UNSUPERVISED MACHINE LEARNING FOR IMPROVED UNDERSTANDING, OPTIMIZATION, AND PROCESS MONITORING, FAULT DIAGNOSIS IN EARLY DEVELOPMENT

NELSON LEE AFANADOR, PH.D., RICHARD BAUMGARTNER, PH.D., DAI FENG, PH.D., SETH CLARK, PH.D.

MERCK

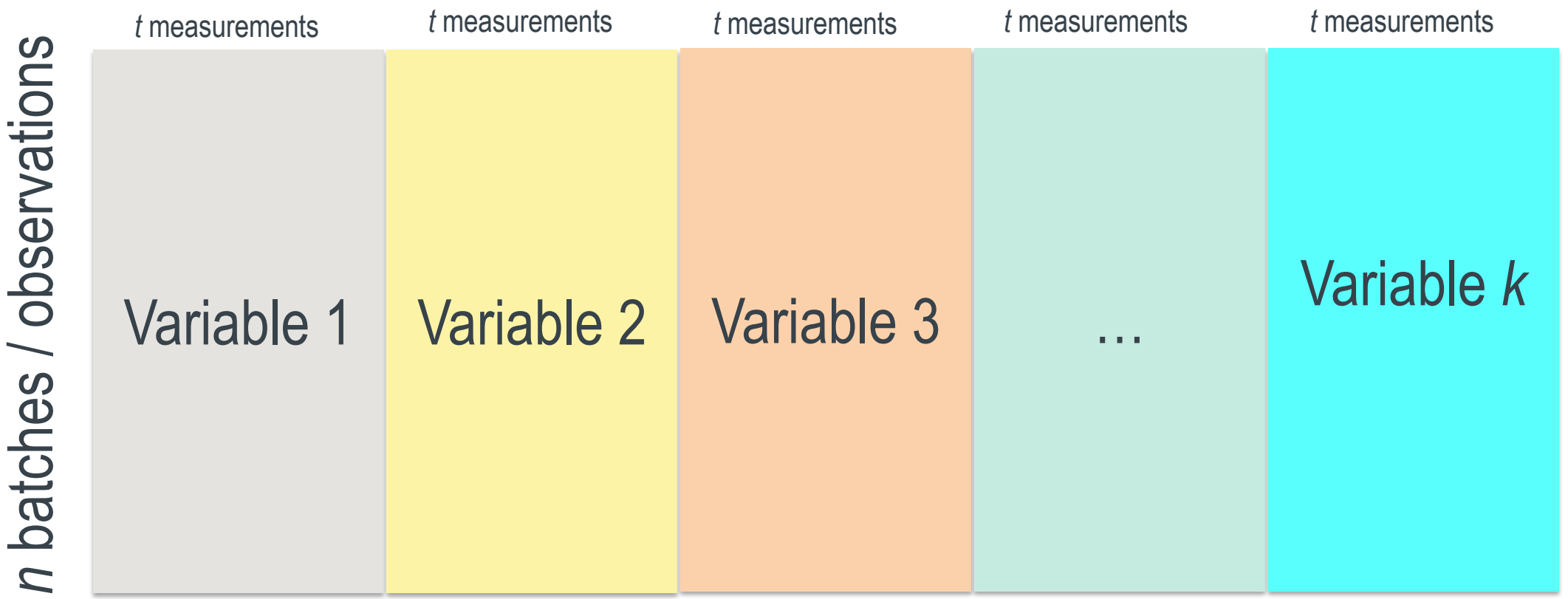
Presentation Date: January 30, 2019

Outline

- Background
- Euclid (L2-norm), Manhattan (L1-norm)
- Unsupervised Random Forest (URF)
- Shallow dive into PCA and URF
- Scale-down model comparison with PCA and URF
- Non-linearly separable clusters – PCA, Kernel-PCA, URF
- Final thoughts

MERCK

‘Multi-Block’ Data Matrix



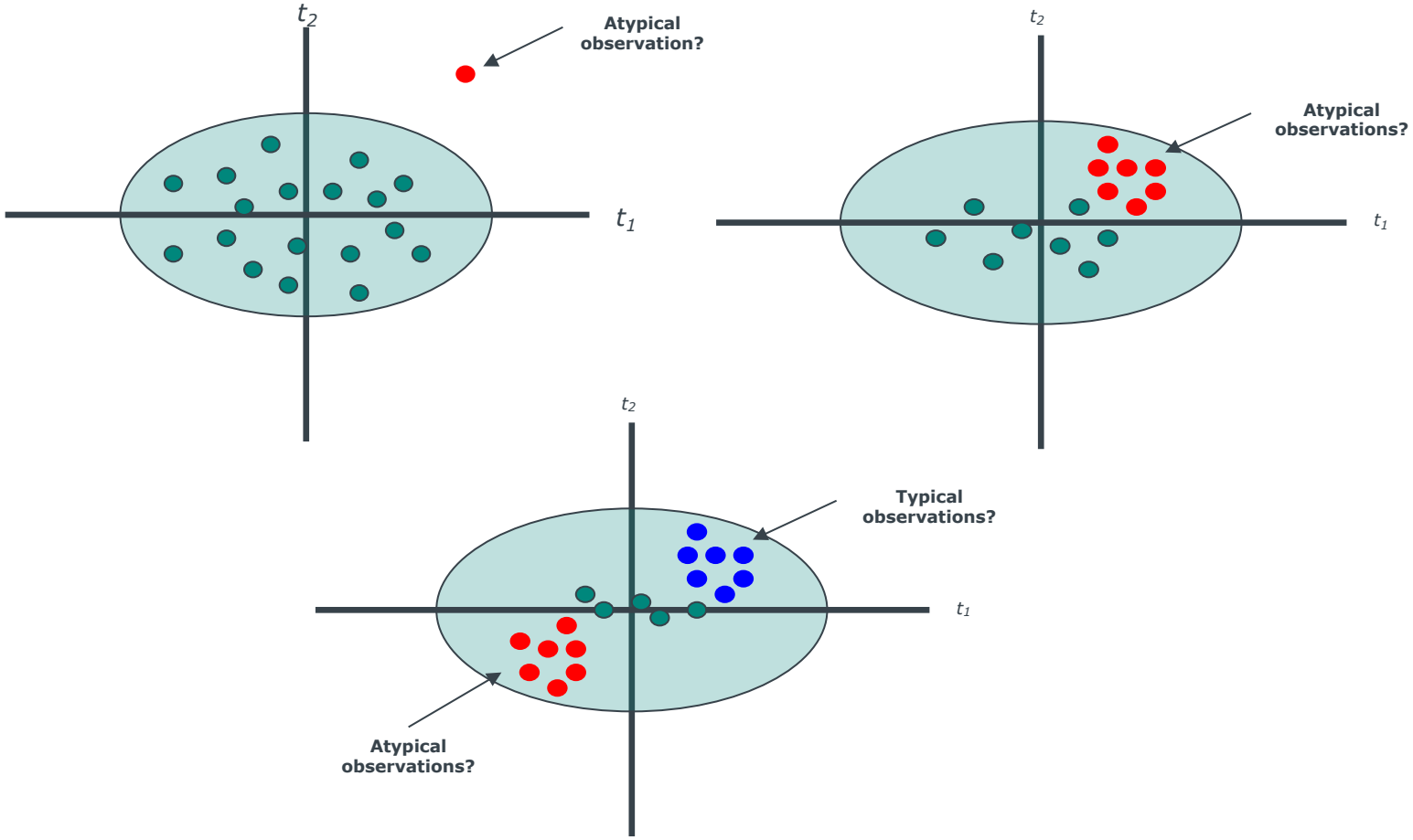
MERCK

Overarching Goal

Dimensionality reduction via latent variable methods, still rule the day.

Both distance and kernel matrices are ways we can to understand the distribution of batches. 'Latent spaces' help us make this effort possible.

But how do we get here? Turns out there's multiple ways…



MERCK

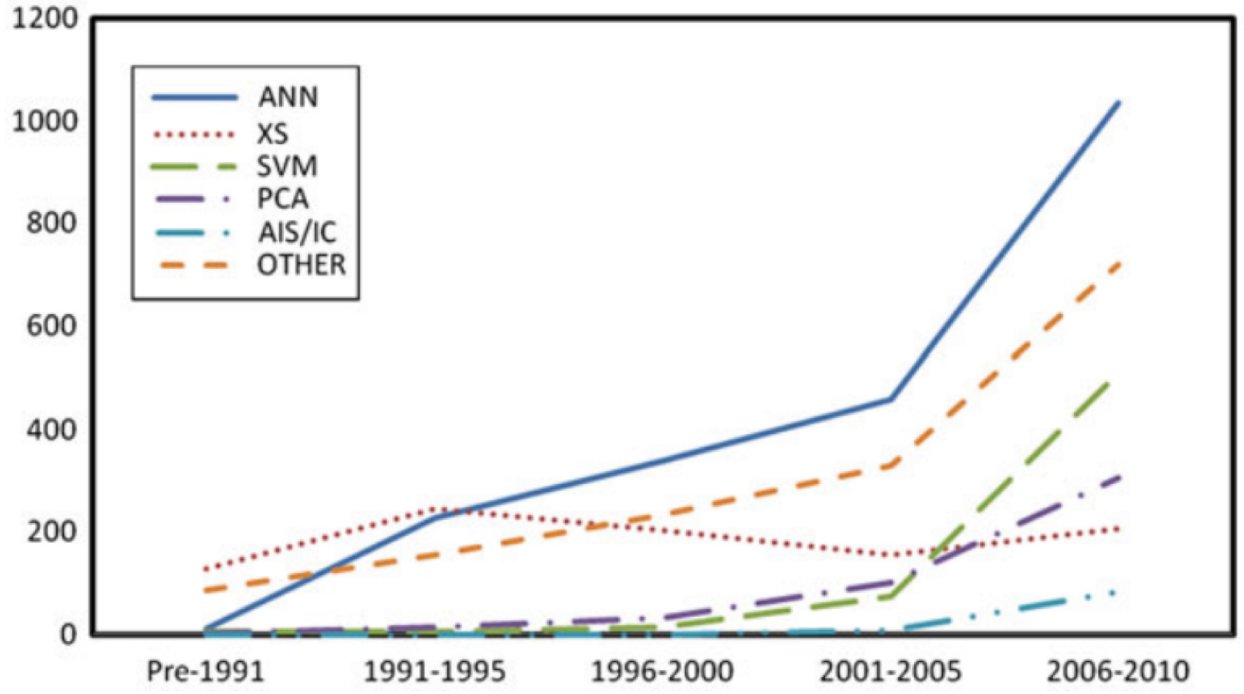
The nature of our multiple objectives

- Process/Analytical understanding
- Improvement/optimization
- Root causes for changes
- Process Monitoring

These objectives have been mode more interesting (or complicated) due to automation and the exponential increase in sampling frequencies.

MERCK

Trends in academic research related to fault diagnosis



"Trends in academic research related to fault diagnosis based on number of publications in the IEEE Xplore digital library from 1991 to 2010" - Aldrich, C., Auret, L., *Unsupervised Process Monitoring and Fault Diagnosis with Machine Learning Methods*, Springer-Verlag London 2013

MERCK

‘Unsupervised’ Machine Learning

- Given the explosion of available process data, due to automation, it may work better in the higher-dimensional spaces we will be forced to work in
- It encompasses a wide range of proven methods
- Flexible modeling options via parameter fine-tuning
- Cross-validation (via reconstruction error) is how one chooses an optimal method for implementation
- Can be computationally expensive
- ***Wherever traditional MVDA is being used, unsupervised machine learning methods can be plugged-in***

MERCK

Common Projection/Dimensionality Reduction Methods

- Principal Component Analysis (PCA) → current most popular Euclidian method
- Unsupervised Random Forest
- Kernel PCA
- Independent Component Analysis
- Locally linear embedding
- t-SNE
- Many others…

MERCK

Common Projection/Dimensionality Reduction Methods

- Principal Component Analysis (PCA) → current most popular Euclidian method
- Unsupervised Random Forest
- Kernel PCA

MERCK

What is a latent variable in everyday life?

Happiness

- How do you measure happiness?
  - Again, it is a combination of factors
    - Work-life balance
    - Health (a latent variable itself)
    - Vacation days

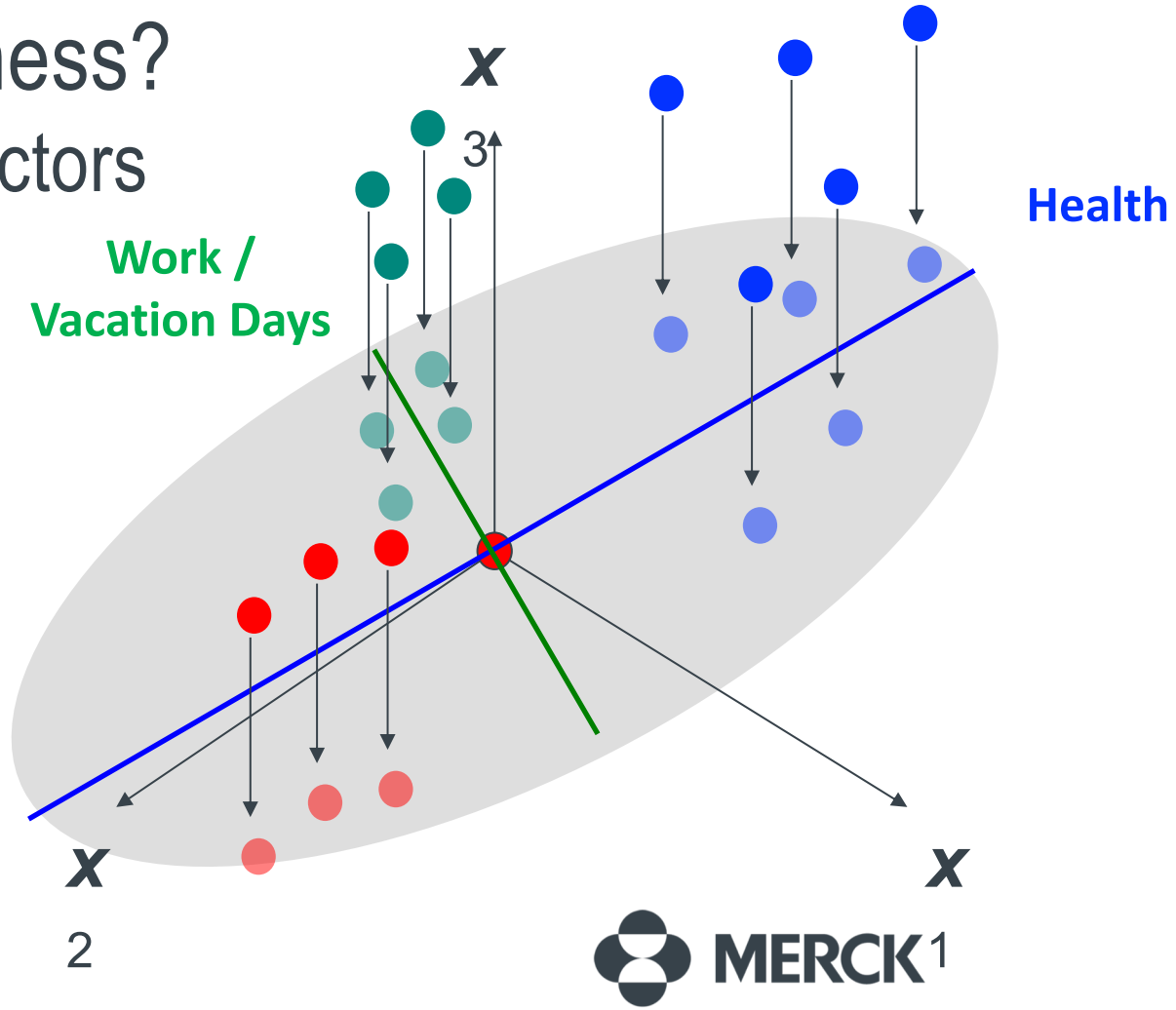
MERCK

What is a latent variable in everyday life?

Happiness

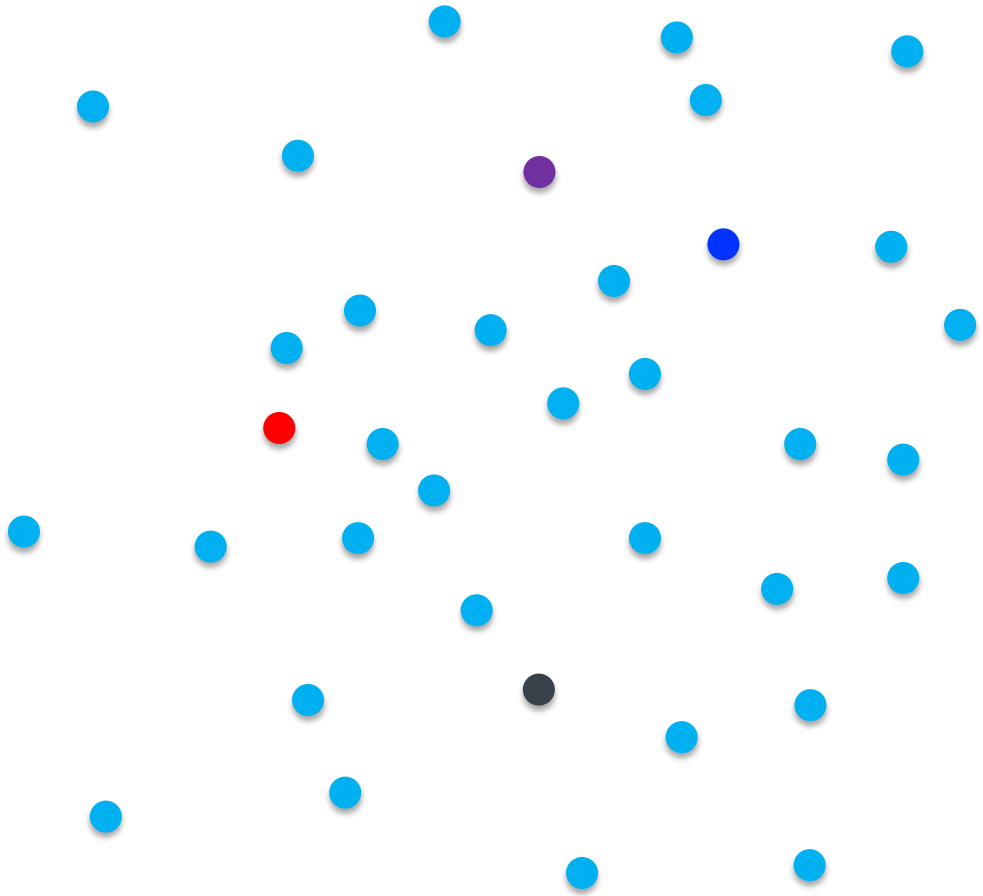
- How do you measure happiness?
  - Again, it is a combination of factors
    - Work-life balance
    - Health (a latent variable itself)
    - Vacation days

Grey shaded area is the ‘latent space’



MERCK

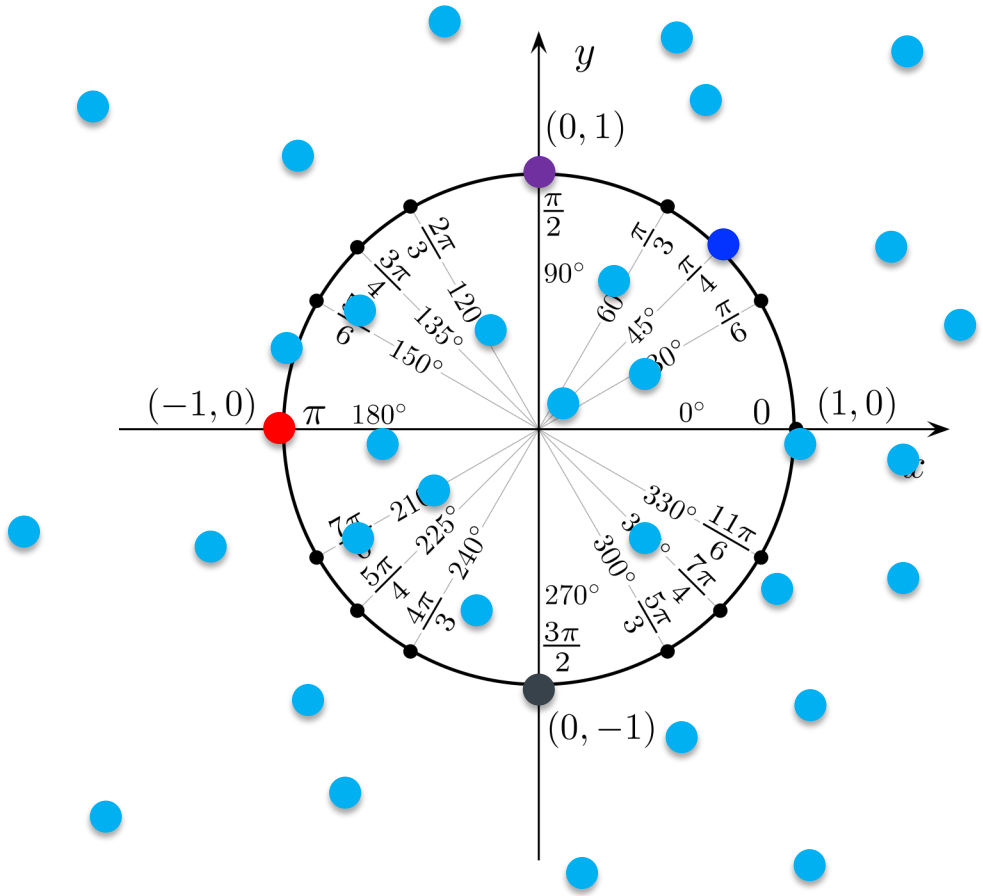
What is the best way to determine
the distance between observations?
Turns out there's quite a few ways…



MERCK

What is the best way to determine
the distance between observations?

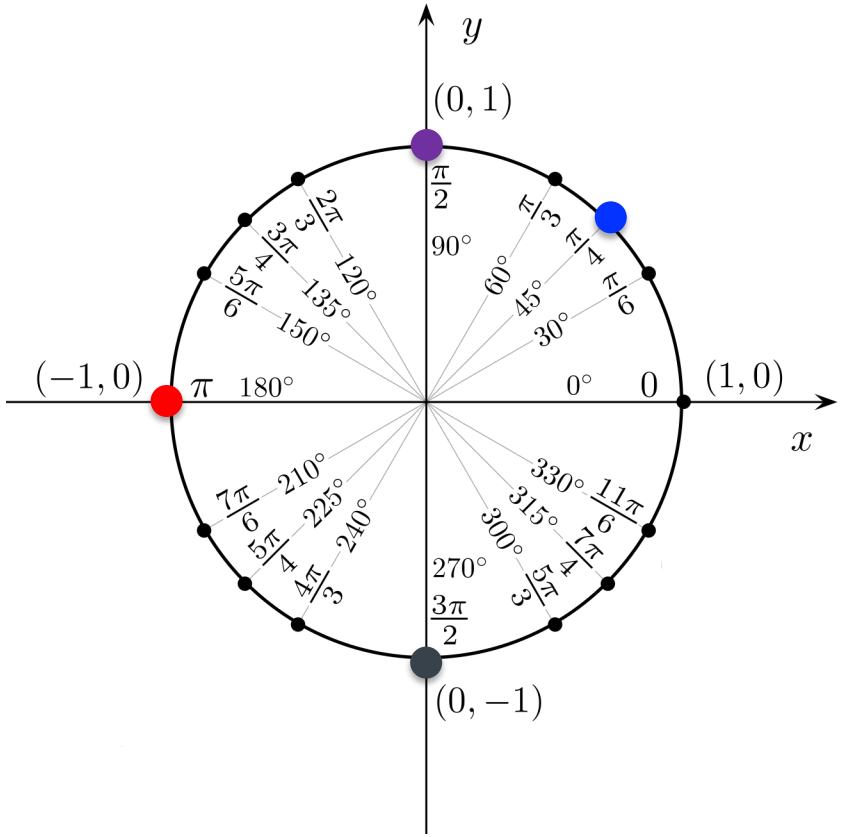
Turns out there's quite a few ways…



MERCK

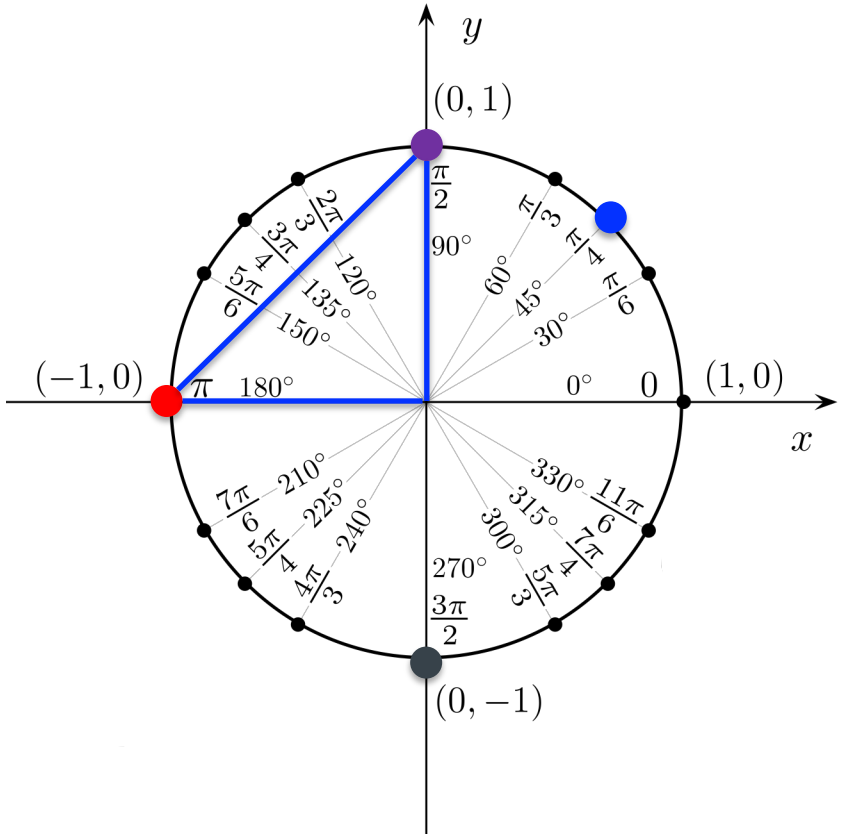
‘Intuitive’ Distance Metrics

Euclid provides a ‘short-cut’ between two points → the predominant method in process monitoring and fault detection

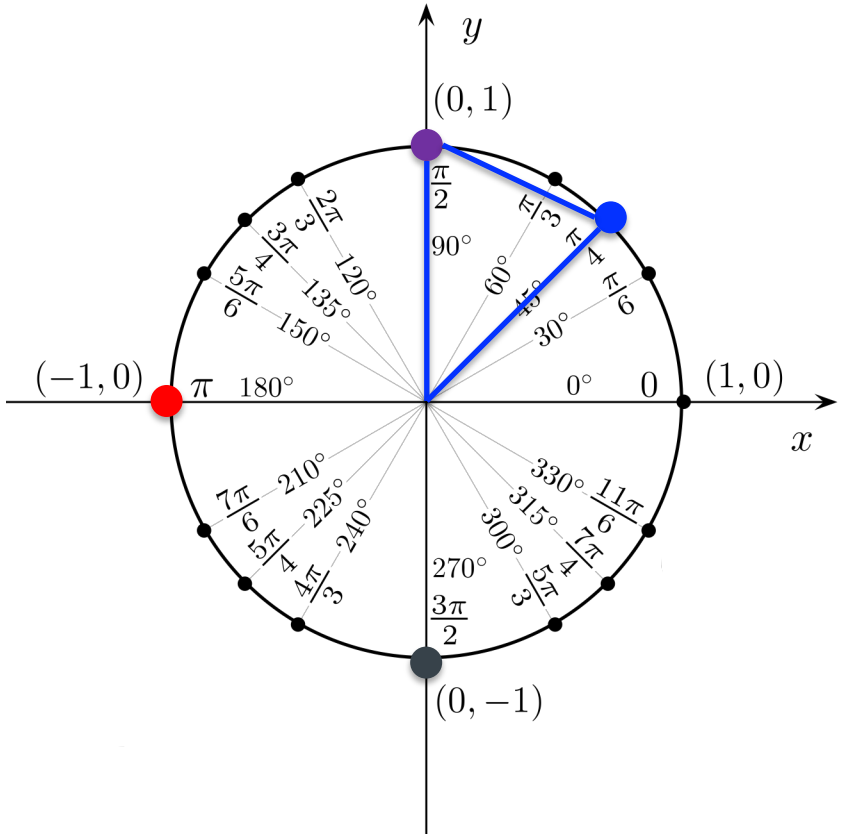


MERCK

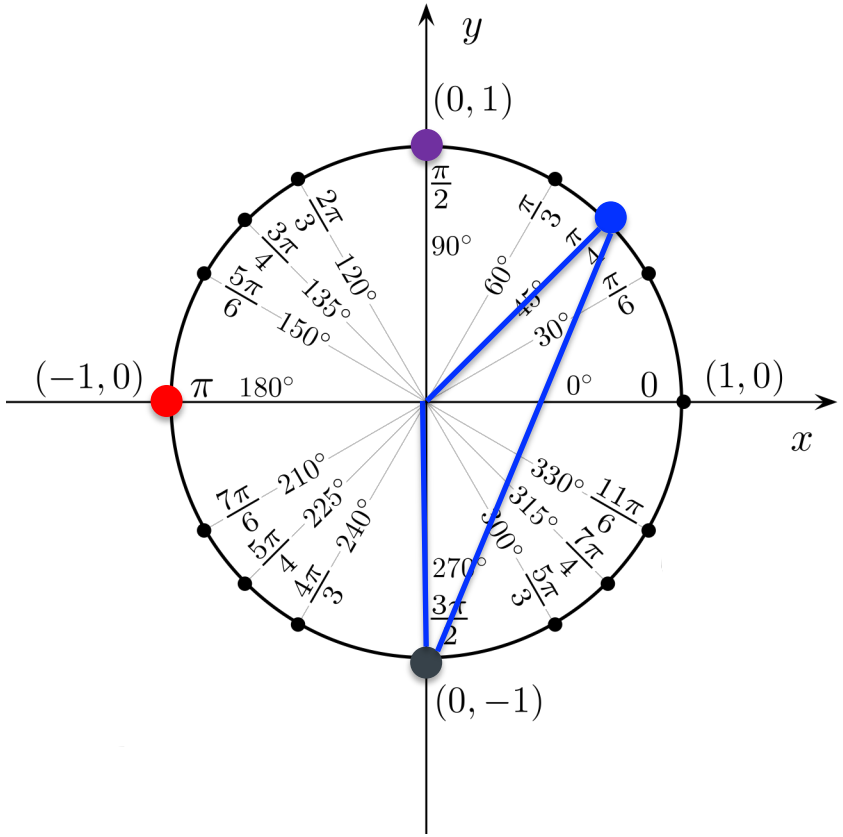
Euclid provides a ‘short-cut’ between two points → the predominant method in process monitoring and fault detection



Euclid provides a ‘short-cut’ between two points → the predominant method in process monitoring and fault detection

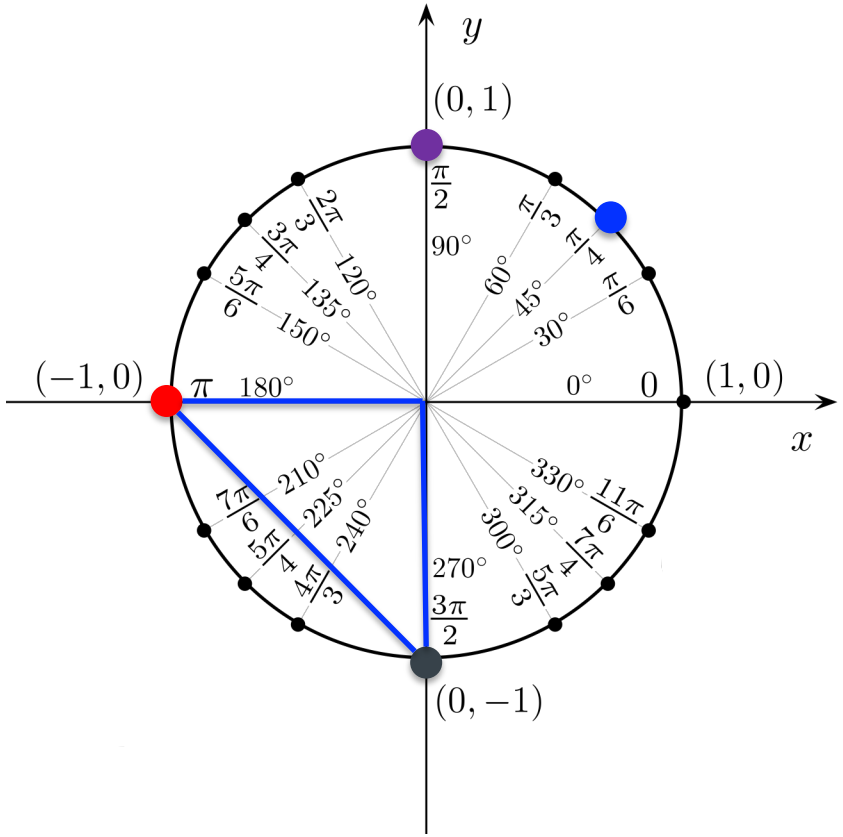


Euclid provides a ‘short-cut’ between two points → the predominant method in process monitoring and fault detection

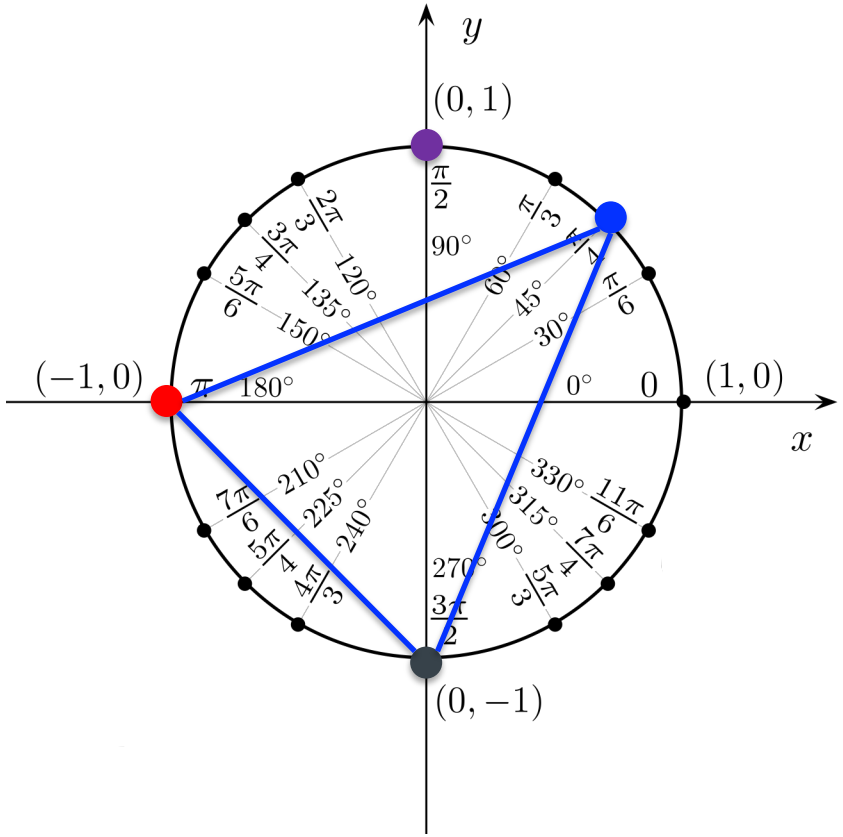


MERCK

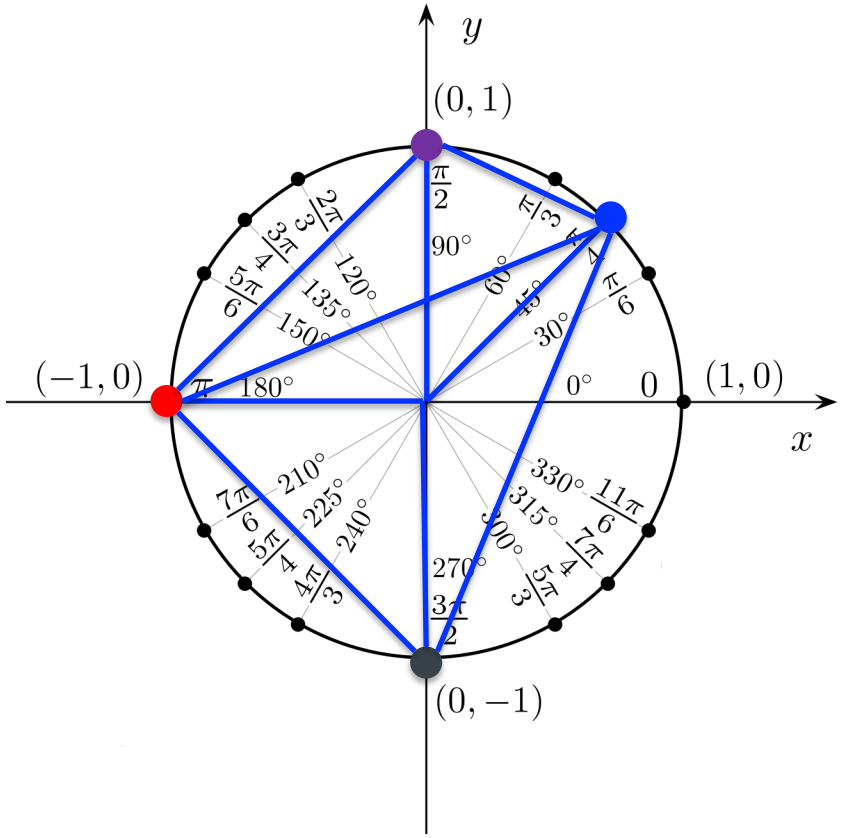
Euclid provides a ‘short-cut’ between two points → the predominant method in process monitoring and fault detection



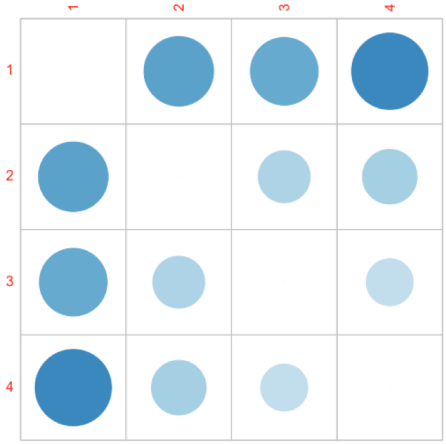
Euclid provides a ‘short-cut’ between two points → the predominant method in process monitoring and fault detection



Euclid provides a ‘short-cut’ between two points → the predominant method in process monitoring and fault detection



Euclidian distance matrix →

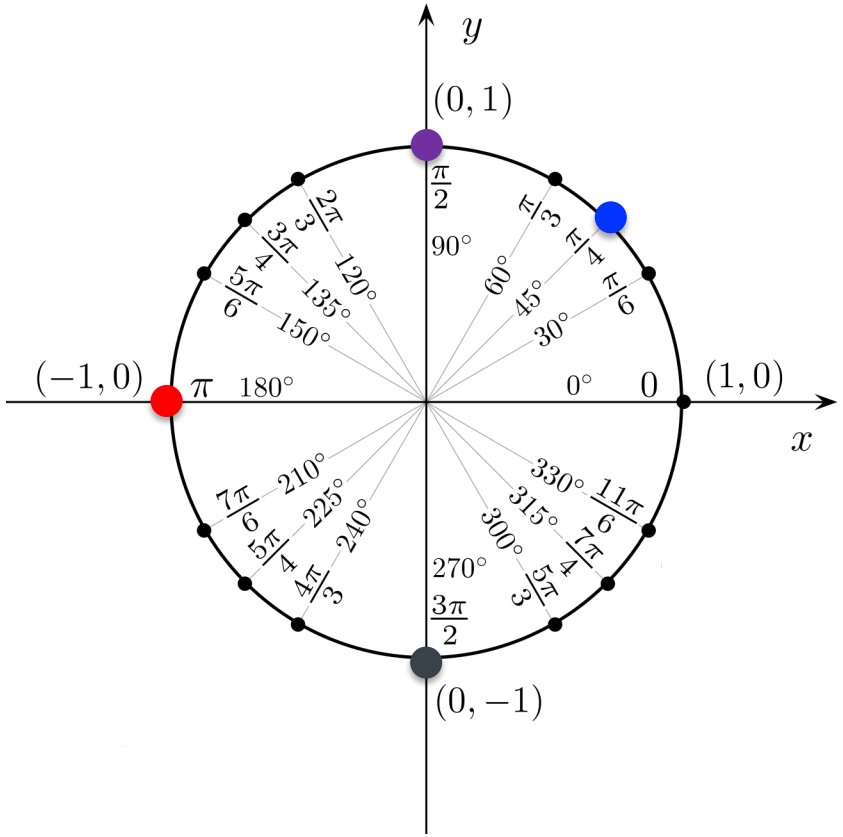


SVD

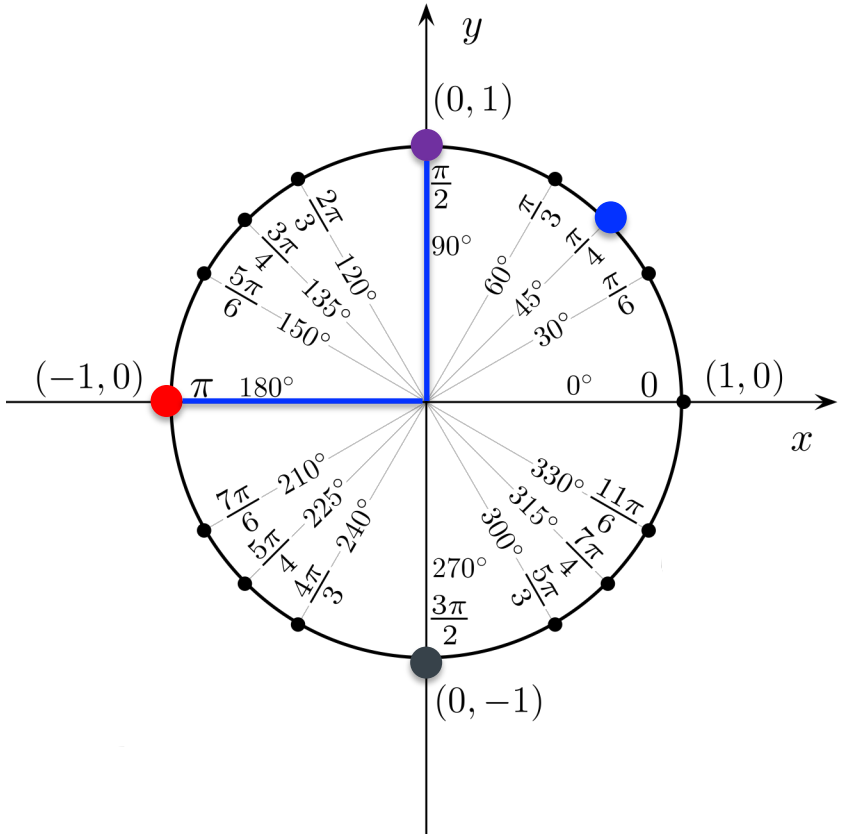
An elegant consequence of using Euclidian distances is that - (double-mean centered)/2 the square of this matrix results in, after mean-centering X, in XX’ → PCA → derive loadings

MERCK

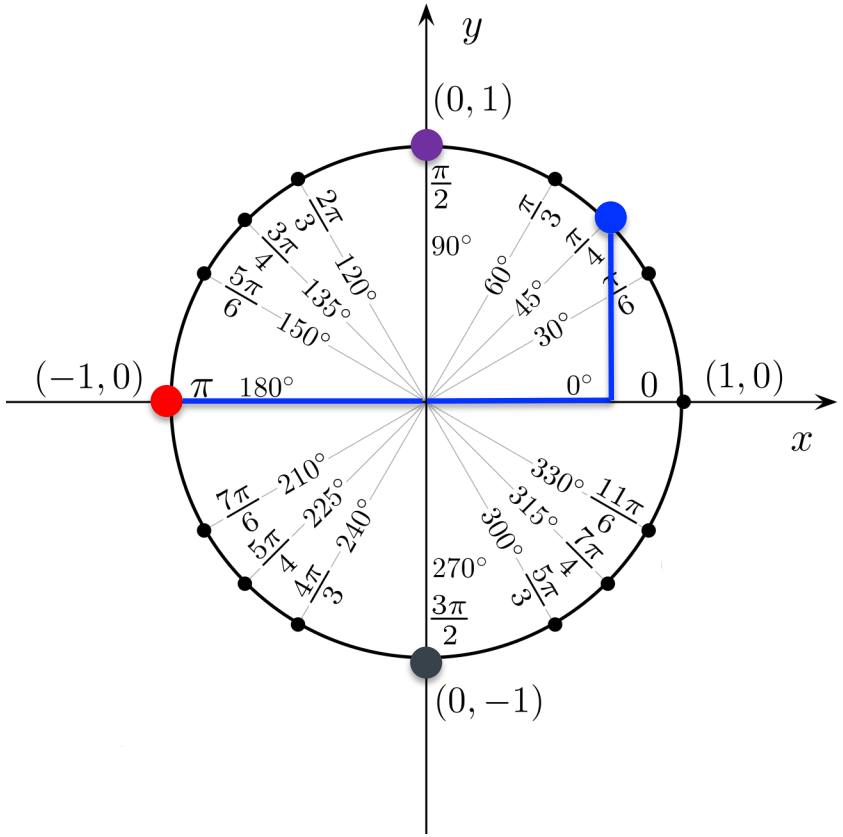
Manhattan (or Minkowski with lambda = 1)
provide the ’taxi-cab’ direction between
two points



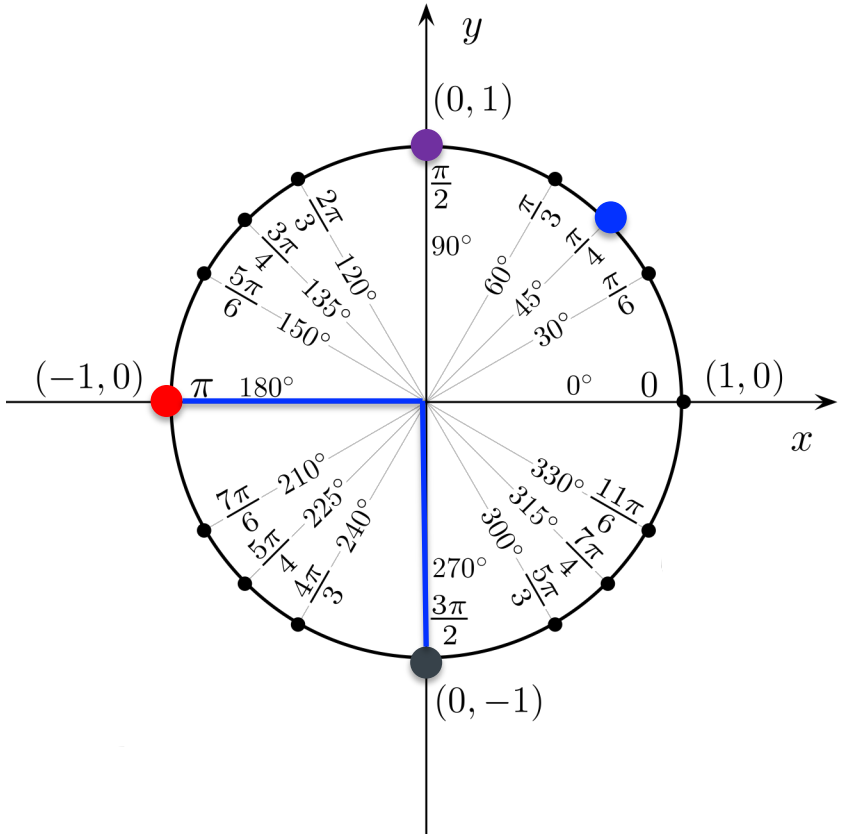
Manhattan (or Minkowski with lambda = 1)
provide the ’taxi-cab’ direction between
two points



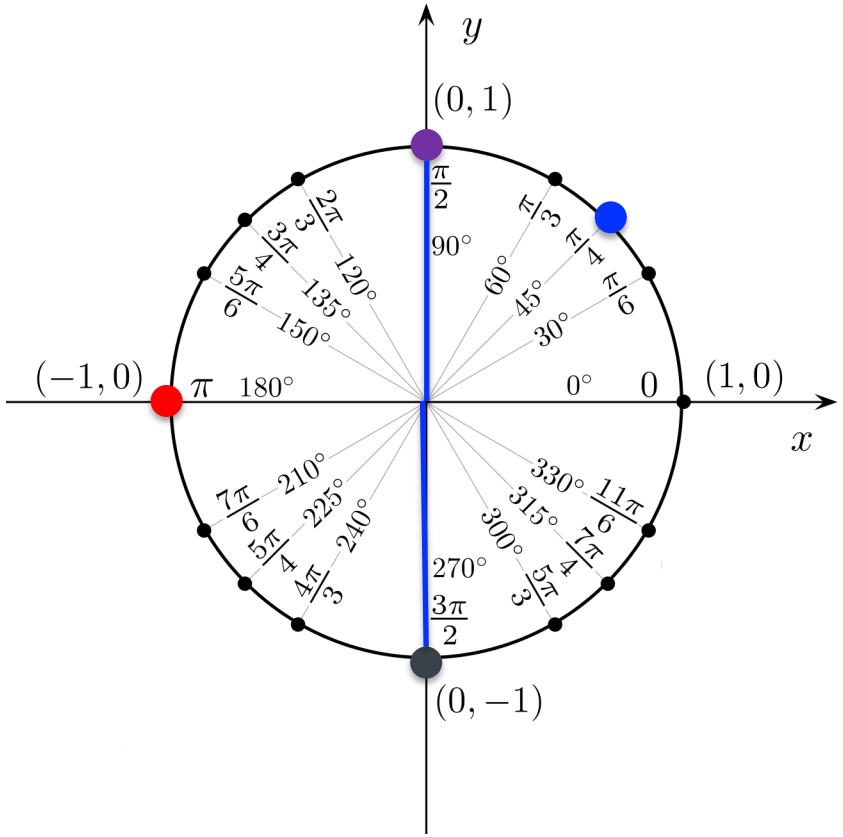
Manhattan (or Minkowski with lambda = 1)
provide the 'taxi-cab' direction between
two points



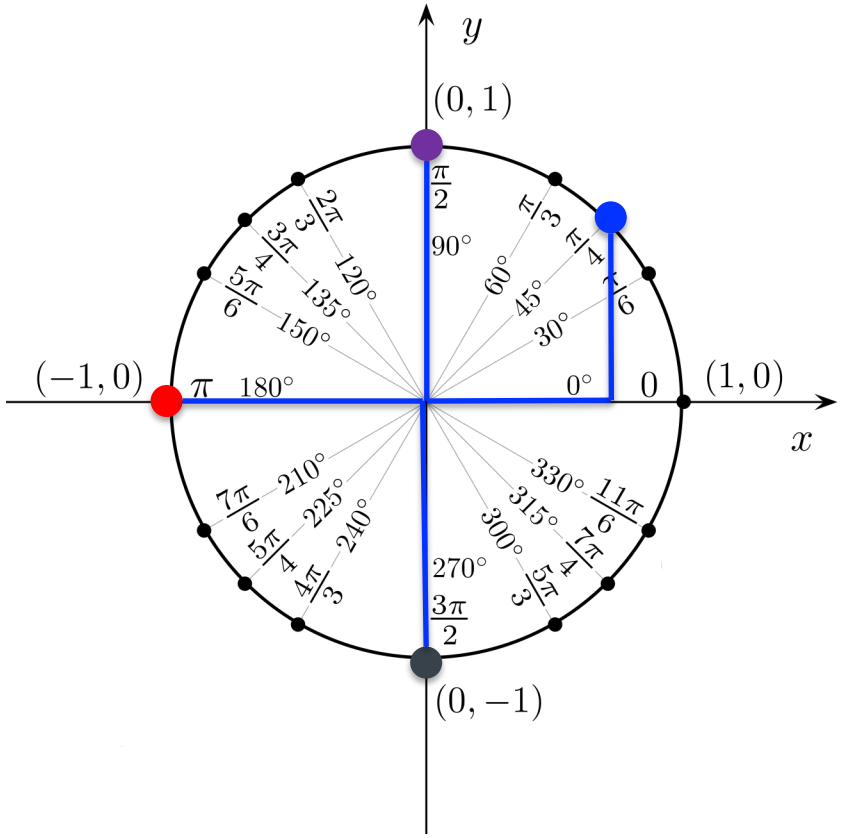
Manhattan (or Minkowski with lambda = 1)
provide the ’taxi-cab’ direction between
two points



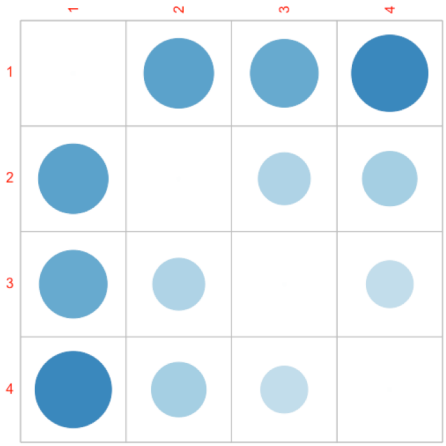
Manhattan (or Minkowski with lambda = 1)
provide the ’taxi-cab’ direction between
two points



Manhattan (or Minkowski with lambda = 1) provide the 'taxi-cab' direction between two points



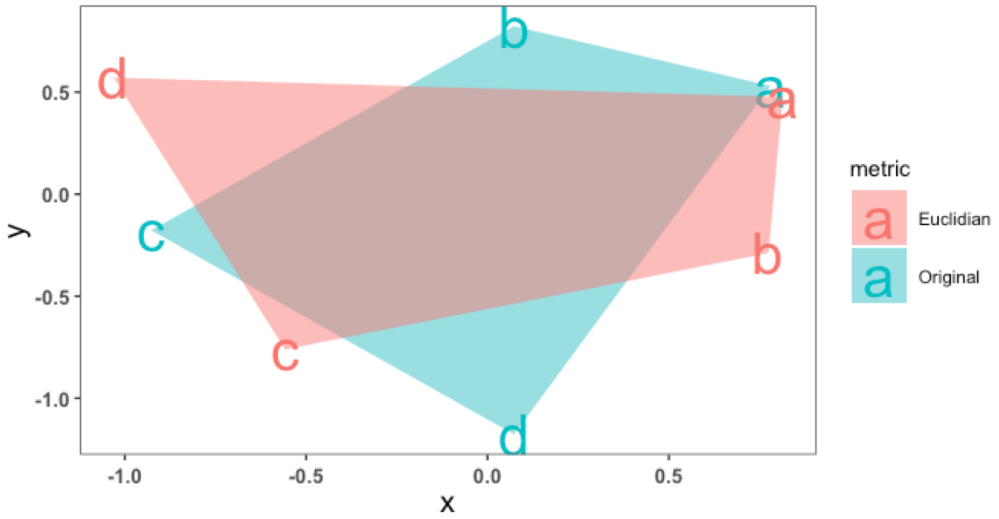
Manhattan distance matrix

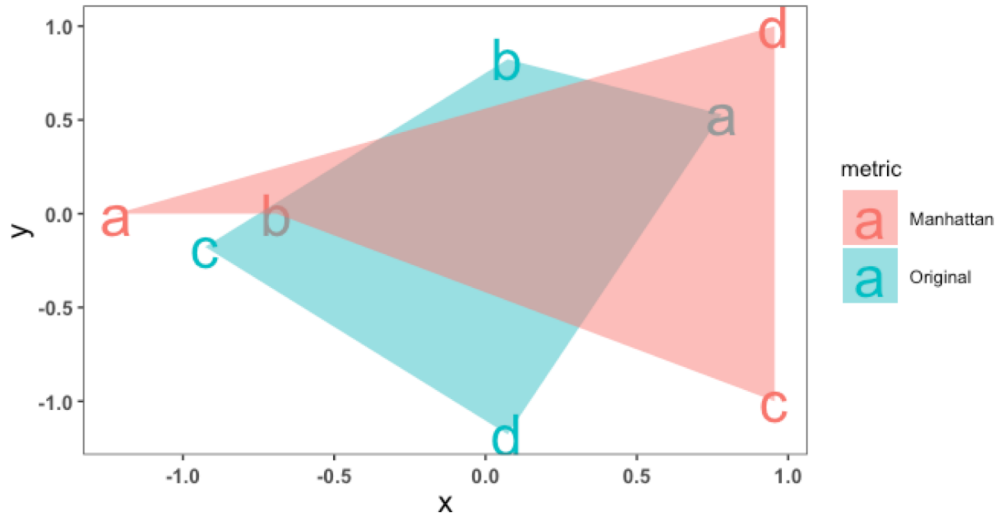


SVD

MERCK

Geometrically what does it all mean?

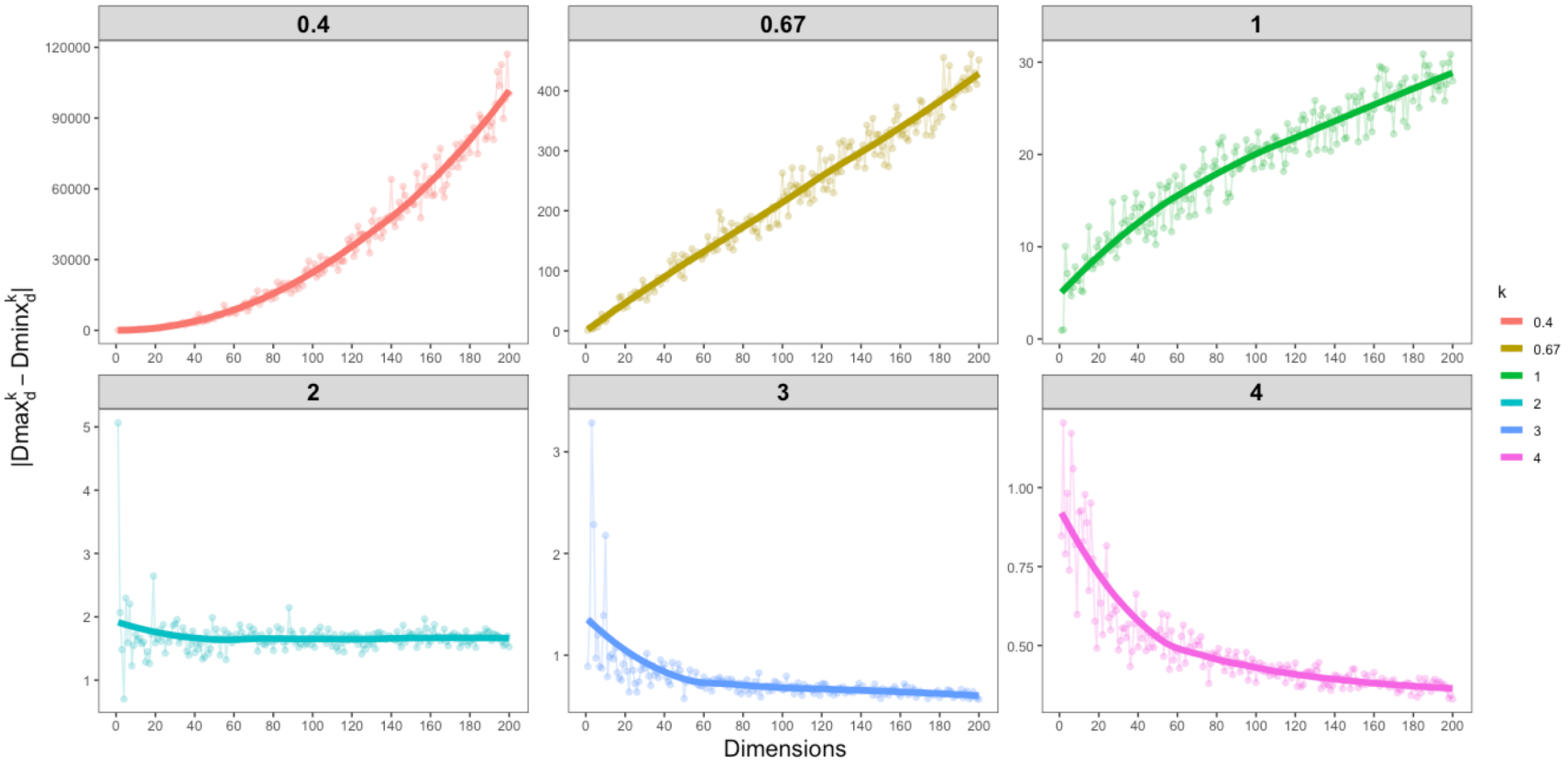




Distance Metrics in High Dimensional Space

$$C_k \leq lim_{d\to\infty} E\left[\frac{Dmax_d^k - Dmin_d^k}{d^{1/k-1/2}}\right] \leq (n-1)\cdot C_k$$

$|Dmax_d^k - Dminx_d^k|$ ~ d; n=2 ~ uniform

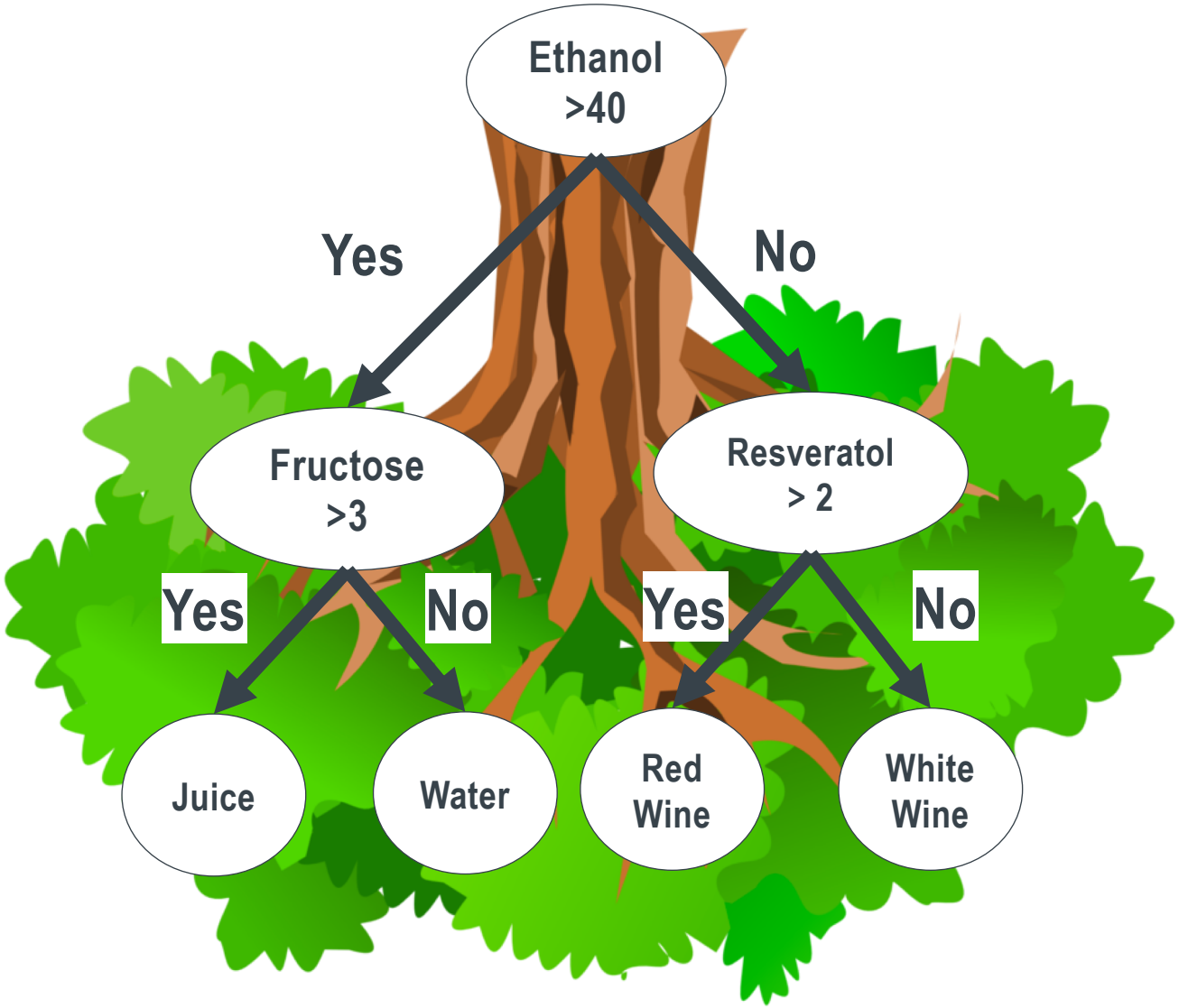


Examples of a not so ‘Intuitive’ Distance Metrics

[Unsupervised] Random Forest

MERCK

[Unsupervised] Random Forest



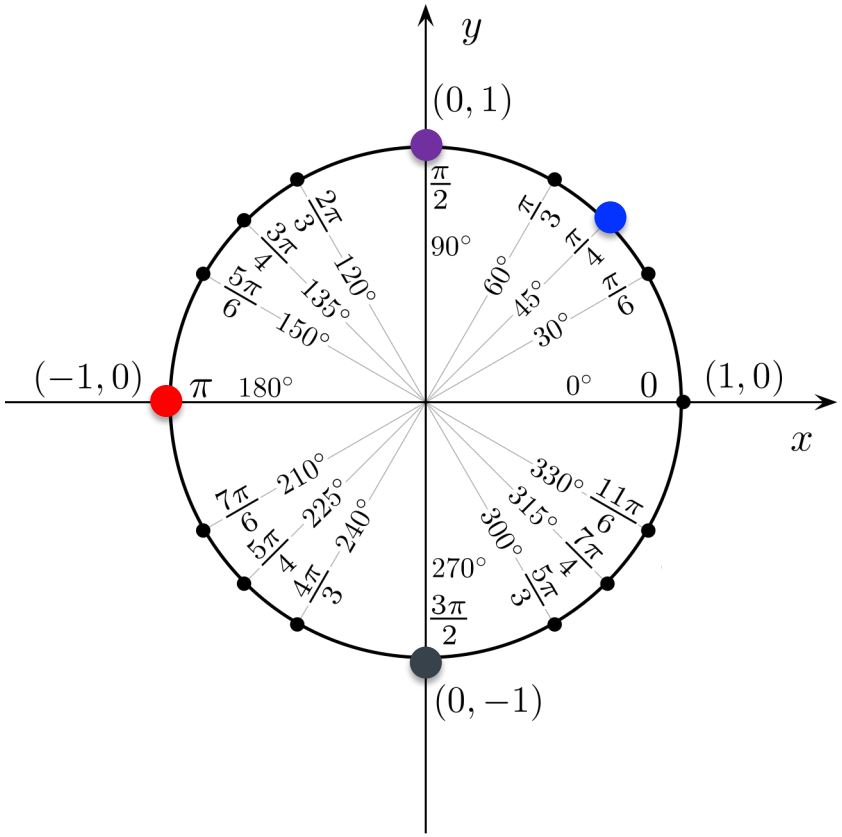
Repeat this process many times but add in

1. bootstrap sample for each tree
2. At each node, randomly select a subset of predictors to determine a split
3. Grow an un-pruned tree
4. Vote across all trees in the forest for classification assignment

In URF, observations that are modeled together in the tree are deemed similar

MERCK

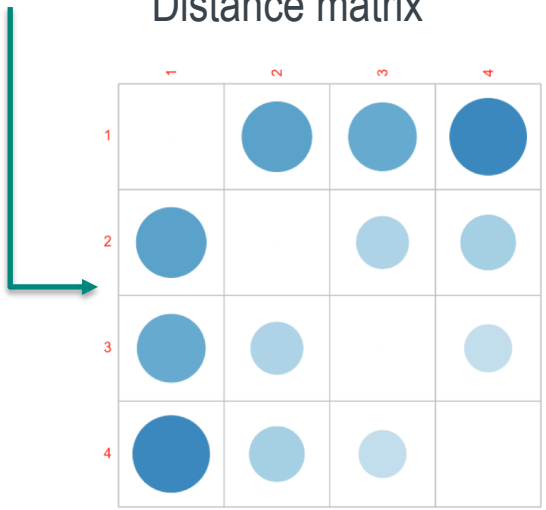
URF – model/probabilistic based distance



Across ***n*** bootstrap samples

	x1	x2	x3	x4	x5	Y
Obs 1	868	252	951	301	344	Actual
Obs 2	78	81	119	855	11	Actual
Obs 3	304	564	247	853	212	Actual
Obs 4	735	197	516	262	115	Actual
Obs 5	216	813	602	587	646	Actual
Obs 6	137	814	30	842	299	Actual
Obs 7	415	201	655	292	269	Actual

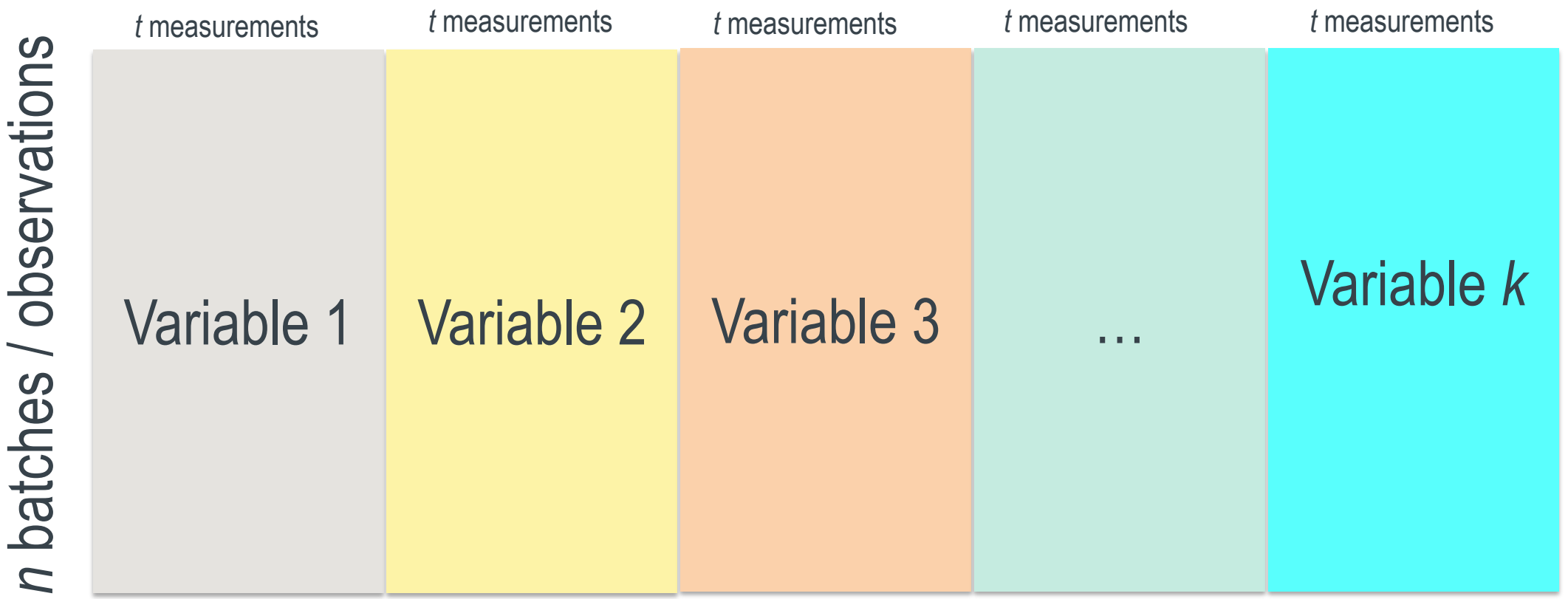
Distance matrix



Biologics Mfg. Simulated Example where we perform a multivariate process scale comparison

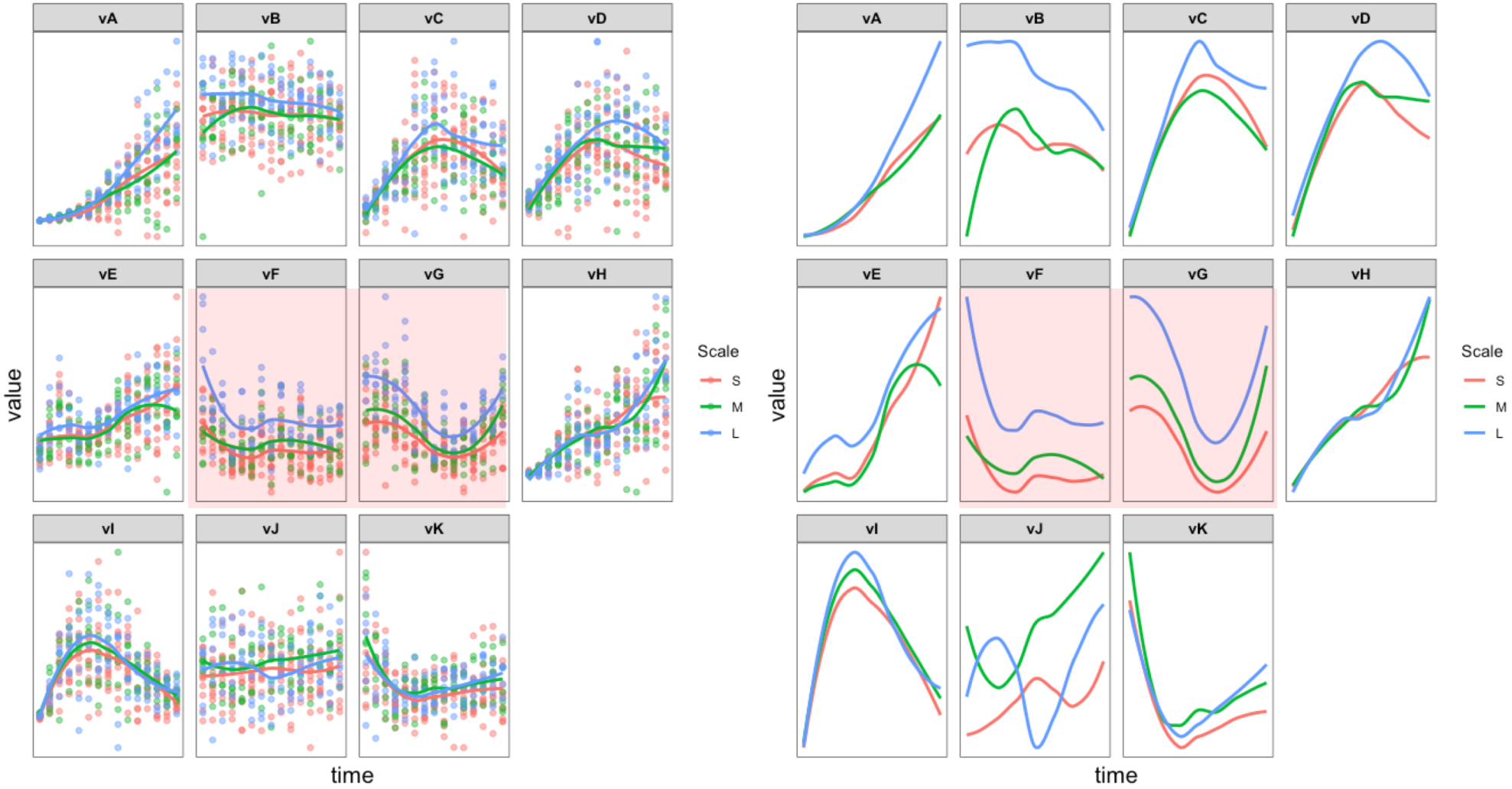
MERCK

‘Multi-Block’ Data Matrix



MERCK

Simulated Process Trajectory Curves



Data Pre-processing

Principal Component Analysis

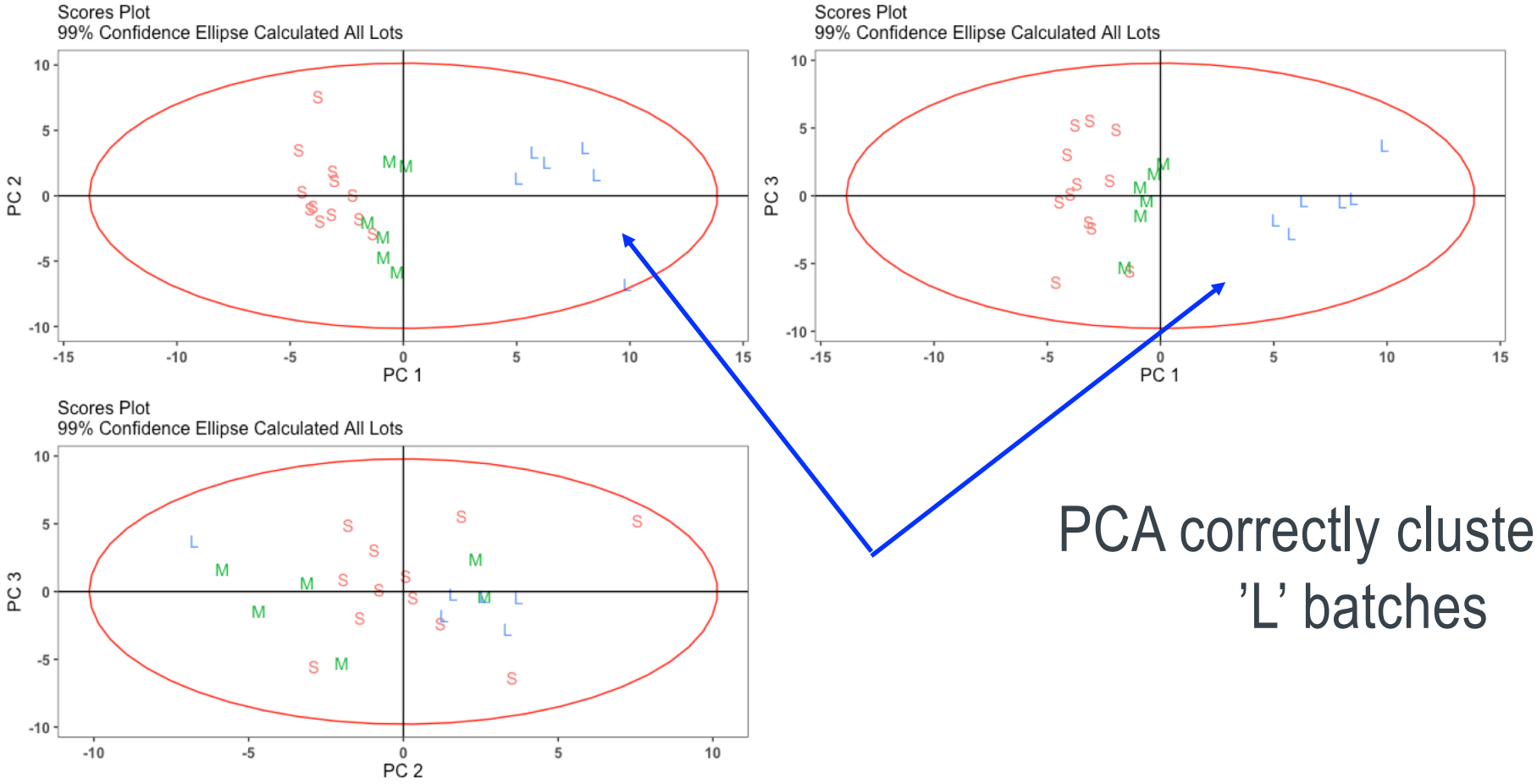
- If categorical variables present, perform one-hot encoding
- Scale to unit-variance
- Ready for analysis

Unsupervised Random Forest

- One-hot encoding not necessary for categorical variables
  - Means that categorical variable are assessed as a single contributor to the model
- Ready for analysis

MERCK

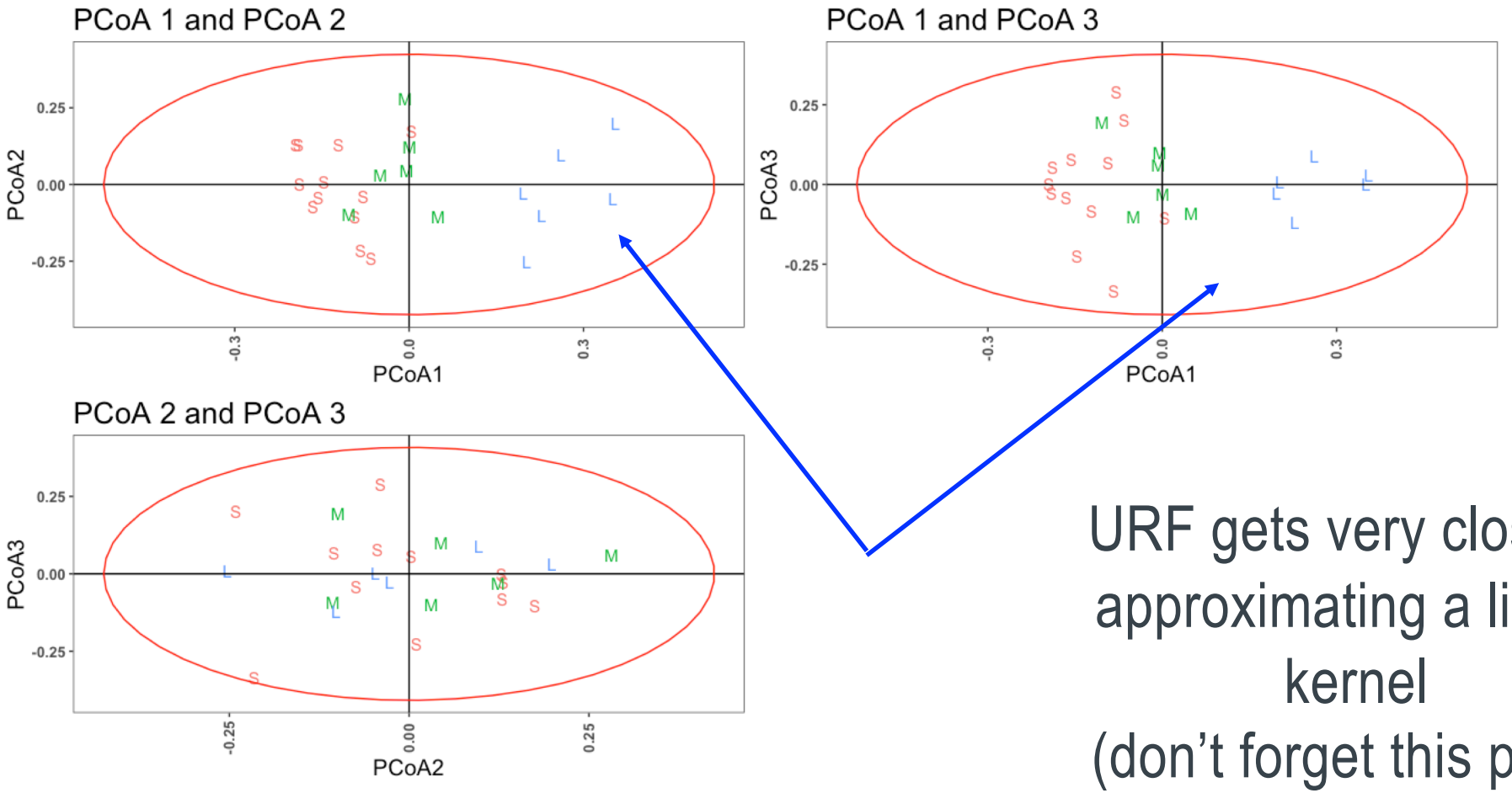
PCA Scores (3 Principal Components)



PCA correctly clusters the 'L' batches

MERCK

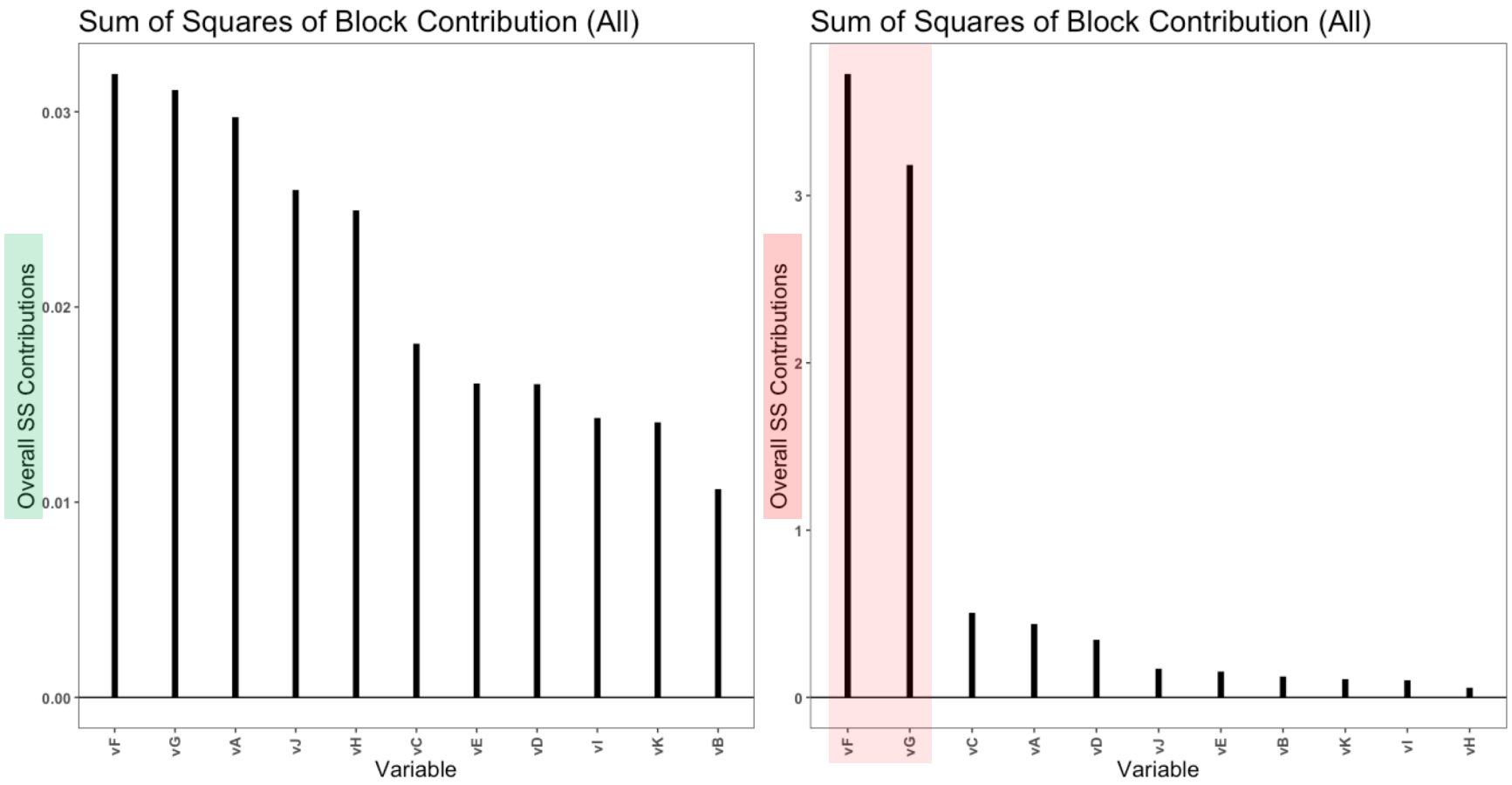
URF PCoA Scores (3 Principal Coordinates)



URF gets very close to approximating a linear kernel
(don't forget this point)

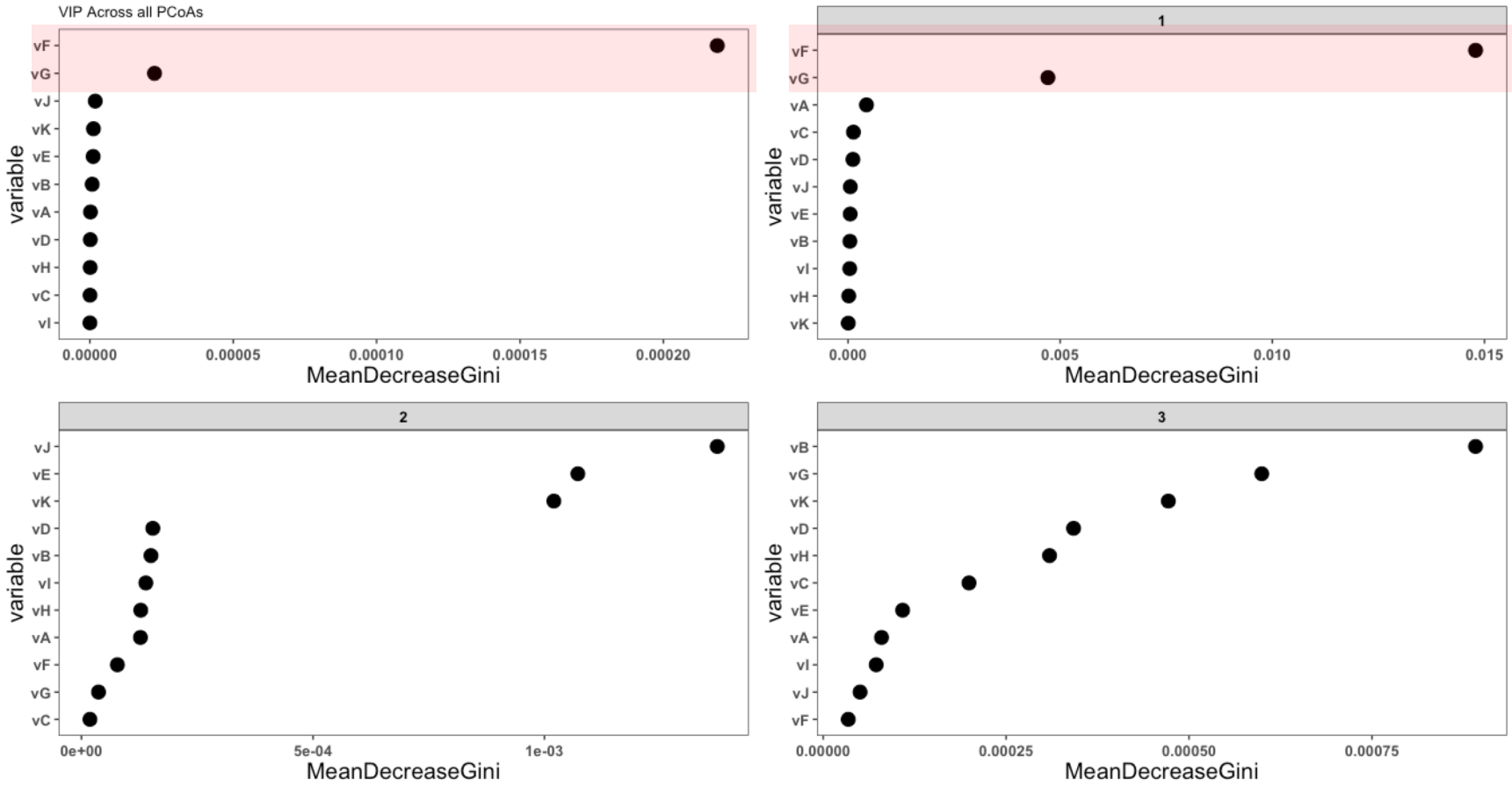
MERCK

Variables Contributing to PCA Clustering

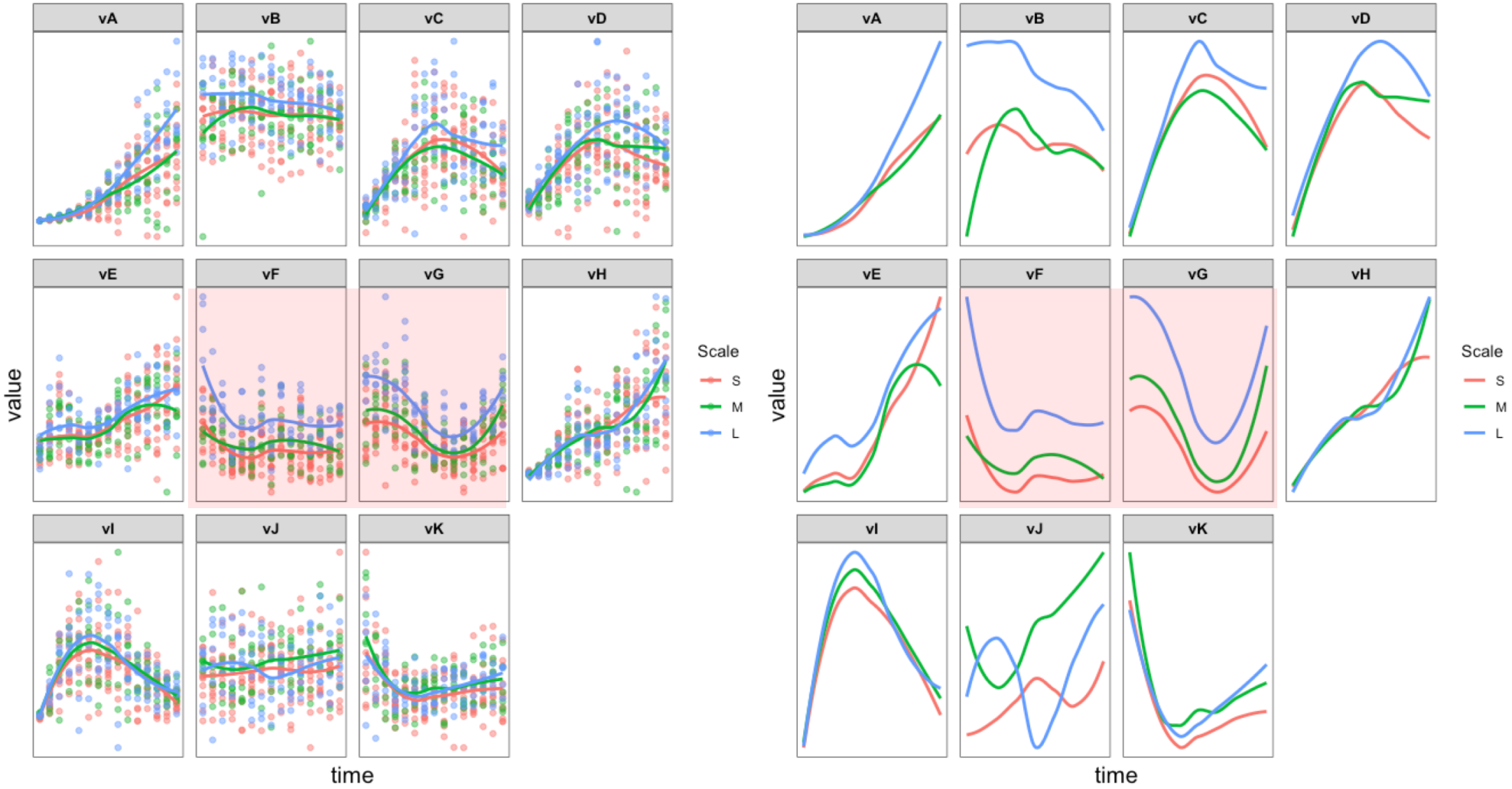


Note differences w.r.t. contribution

Variables Contributing to PCoA Clustering



What did PCA & URF discover?



Note that if we looked at the unscaled data using PCA the results could change dramatically

MERCK

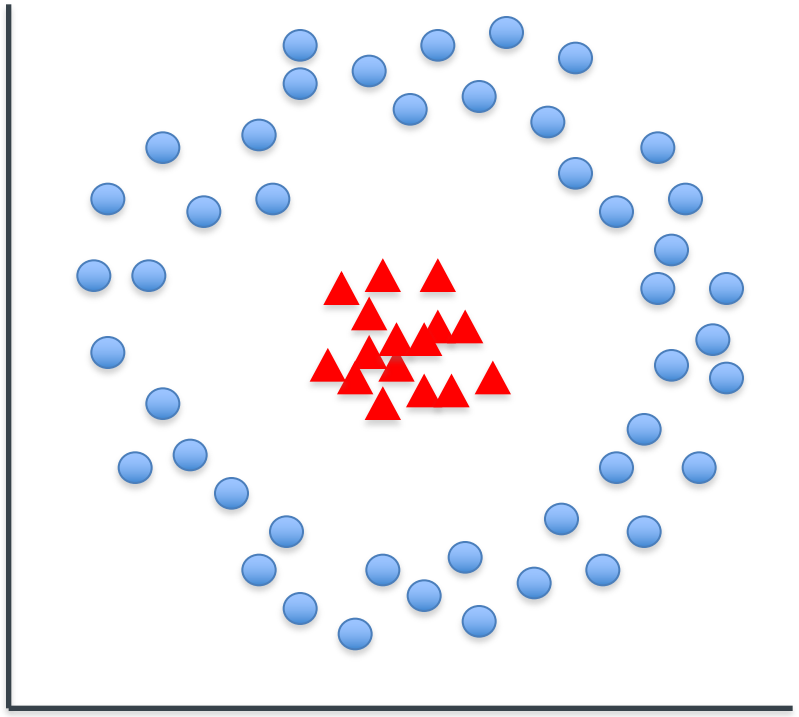
But wait, there's more & here is where it gets more interesting (i.e. challenging)!

MERCK

What if I’m dealing with highly non-linearly separable original spaces?

Here we bring in Kernel PCA

Kernel PCA



$$\phi:$$

$$\Re^2 \qquad \Re^3$$

$$\mapsto (z_1, z_2, z_3) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$$

$$(x_1, x_2) \quad k(x_i, x_j) = x_{i1}x_{j1} + x_{i2}x_{j2} + \ldots + x_{ik}x_{jk} = XX'$$

*Assuming a centered homogenous kernel

$$(XX')^2$$

Voila, the 'kernel trick'!

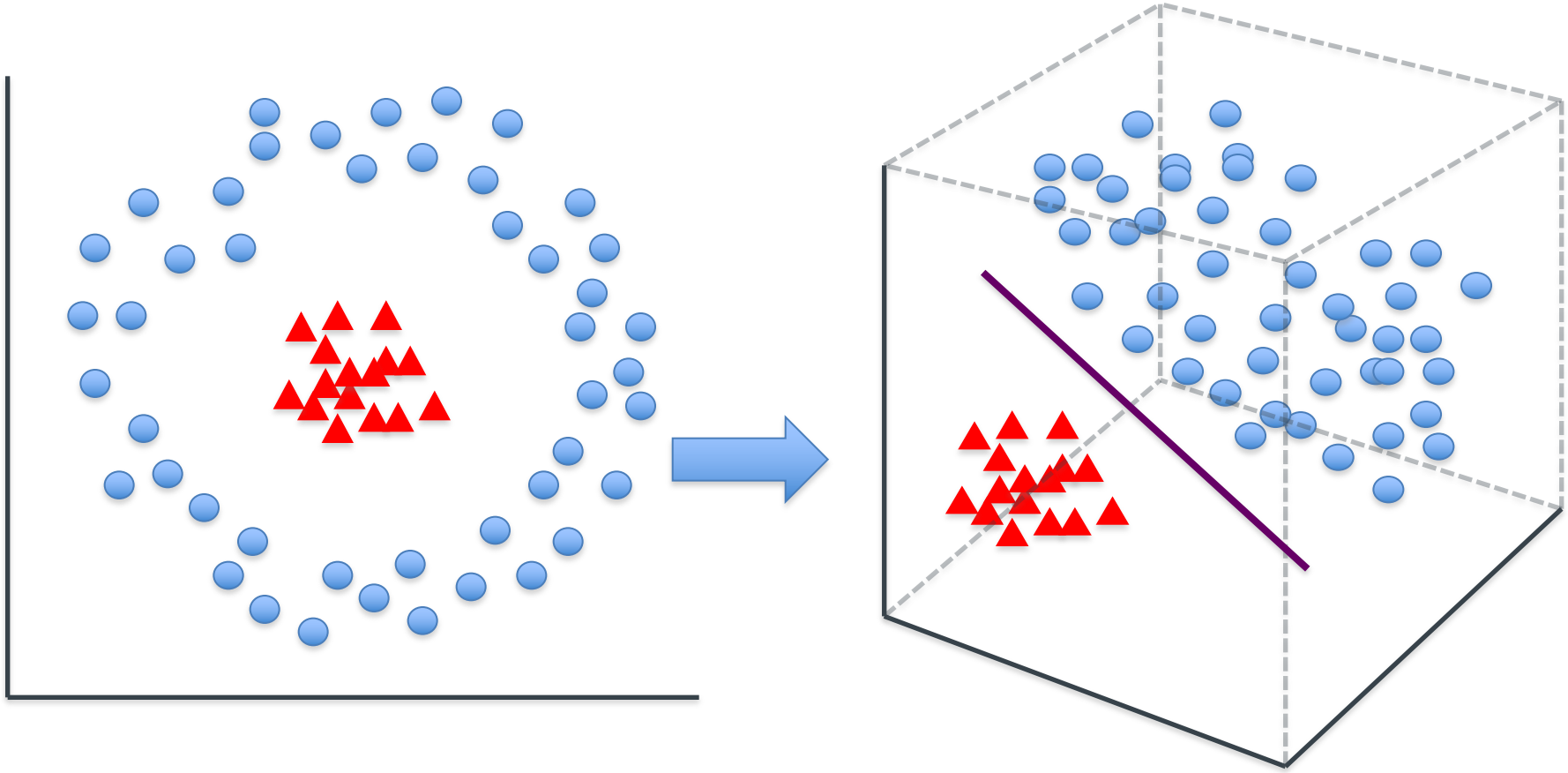
$$k(x_i, x_j)^2 = x_i^2, \sqrt{2}x_ix_j, x_j^2$$

$$svd(XX')$$

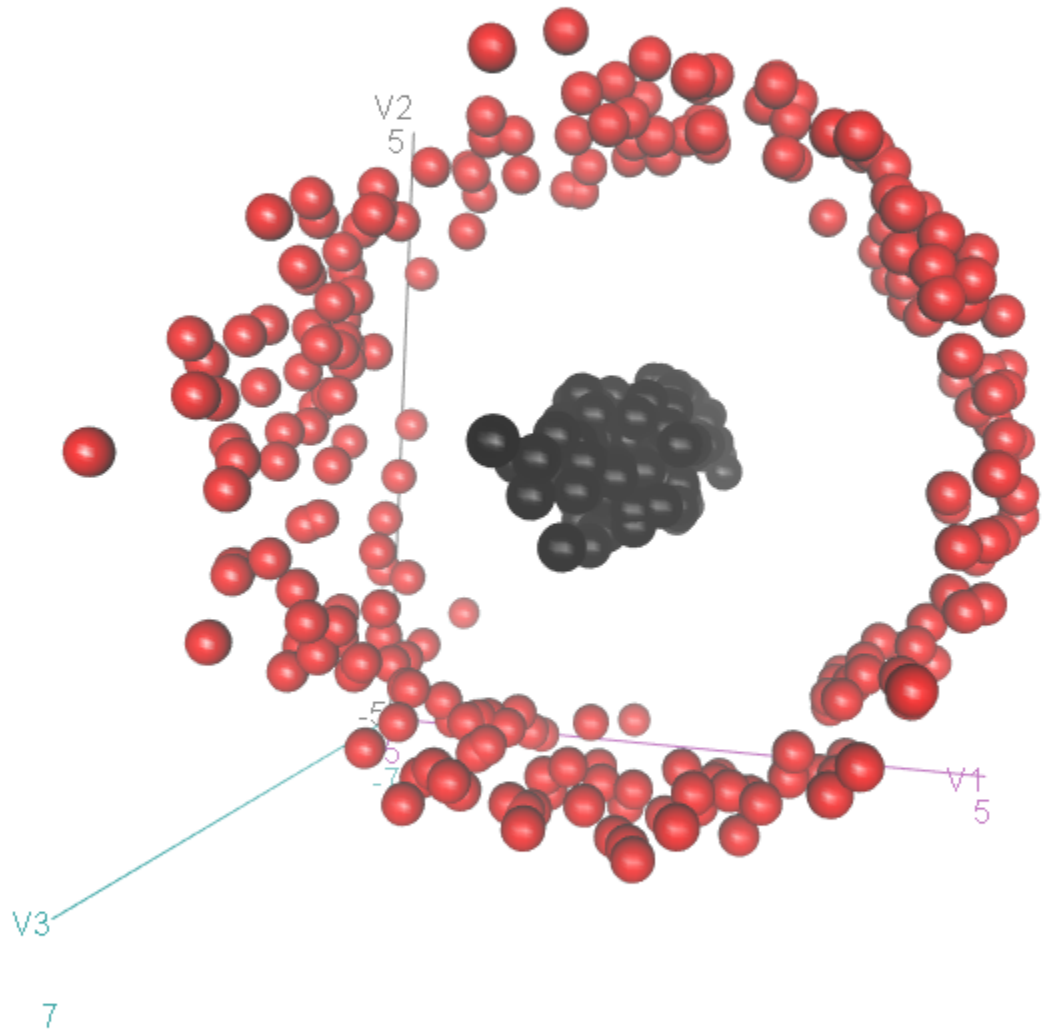
Why do we care? Because now all we need to do is simply calculate the inner product and *implicitly* map to the higher-dimensional space, i.e., we do not have to map to ϕ and THEN compute the inner product.

MERCK

Kernel PCA



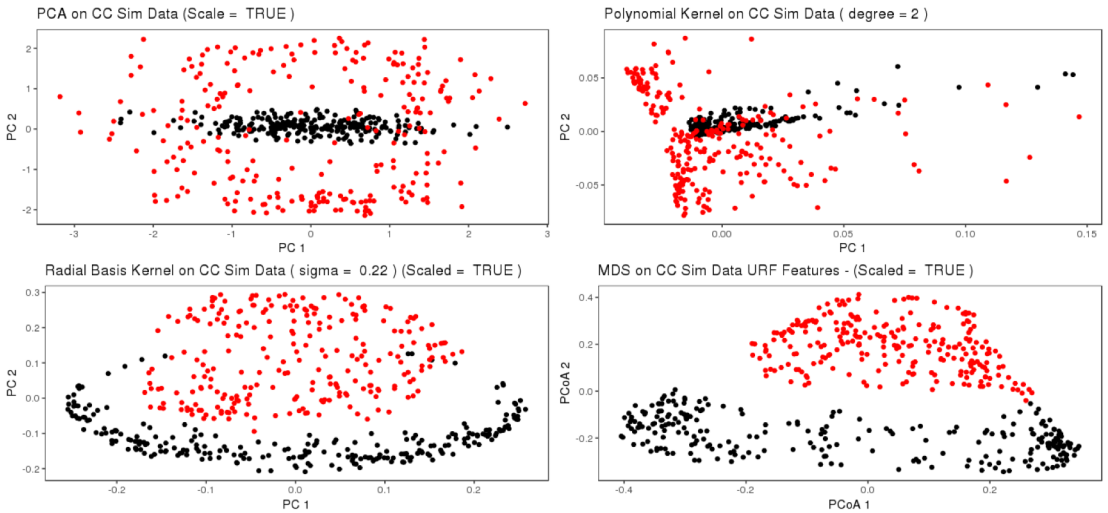
MERCK



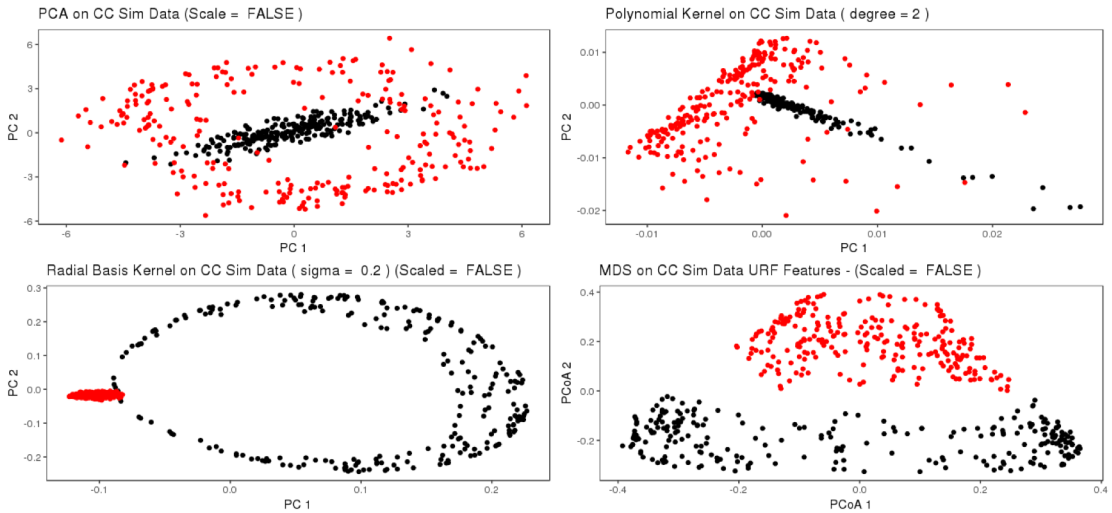
MERCK

2D Projections

Scaled to Unit Variance



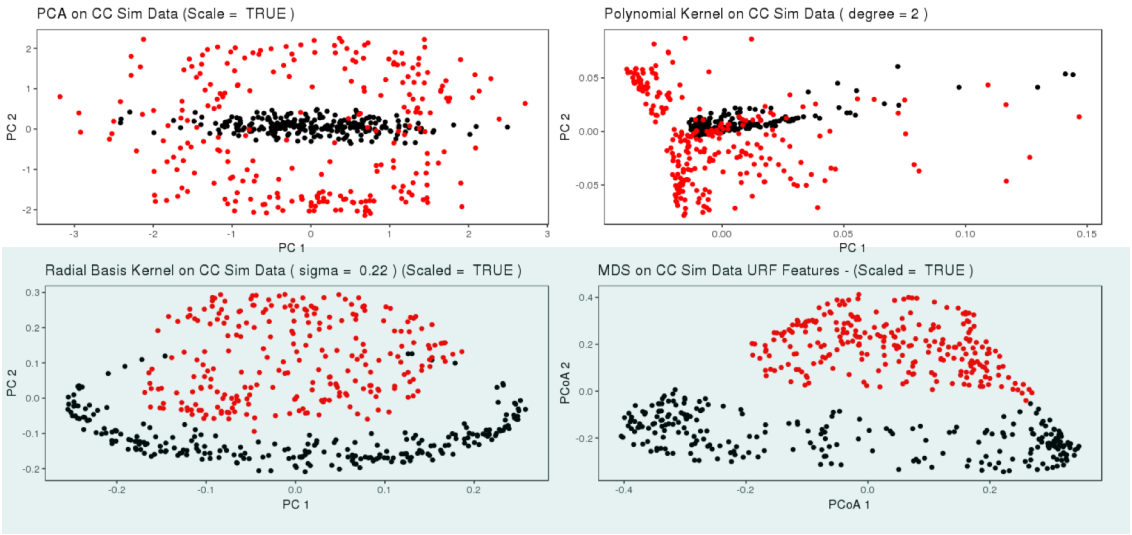
Not Scaled to Unit Variance



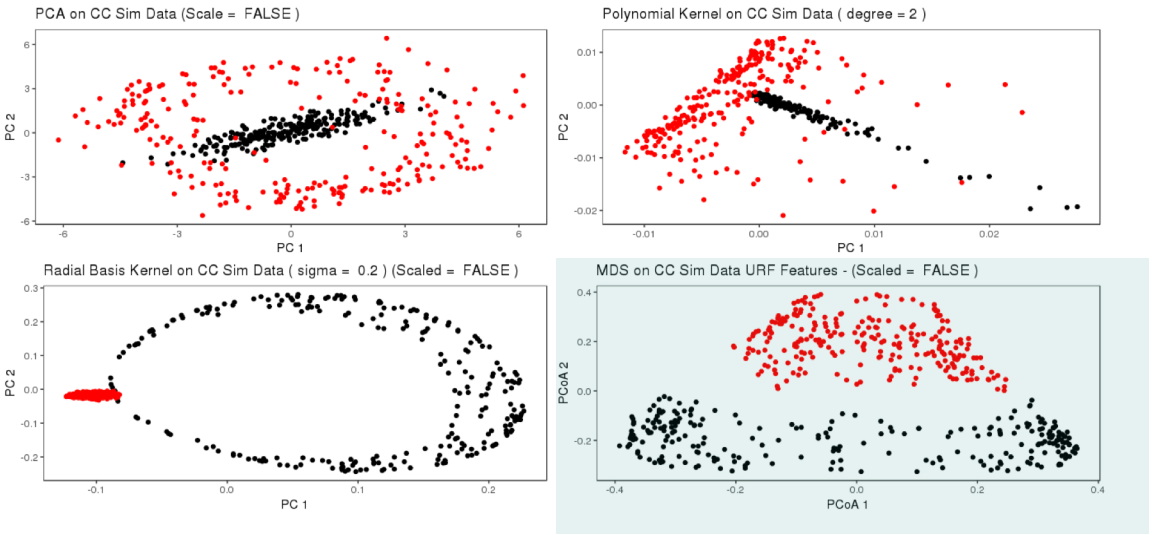
MERCK

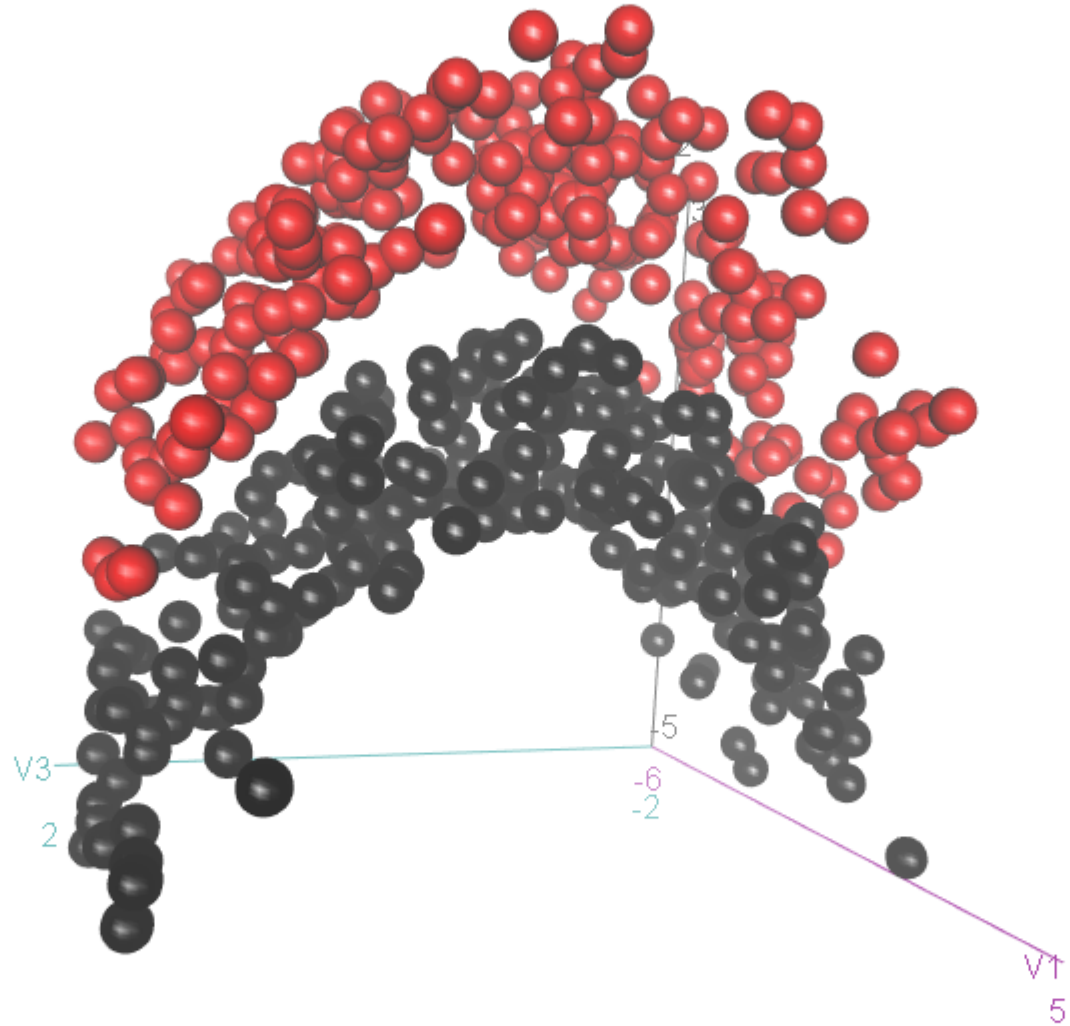
2D Projections

Scaled to Unit Variance



Not Scaled to Unit Variance

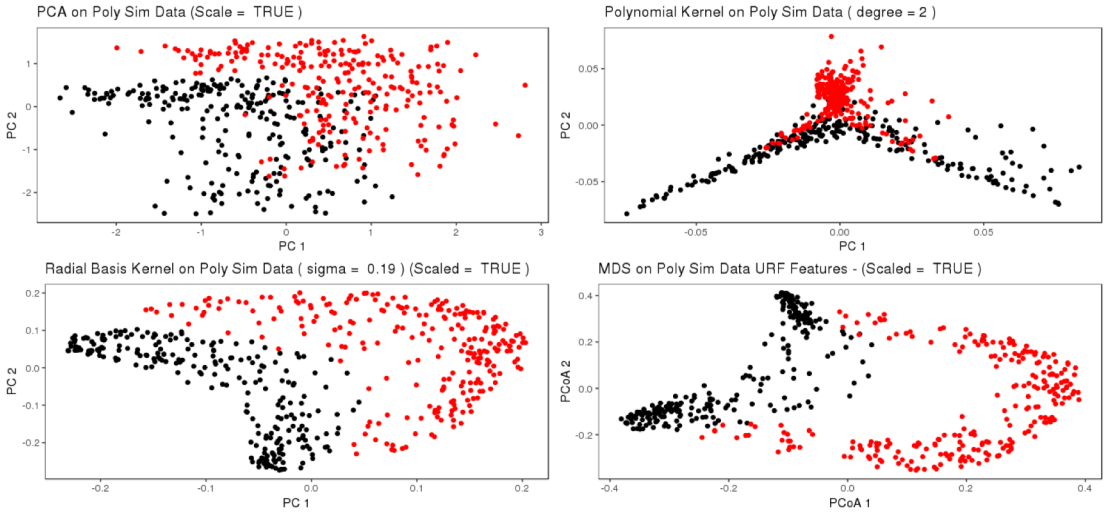




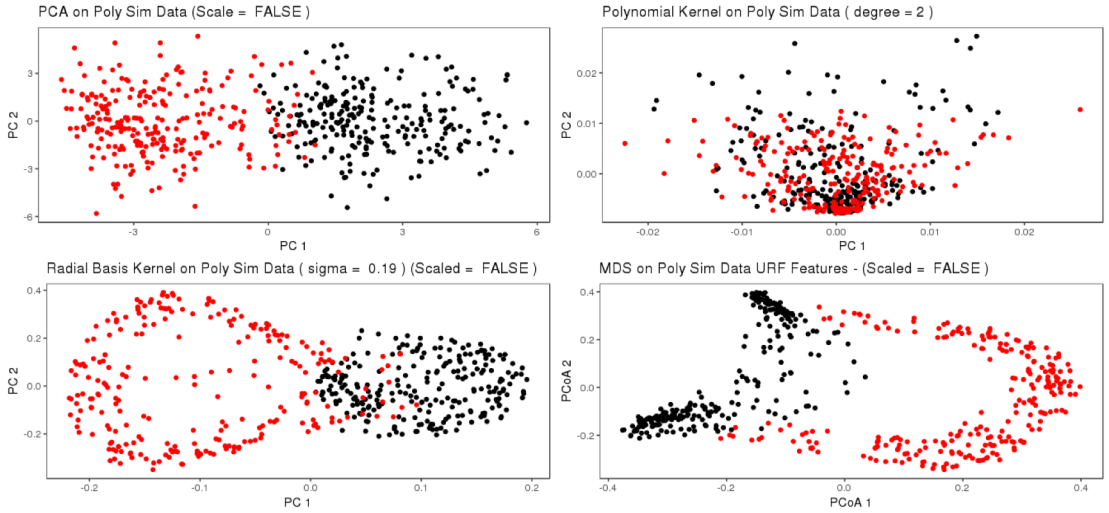
MERCK

2D Projections

Scaled to Unit Variance



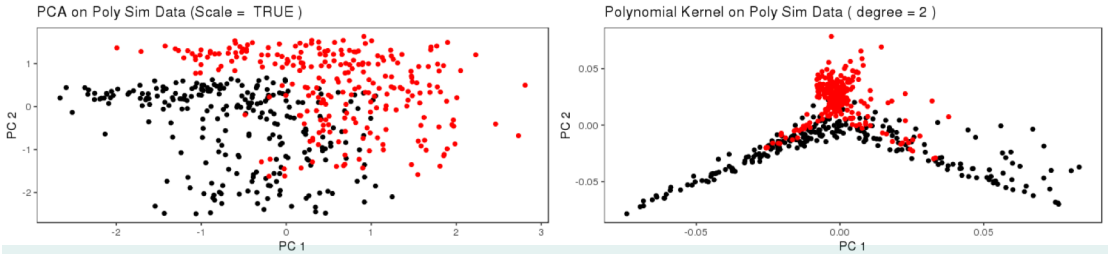
Not Scaled to Unit Variance



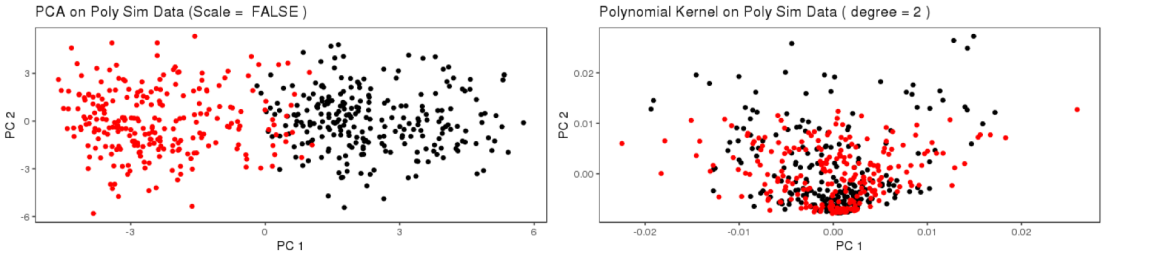
MERCK

2D Projections

Scaled to Unit Variance



Not Scaled to Unit Variance



MERCK

Final Thoughts

- Distance metrics abound
- Successful dimensionality reduction is dependent on finding a suitable distance metric in order to carry out the 'optimal' embedding of the original data unto a lower-dimensional space
- The singular use of Euclidian distances (among other L_k-Norms) may be suboptimal
  - scale dependent (like others)
  - curse of dimensionality
  - non-linearly separable structures
  - forces an increase in dimensionality in the presence of categorical variables due to one-hot encoding
- Method in unsupervised machine learning should be explored in parallel with current methods to determine their usefulness

MERCK

References

Aldrich, C., Auret, L., *Unsupervised Process Monitoring and Fault Diagnosis with Machine Learning Methods*, Springer-Verlag London 2013

Borg, I., Groenen, P., *Modern Multidimensional Scaling – Theory and Applications*, Springer Science+Business Media New York, 1997

Aggarwal C.C., Hinneburg A., Keim D.A. (2001) On the Surprising Behavior of Distance Metrics in High Dimensional Space. In: Van den Bussche J., Vianu V. (eds) Database Theory — ICDT 2001. ICDT 2001. Lecture Notes in Computer Science, vol 1973. Springer, Berlin, Heidelberg

MERCK

R Packages

Nelson Lee Afanador, Thanh Tran, Lionel Blanchet, and Richard Baumgartner* (2016). mvdalab: Multivariate Data Analysis Laboratory. R package version 1.2.

Andy Liaw, and Matt Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Alexandros Karatzoglou, Alex Smola, Kurt Hornik, Achim Zeileis (2004). kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9), 1-20.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

MERCK

THANK YOU!

MERCK