, T.; Hudson, J.; Lu, L.; Lewis, D.; Tibshirani, R.; Sherlock, G.; Chan, W.; Greiner, T.; Weisenburger, D.; Armitage, J.; Warnke, R.; Levy, R.; Wilson, W.; Grever, M.; Byrd, J.; Botstein, D.; Brown, P. & Staudt, L.
  Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling
  Nature, 2000*, 403*, 503-511

- Andersson, A.; Olofsson, T.; Lindgren, D.; Nilsson, B.; Ritz, C.; Eden, P.; Lassen, C.; Råde, J.; Fontes, M.; Morse, H.; Heldrup, J.; Behrendtz, M.; Mitelman, F.; Höglund, M.; Johansson, B. & Fioretos, T.
  Molecular signatures in childhood acute leukemia and their correlations to expression patterns in normal hematopoietic subpopulations
  Proc. Natl. Acad. Sci. USA, 2005*, 102*, 19069-19074

- Belkin, M. & Niyogi, P.
  Laplacian Eigenmaps for Dimensionality Reduction and Data Representation
  Neural Computation, 2003*, 15*, 1373-1396

- Donoho, D. L. & Grimes, C.
  Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data
  Proceedings of the National Academy of Sciences USA, 2003*, 100*, 5591-5596

- Fukumizu, K.; Bach, F. R. & Jordan, M. I.
  Kernel Dimension Reduction in Regression
  Department of Statistics, University of California, Berkeley, 2006

- Fukumizu, K.; Bach, F. R. & Jordan, M. I.
  Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces
  Journal of Machine Learning Research, 2004, *5*, 73-99

- Li, K.
  Sliced Inverse Regresion for Dimension Reduction
  Journal of the American Statistical Association, 1991, *86*, 316-327

- Nilsson, J.; Fioretos, T.; Höglund, M. & Fontes, M.
  Approximate geodesic distances reveal biologically relevant structures in microarray data
  Bioinformatics, 2004, *20*, 874-880

- Nilsson, J.; Sha, F.; Andersson, A.; Fioretos, T.; Fontes, M. & Jordan, M.I.
  Variable importance assessment in manifold learning, in
  Nilsson J. ”Manifold Learning in Computational Biology”, PhD thesis, Lund University 2008

- Roweis, S. & Saul, L.
  Nonlinear dimensionality reduction by locally linear embedding
  Science, 2000, *290*, 2323-2326

- Weinberger, K.; Sha, F. & Saul, L.
  Learning a kernel matrix for nonlinear dimensionality reduction
  2004

LUNDS UNIVERSITET

Lunds Tekniska Högskola