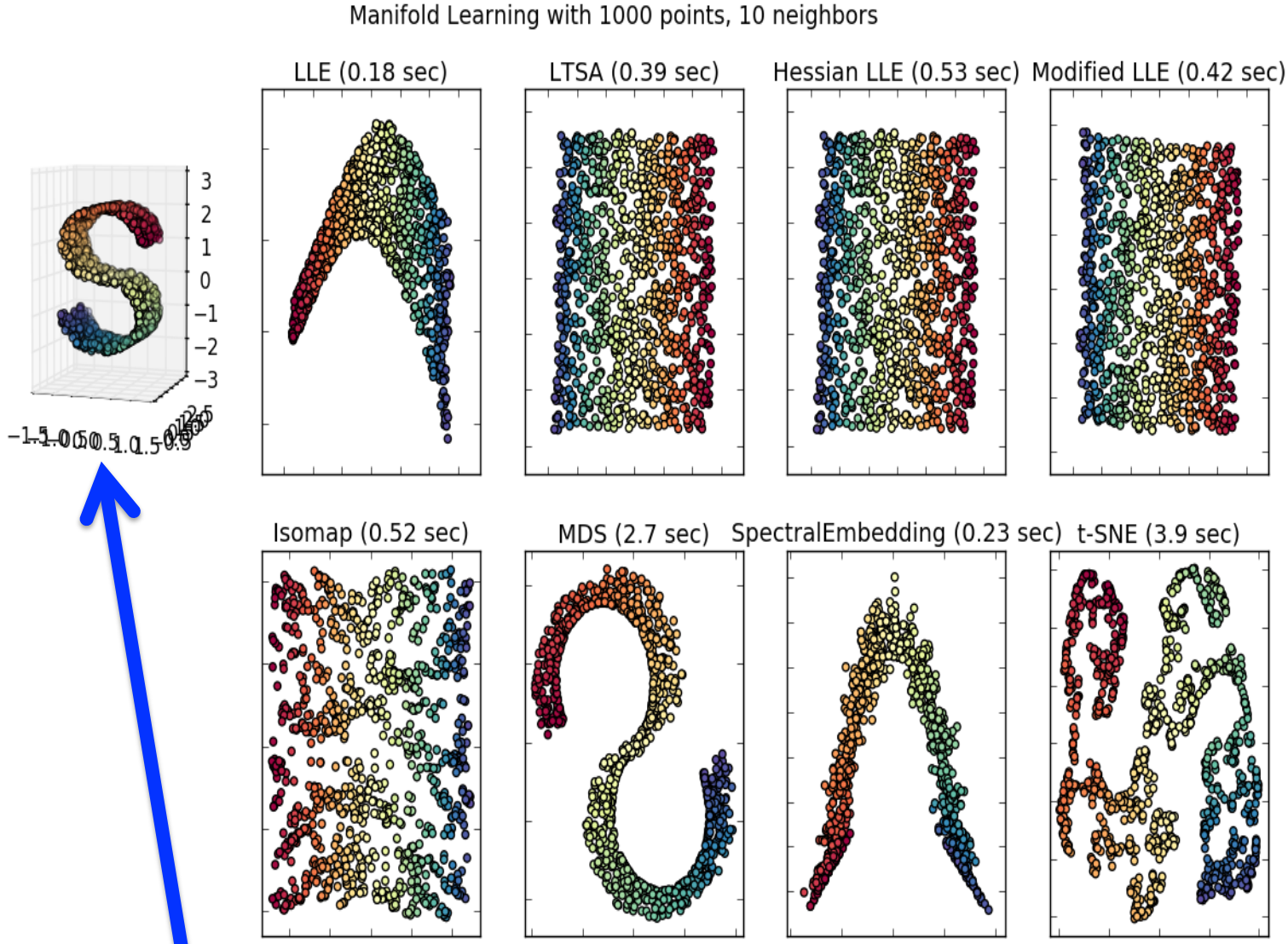
MANCHESTER 1824
The University of Manchester

Visualising high dimensional data using non-linear methods

Alexis Boukouvalas and Magnus Rattray
Faculty of Biology, Medicine and Health
University of Manchester

- Problem with linear PCA
- Other methods in python
- Stochastic Neighborhood embedding
- Gaussian processes
  - Regression
  - Gaussian process latent variable model (GPLVM)



http://scikit-learn.org

Linearity can be quite restrictive – in 2-D it would be a mess here!

BEYOND PCA IN PYTHON

- Isomap: maintains geodesic distances between all points. Builds nearest neighbor graph in data space, computes shortest path graph.
- Locally linear embedding (LLE): preserves distances within local neighborhoods.
- Spectral Embedding aka Laplacian Eigenmaps: estimation of graph in data space, minimization of a cost function so points close to each other on the manifold are mapped close to each other in the low dimensional space, preserving local distances.
- Multidimensional scaling (MDS): that distances in the original high-dimensional space are respected by minimizing cost: $sum_{i<j} d_{ij}(X) - \hat{d}_{ij}(X)$
- Independent component analysis (ICA): Linear projection to decompose signal to independent non-Gaussian sources.

http://scikit-learn.org

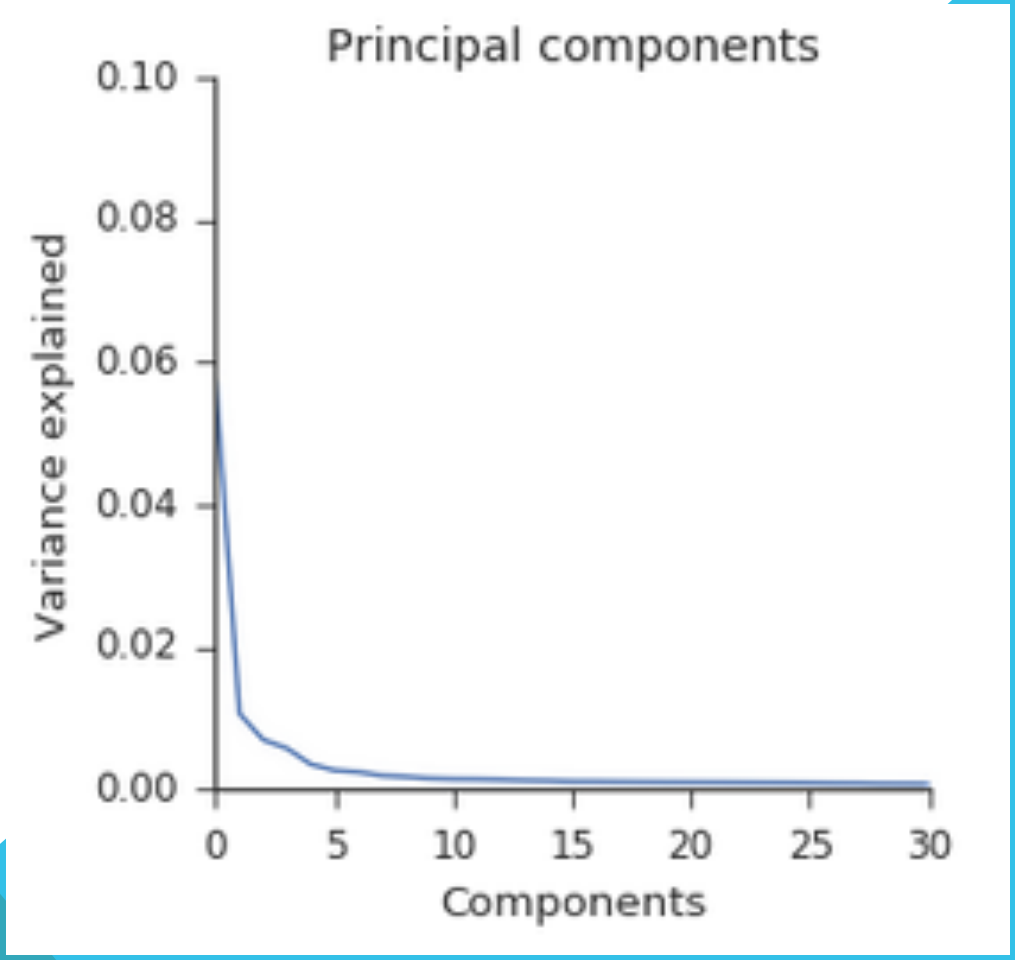
T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING (T-SNE)

- t-distributed Stochastic Neighbor Embedding (t-SNE): minimize the difference between Gaussian joint probabilities in data space and the Student's t-distributions in latent-space. Very populare in single-cell.
- Demo: http://distill.pub/2016/misread-tsne/
- One parameter: perplexity to balance preservation of local and global aspects. Low value (e.g. 2) local variation dominates and vice versa for larger values. Should be smaller than number of points.
- Good at picking out clusters of data that lie in separate manifolds -> well separated clusters in latent space.
- Standard implementation limited to 2-D and 3-D latent spaces.
- Multiple restarts needed to avoid local minima using objective to select best one.
- Often run on PCA latent space to preserve global structure (see later example).

Single cell RNA-sequencing example

- Differentiation of myeloid and erythroid precursors from hematopoietic stem cells in the mouse bone marrow.
- RNA-seq data with 4423 cells x 2312 genes

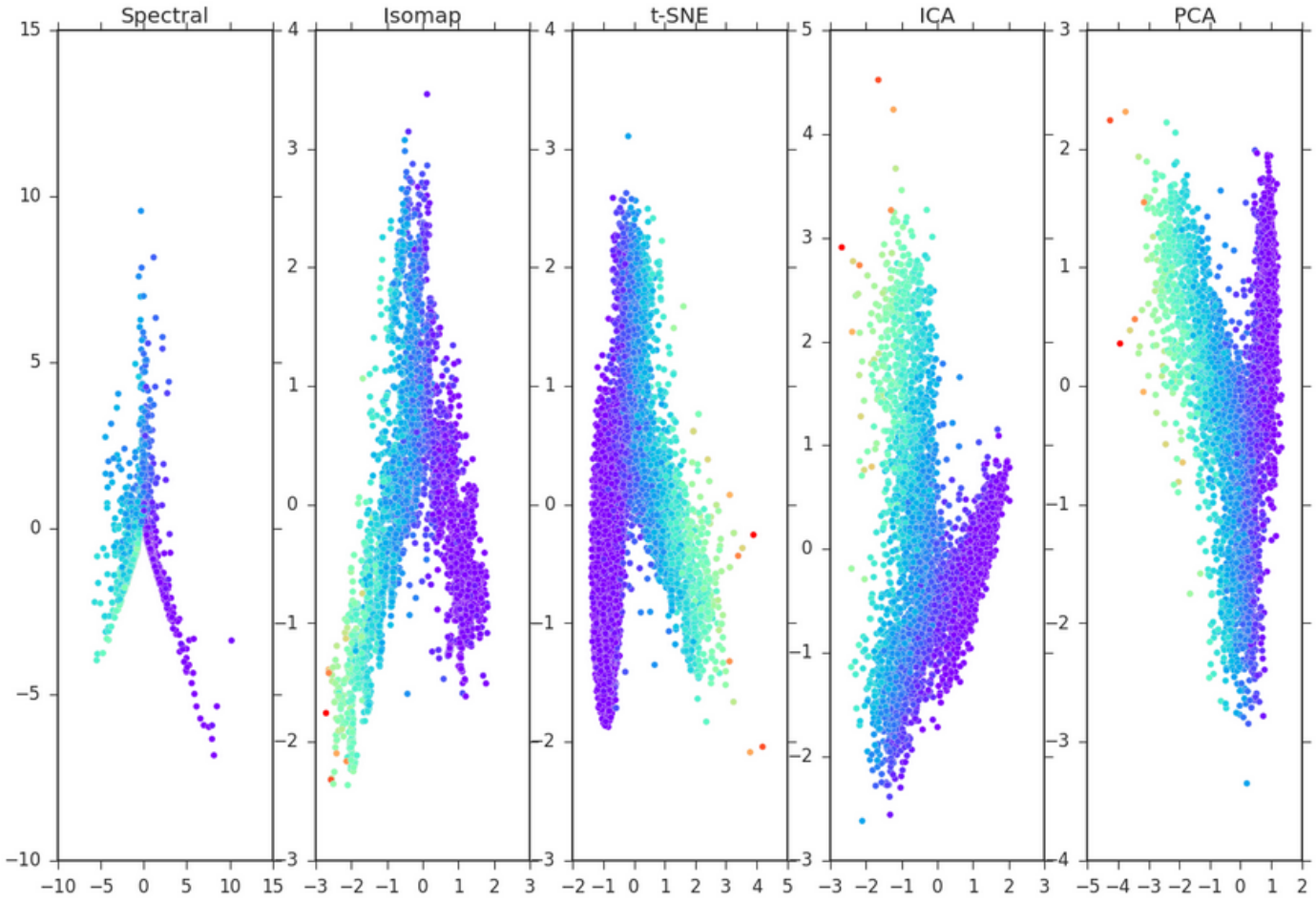
Principal Components Analysis: How many components to use?



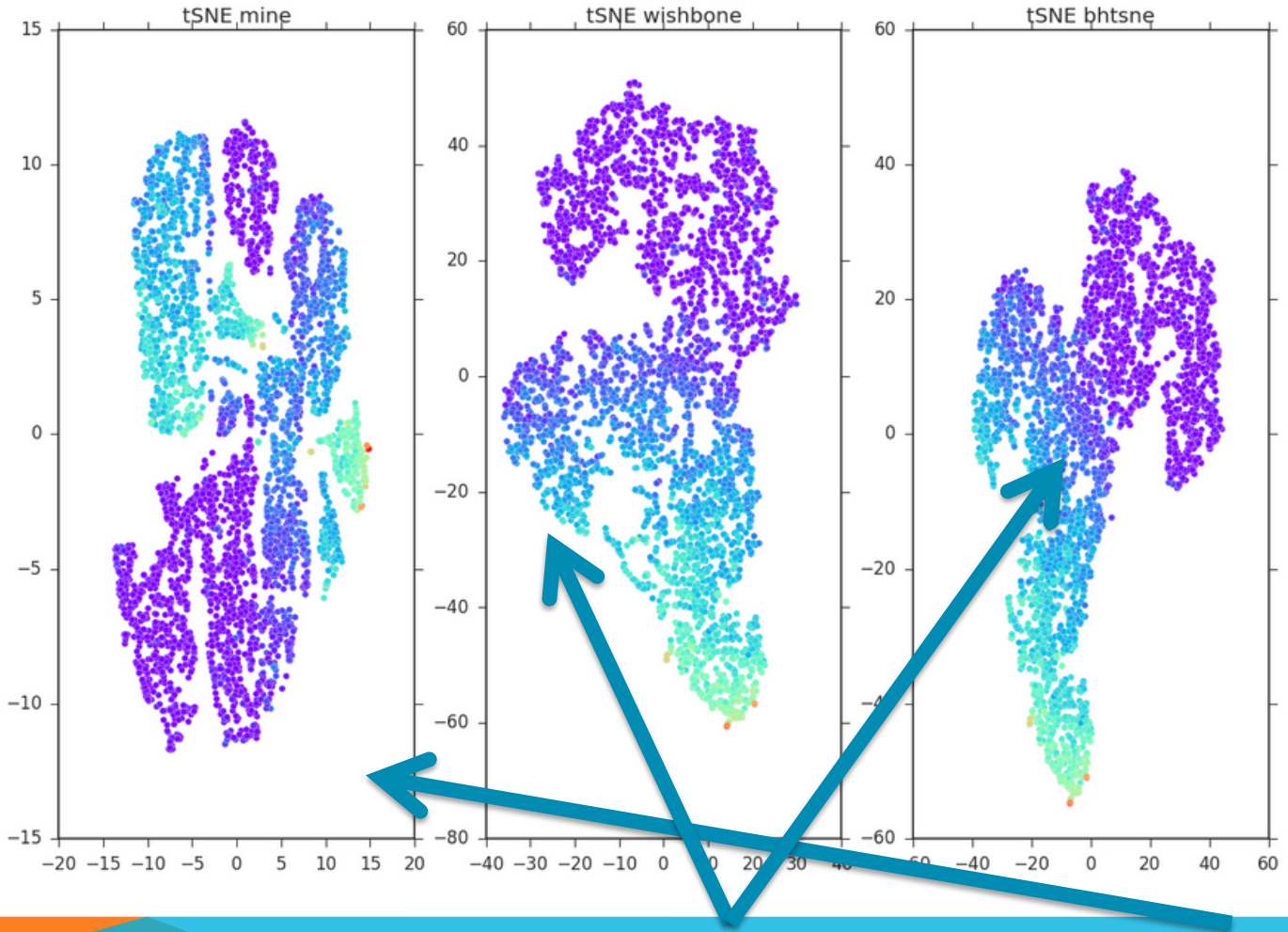
https://github.com/ManuSetty/wishbone

Looking at the elbow? 5 components

Color by myeloid gene MPO:

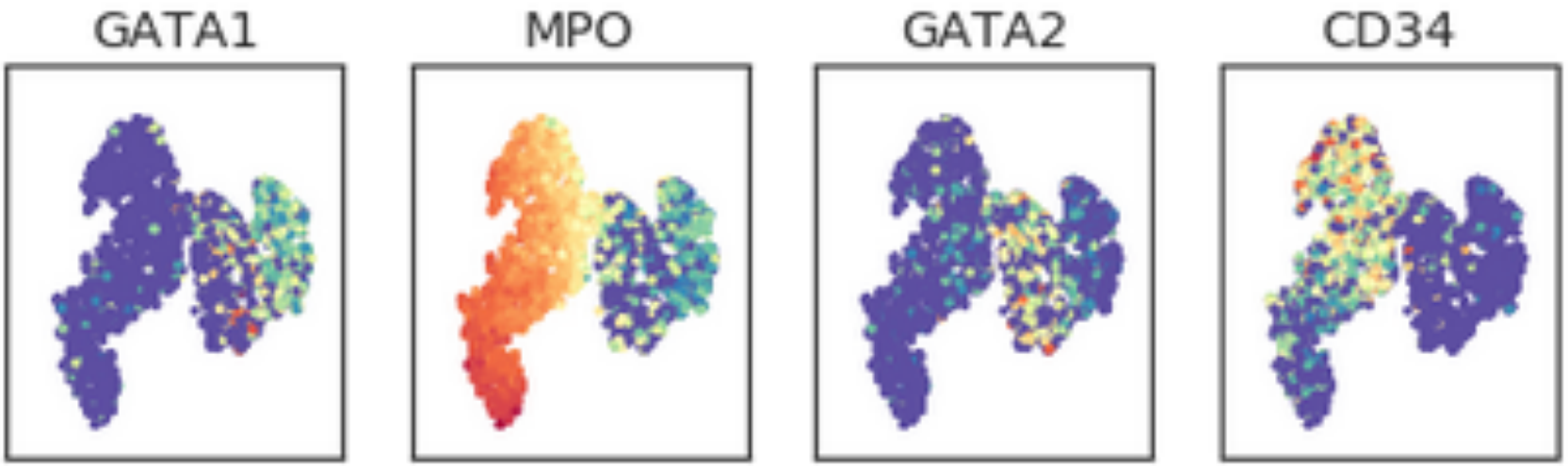


Stacking PCA-> tSNE to reveal global structure but still prone to local minima.



Global structure clearer when PCA->tSNE, not tSNE directly.

Color tSNE latent space by gene expression: HSC gene CD34, myeloid gene MPO and erythroid precursor genes GATA2 and GATA1.

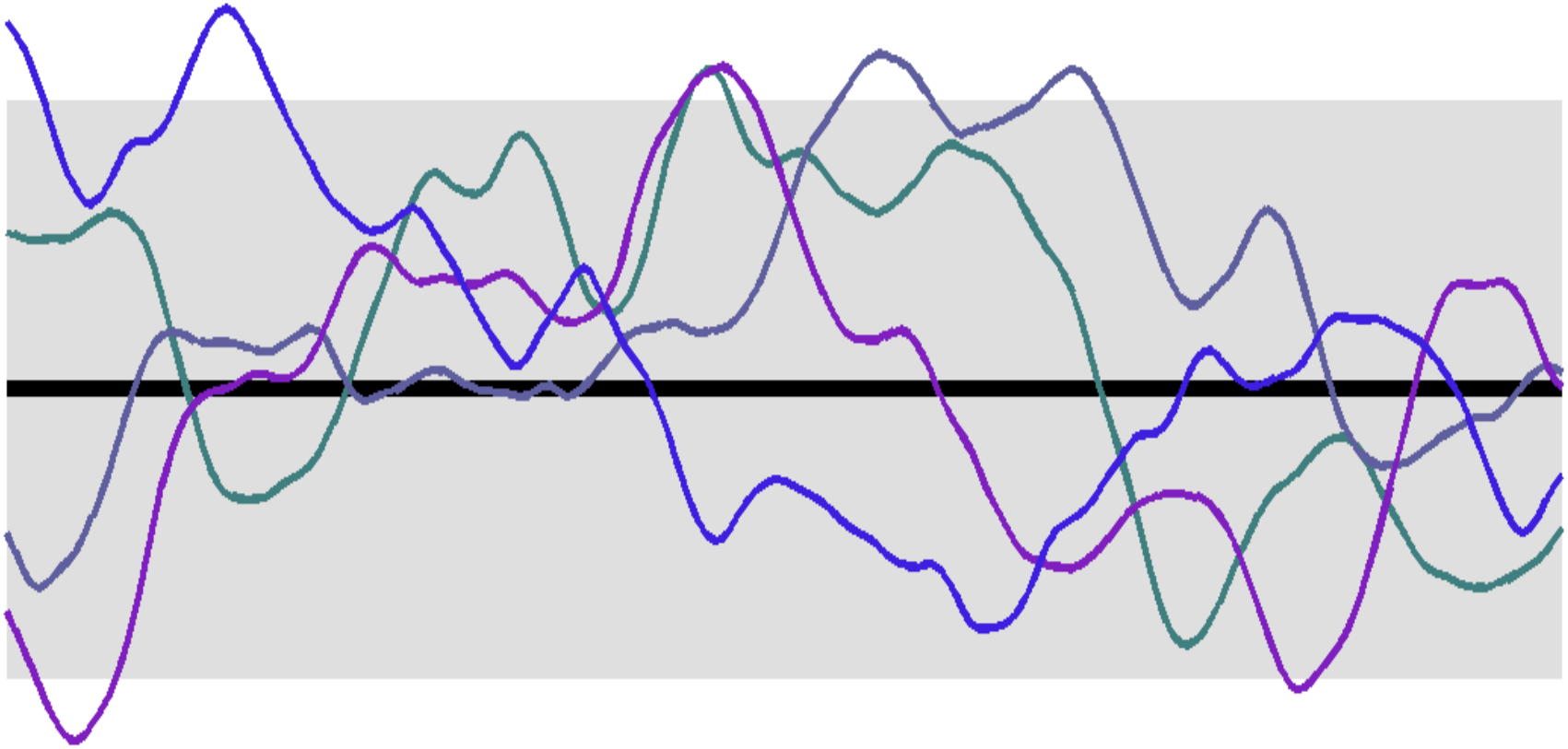


https://github.com/ManuSetty/wishbone

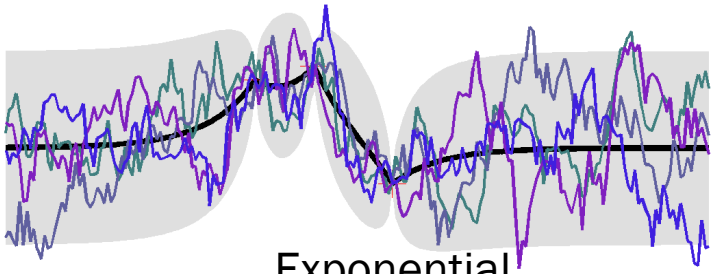
GAUSSIAN PROCESS LATENT VARIABLE (GPLVM)

- A probabilistic non-linear dimension reduction method based on Gaussian processes.
- Gaussian process: A prior over functions.
- Defined by the mean function, m(x), and covariance function k(x, x′).

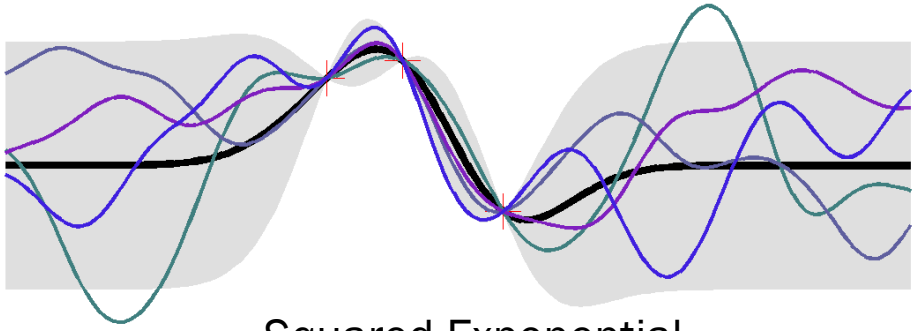
DIFFERENT COVARIANCES=DIFFERENT BELIEFS ON FUNCTION SHAPE: PRIOR



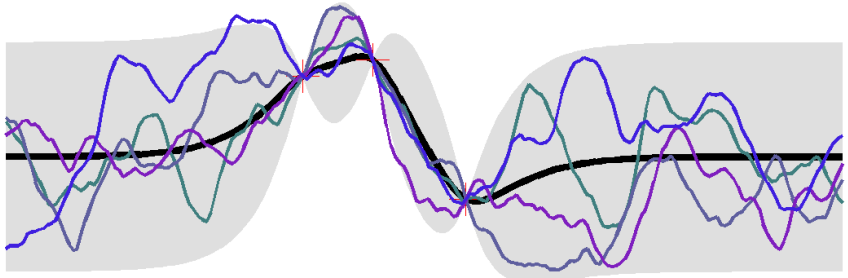
Matern 5/2



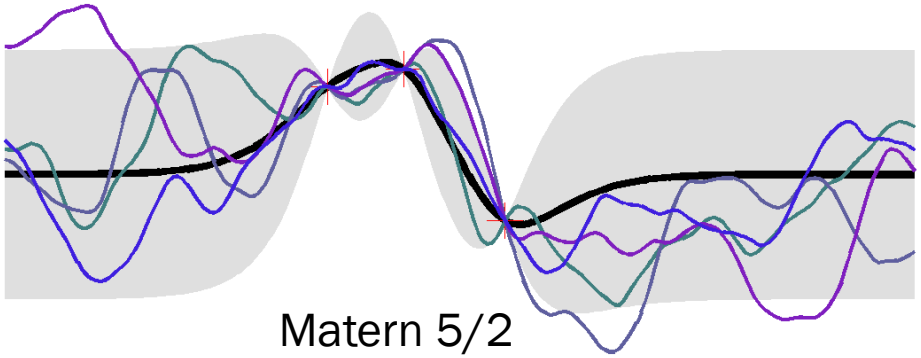
Exponential



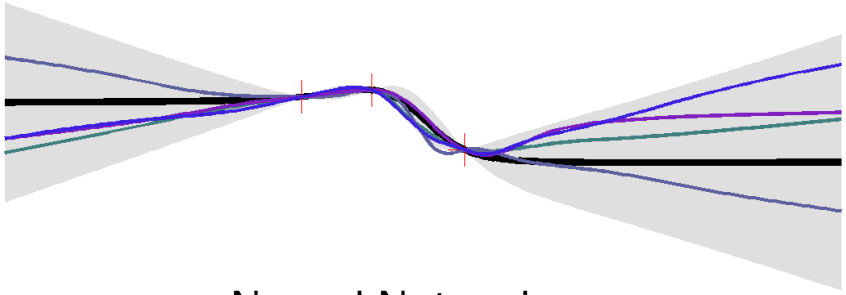
Squared Exponential



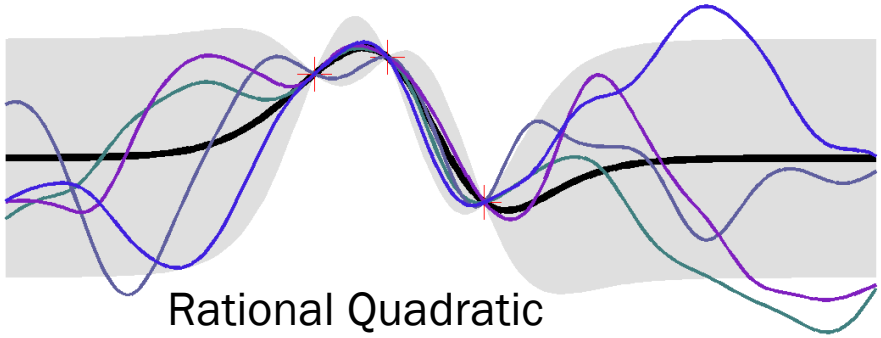
Matern 3/2



Matern 5/2



Neural Network



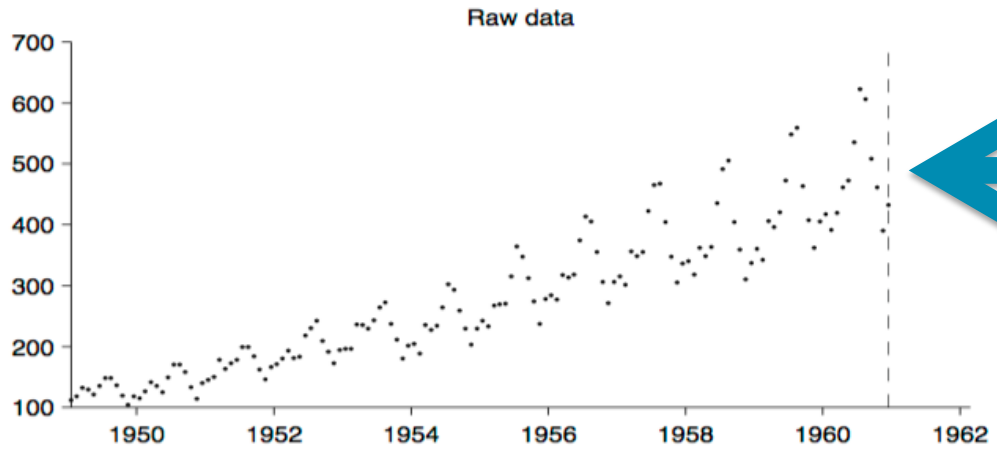
Rational Quadratic

Covariance functions conditioned on 3 points.

TYPES OF COVARIANCE FUNCTIONS

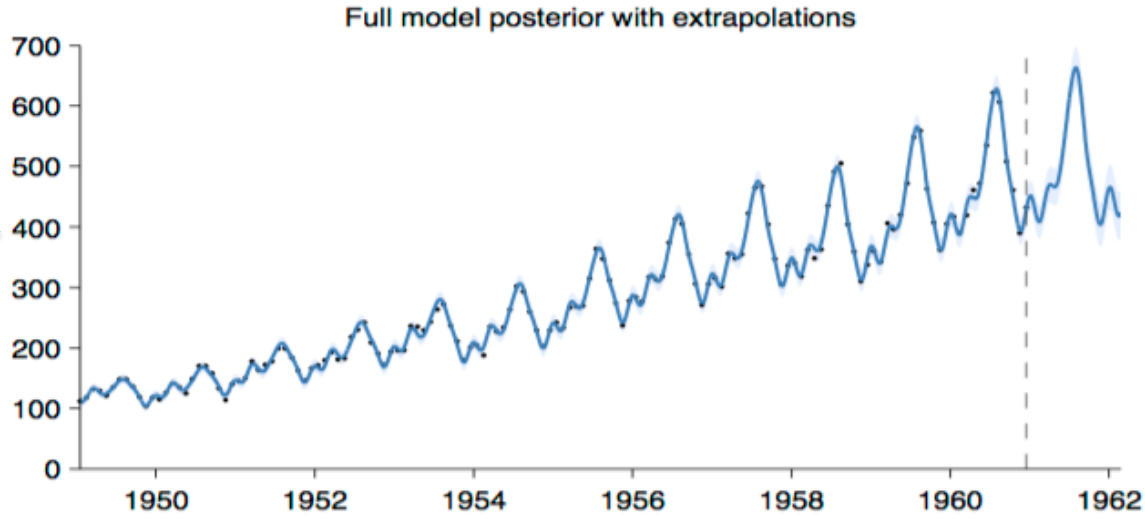
1. Stationary: covariance between two points only depends on their distance. Can be easier to estimate than non-stationary covariances.
2. Not all functions are valid covariance functions: the function must be positive semi-definite to produce valid covariance matrices.
3. We can create new covariance functions since the sum or product of any two covariance functions is also a valid covariance function.
4. A likelihood can be calculated: how well can the model explain the data?
5. A likelihood is a good way to compare models and even select the covariance structure using rules like (3) and evaluating the likelihood -> Automatic statistician

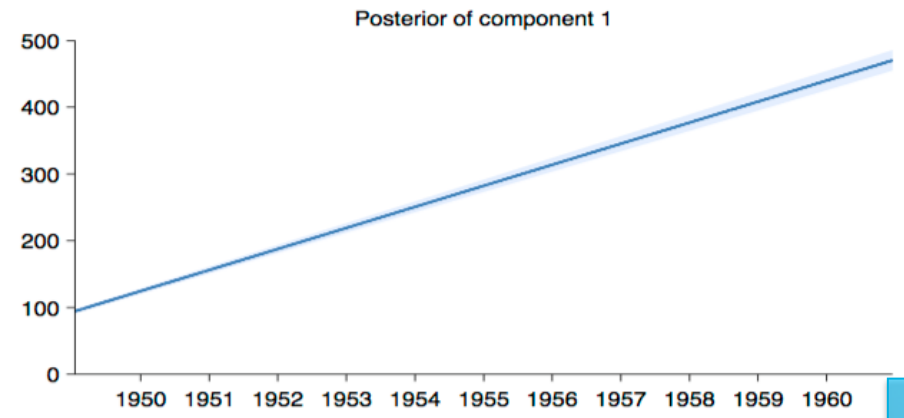
WWW.AUTOMATICSTATISTICIAN.COM



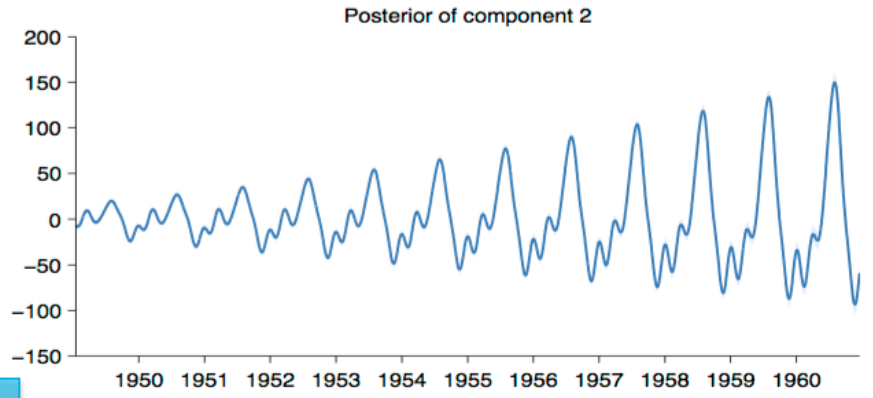
Raw data

Model fit

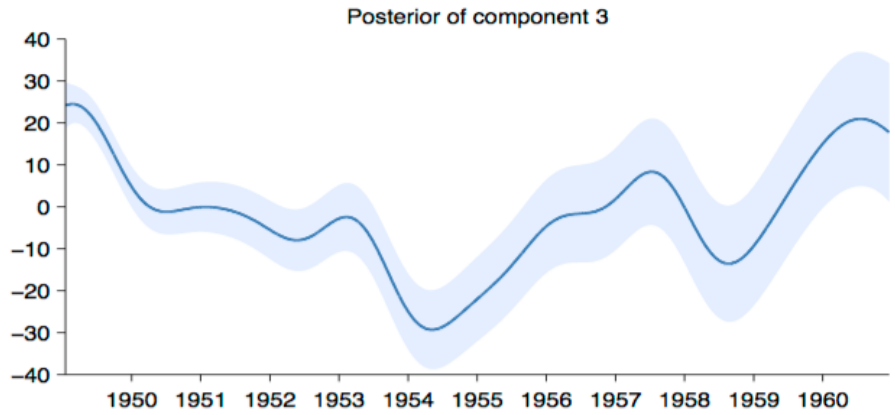




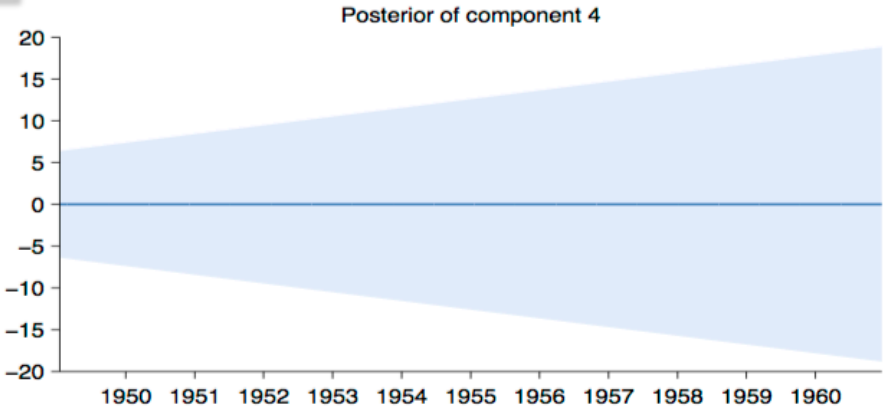
A linearly increasing function



periodic function with
linearly increasing amplitude



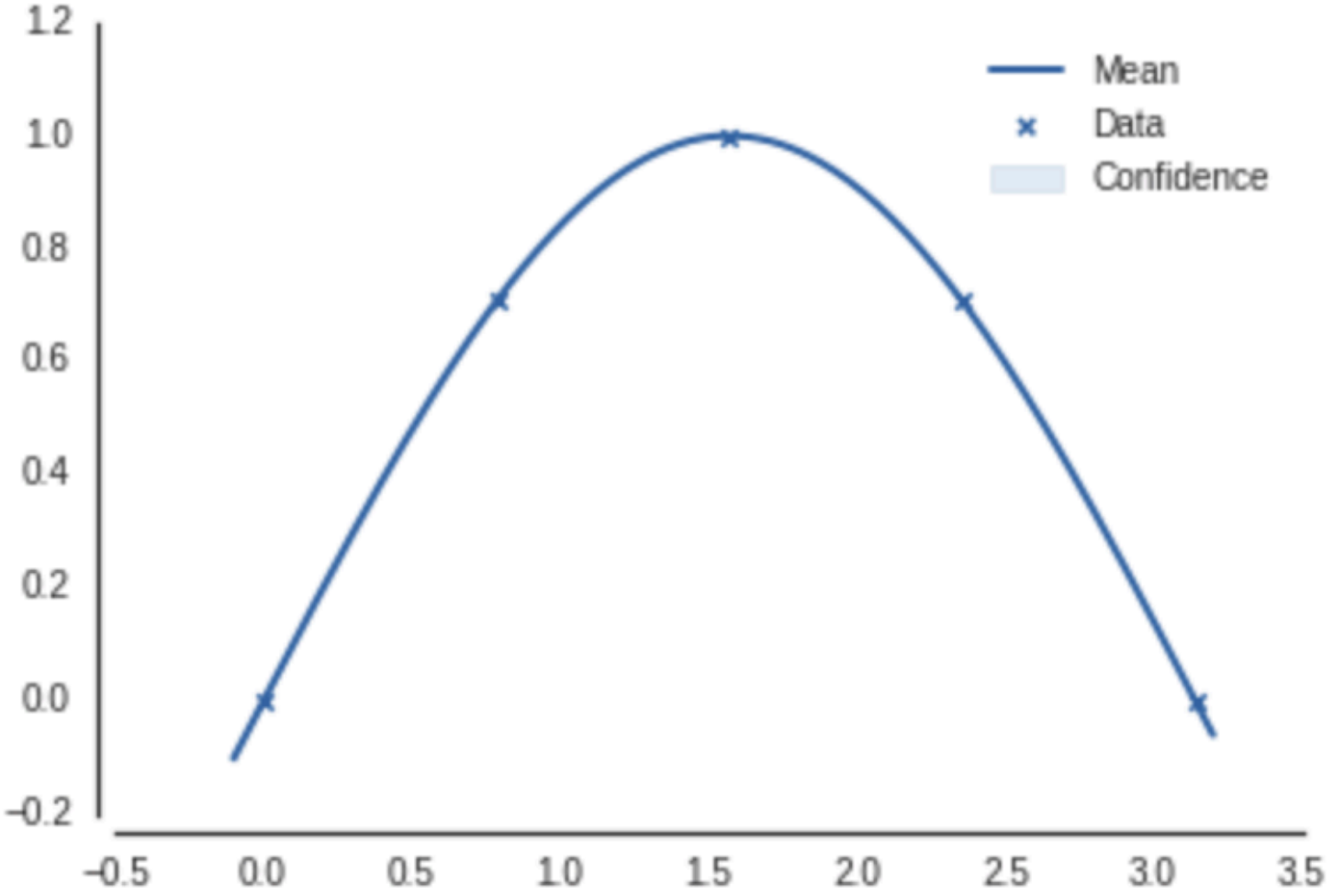
smooth function



Uncorrelated noise with linearly increasing standard deviation

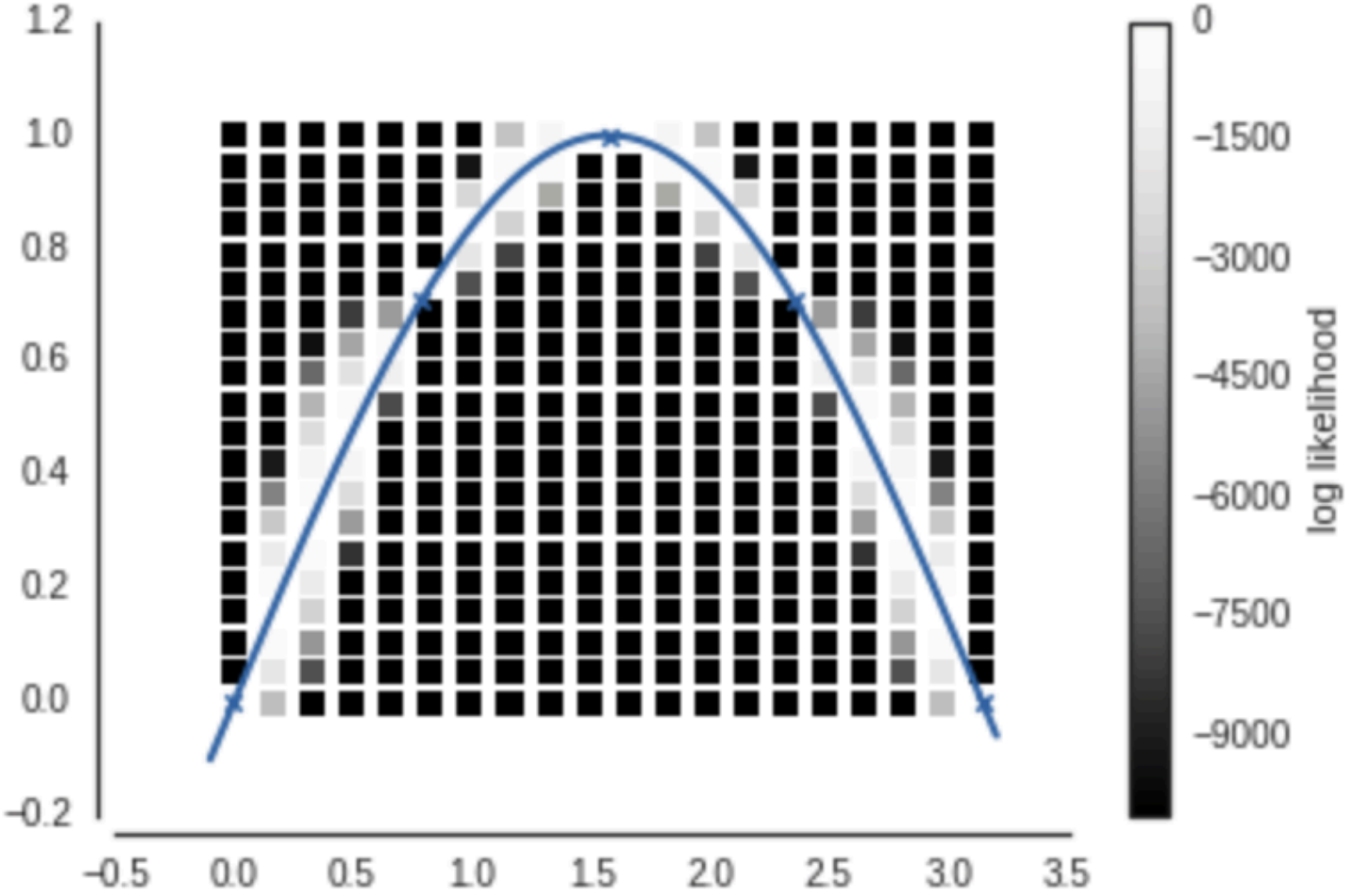
GAUSSIAN PROCESS LATENT VARIABLE MODEL

- A Gaussian process where we only know Y but not X. GP(X) -> Y
- X is not known and has to be estimated – the latent space.
- A GPLVM with a linear covariance function is the same as PCA!
- A non-linear covariance function like the ones we have seen so far -> non-linear dimension reduction.



Given 5 points X, Y we learn the model by Gaussian process regression

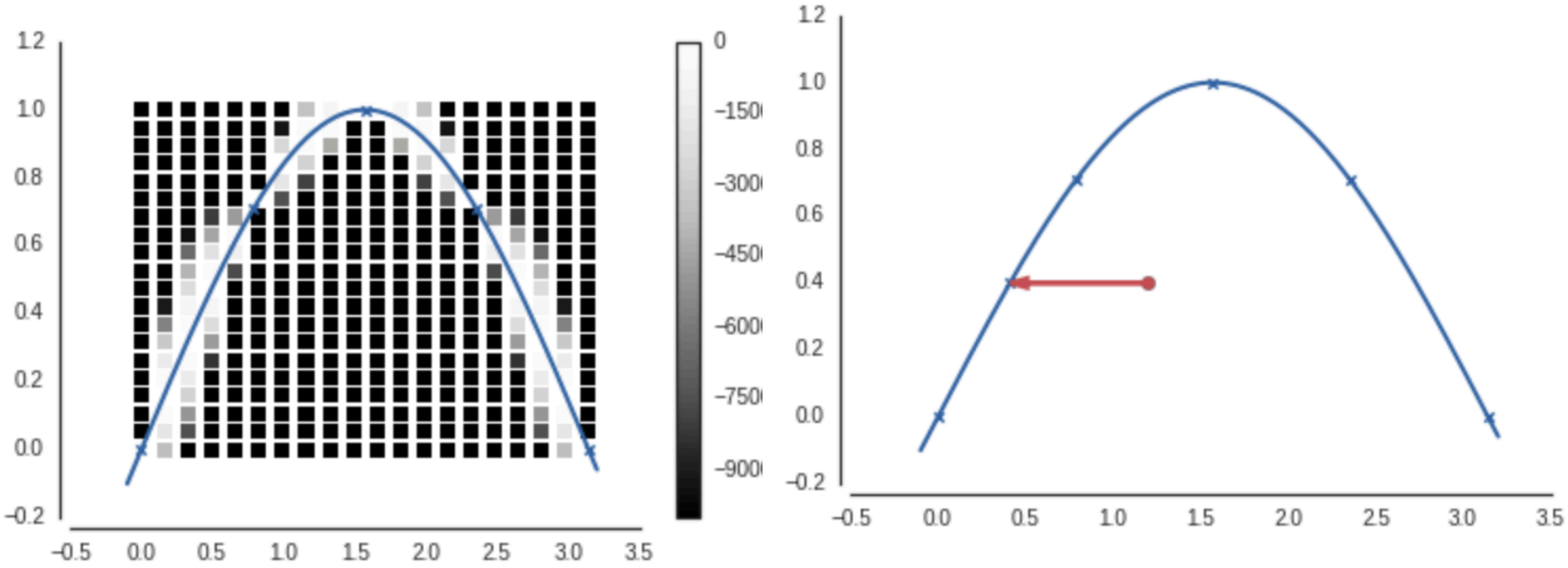
http://www.nxn.se/valent/some-intuition-about-the-gplvm



Where could the 6th (X,Y) point appear?

More likely close to the curve.

http://www.nxn.se/valent/some-intuition-about-the-gplvm



For Y=0.4, where could the X be?

Answer depends on initial point.

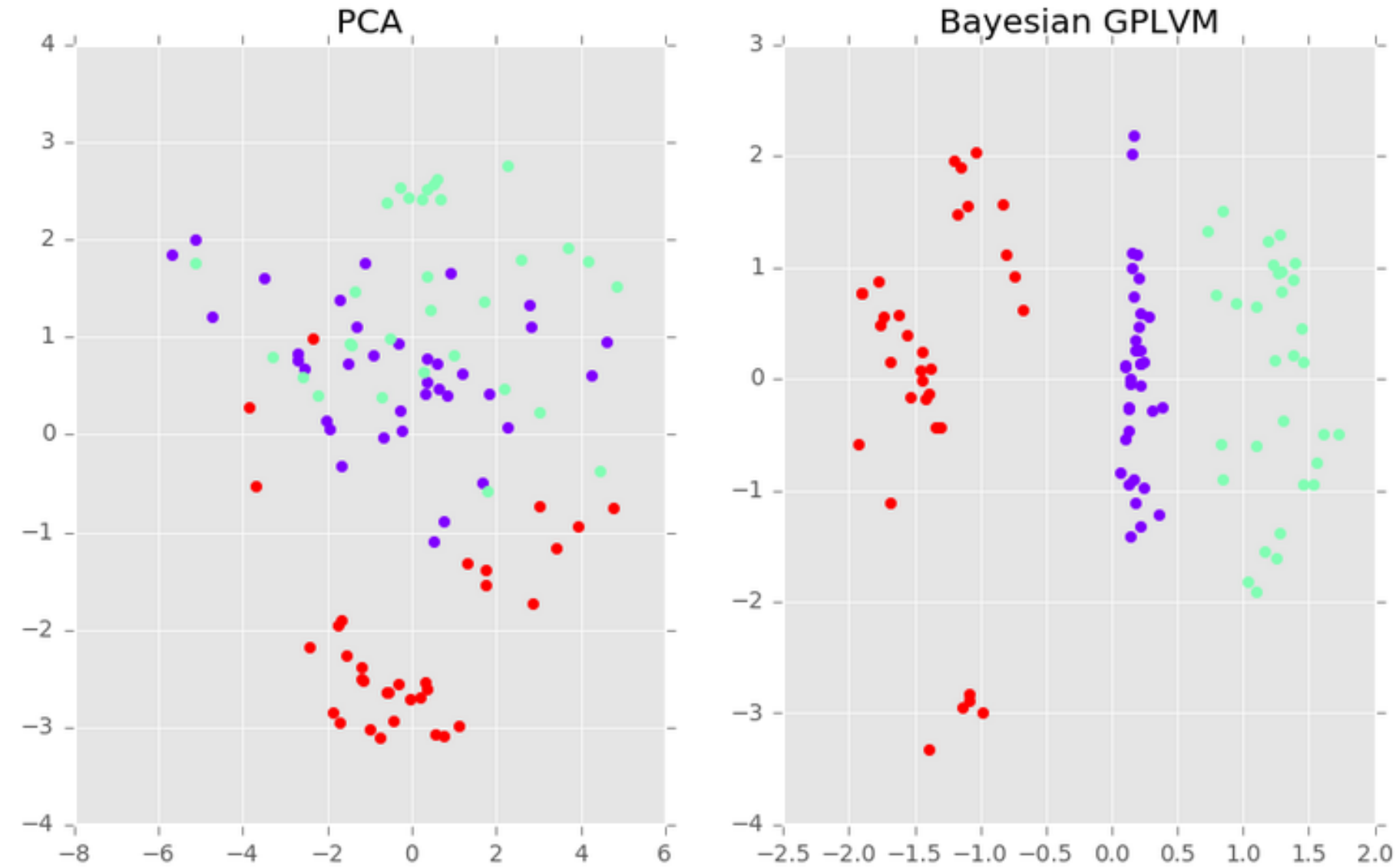
For higher-dimensional Y, easier!

http://www.nxn.se/valent/some-intuition-about-the-gplvm

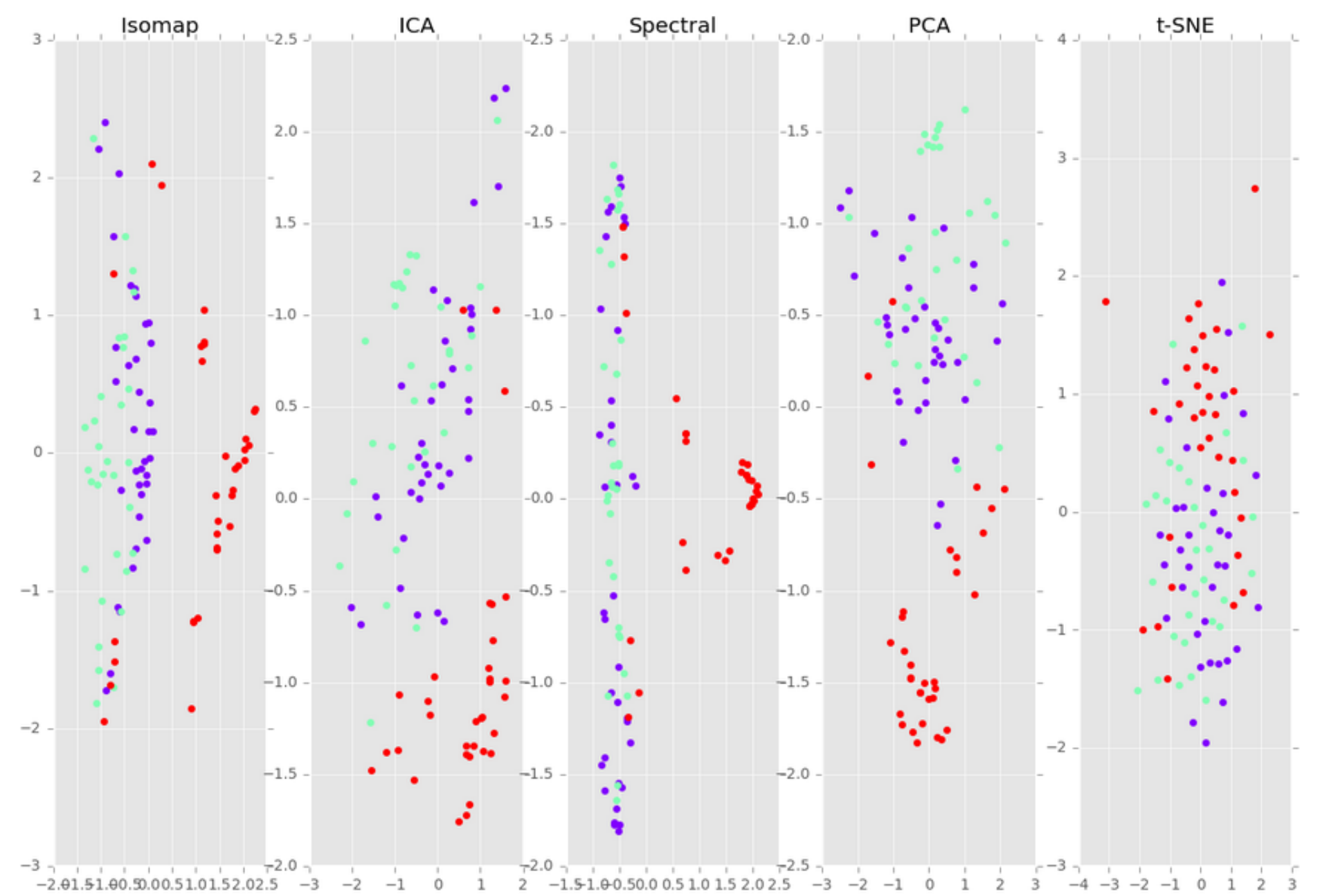
OIL DATASET

- Standard benchmark dataset generated from oil flow data containing three phases.
- Bishop, C. M. and G. D. James (1993). Analysis of multiphase flows using dual-energy gamma densitometry and neural networks. Nuclear Instruments and Methods in Physics Research A327, 580-593

How well can unsupervised methods separate the three phases without seeing the phase labels?


PCA
4
3
2
1
0
−1
−2
−3
−4
−8
−6
−4
−2
0
2
4
6

Bayesian GPLVM
3
2
1
0
−1
−2
−3
−4
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0



Other methods cannot separate the three classes.

GPLVM IN COMPUTATIONAL BIOLOGY

- Pseudotime inference

A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. Buettner F1, Theis FJ., Bioinformatics, 2012.

- Account for confounding factors

Joint Modelling of Confounding Factors and Prominent Genetic Regulators Provides Increased Accuracy in Genetical Genomics Studies Fusi N, Stegle O, Lawrence ND, PLOS Computational Biology, 2012.

- Incorporating prior information

Pseudotime estimation: deconfounding single cell time series. Reid JE, Wernisch L., Bioinformatics, 2016.

- Assessing uncertainty

Order Under Uncertainty: Robust Differential Expression Analysis Using Probabilistic Models for Pseudotime Inference. Campbell KR, Yau C, PLOS Computational Biology, 2016.

THE END: GO TRY IT OUT!

- Non-probabilistic non-linear methods:

http://scikit-learn.org/stable/modules/manifold.html

- GPLVM:

http://www.nxn.se/valent/some-intuition-about-the-gplvm