Microsoft Research

Each year Microsoft Research hosts hundreds of influential speakers from around the world including leading scientists, renowned experts in technology, book authors, and leading academics, and makes videos of these lectures freely available.

2016 © Microsoft Corporation. All rights reserved.
Manifold-Learning Yields Insight into Complex Cellular State Space

Smita Krishnaswamy, David van Dijk
Yale University
September 10, 2018
High Dimensional

High Throughput

Heterogeneous
Challenges

- Sparsity
- Noise
- Non-linearity
- Scale

How do we ‘look’ at large, high-dimensional, and noisy datasets?
High Dimensional

- 72 channels
- 25,000 genes
- 100,000s of genomic locations
Low dimensional structure
Low dimensional structure

- Many genes but high degree of redundancy
Low dimensional structure

- Many genes but high degree of redundancy
- Cell types / states are constrained
Manifolds: Low Dimensional, Smooth Patches

- The data has a shape
- The shape is learned by non-linear dimensionality reduction
- Methods include:
  - Graph signal processing
  - Deep Learning
How do we learn global structure?
How do we learn global structure?

• Walk via small local steps in the data and compute probability of reaching every node
How do we learn global structure?

- Walk via small local steps in the data and compute probability of reaching every node.
- Random walk = diffusion.
MAGIC
Markov Affinity-based Graph Imputation of Cells
(van Dijk et al. Cell 2018)

David van Dijk
Kevin Moon
Kristina Yim
Guy Wolf
Dana Pe’er
MAGIC
Markov Affinity-based Graph Imputation of Cells
(van Dijk et al. Cell 2018)

Main idea: cells learn values from their neighbors, gaining information from similar cells while retaining their individual identity.
Data is noisy and sparse (scRNA-seq)
Data is noisy and sparse (scRNA-seq)
HMLE breast cancer cell line, TGF-6 induced EMT

 dropout obscures gene-gene relationships
i. Original Data

ii. Calculate Distances
   Euclidean distance
   
   \[ \text{dist}(C_i, C_j) = \sqrt{\|C_i - C_j\|^2} \]

iii. Calculate Affinities
   Gaussian kernel
   
   \[ a.f.(C_i, C_j) = \exp\left(-\frac{\text{dist}(C_i, C_j)}{\sigma^2}\right) \]

iv. Markov normalization
   
   \[ \sum_{j=1}^{S} P_{i,j} = 1. \]
Diffusing Values to Impute
missing values

original data with dropout
imputation with MAGIC
imputation with MAGIC
imputation with MAGIC
Before MAGIC
After MAGIC
Data before MAGIC
Data before MAGIC

Data after MAGIC

Fit polytope

Infer archetypes
Complex Experimental Designs: Multiple Conditions

Cells + Perturbation → Cells
Complex Experimental Designs: Multiple Conditions

Cells + Perturbation → Cells

Gene Expression Matrix

Genes → Cond. → PC2

PC1
Complex Experimental Designs: Multiple Conditions

Problem: Effect size is small relative to biological and technical noise
MELD
(in preparation)

David van Dijk
Jay Stanley
Dan Burkhardt
MELD
(in preparation)

Main idea: smooth an external, experimental label on the data graph to impute continuity and infer associations
Experimental Label Smoothing

- Genes
- Condition Labels
- Interpolated labels

- Cell from Condition 1
- Cell from Condition 2
Enhanced ability to correlate or regress variables with experimental condition

Gene A  
Condition  |  MELD dimension
---|---
1  | 1
1.5  | 2

Gene B  
Condition  |  MELD dimension
---|---
1  | 1.5
2  | 2

Gene C  
Condition  |  MELD dimension
---|---
1  | 1.5
2  | 2

Ability to identify emerging cell types

Sample labels  

MELD scores  

PHATE 2  

PHATE 1  

PHATE 2
Graph Signal Processing

1 = \lambda_0 \geq \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_\delta > 0

\{ \phi_0, \phi_1, \phi_2, \ldots, \phi_\delta \}

x \mapsto \Phi(x) \triangleq [\lambda_0 \phi_0(x), \lambda_1 \phi_1(x), \lambda_2 \phi_2(x), \ldots, \lambda_\delta \phi_\delta(x)]^T
Lyme Disease Analysis (HIPC)

T cells gated by CD3+
Cell Type Changes Over Time

- Memory T cells up-regulated in v2
- Tregs appear in v2

Cluster proportion changes over MELD time.
Enhanced ability to correlate or regress variables with experimental condition

Ability to identify emerging cell types
PHATE
(Moon, van Dijk, et al. 2017)
PHATE
(Moon, van Dijk, et al. 2017)

Main idea: a metric-preserving dimensionality reduction algorithm that naturally emphasizes trajectory structure
Artificial Tree

Embryoid Bodies

Embryoid Body
- Ectoderm
- Mesoderm
- Endoderm

Progenitors
- Cardiac
- Hematopoietic
- Neural

Stem cells

Colony in suspension
Visualization Methods

- Artificial Tree
- PCA
- tSNE
- PHATE
A. Data

B. Distances

\[ \text{dist}_i = \sqrt{\sum \text{Cell}_i \cdot \text{Cell}_j}^2 \]

C. Affinities

D. Diffusion Probabilities

E. Informational Distance

F. PHATE

Embed information distance with MDS

Distance between probability distributions, e.g., potential distance

\[ \text{Normalized Affinities}^+ \]
Preserving Information Geometry

Statistical manifold

Each point is a probability distribution

An information theoretic distance (divergence) is preserved between points
Metric MDS instead of Eigendecomposition

Preserves variance in two dimensions using gradient descent or SMACKOFF
PHATE preserves local as well as global structure
SAUCIE
Sparse Autoencoder for Clustering Imputation and Embedding
(Amodio, van Dijk, Srinivasan et al. 2017)
SAUCIE
Sparse Autoencoder for Clustering Imputation and Embedding
(Amodio, van Dijk, Srinivasan et al. 2017)

Main idea: Let a neural network find emergent patterns in the data
One-stop scalable multi-sample data analysis
[Hinton, Salakhutdinov, Science 2006]
Representation for Clustering
Representation for Clustering

Embedding layer by cluster
Representation for Clustering

Embedding layer by cluster
Representation for Clustering

We want: codes for subspaces
Binary activation, good spacing
Clustering: Information Dimension Regularization

High activation entropy

Low activation entropy

Penalty:

\[-\sum a_i \log(a_i)\]

\[a_i = \frac{A_i}{\sum A_i}\]
Information Dimension (ID) Regularization

ID regularization enforces a binary-like code
Batch Effects

Patient 1

Patient 2
Maximal Mean Discrepancy

$$MMD(p, q) = \frac{1}{m^2} \sum_{i,j \in m} K(p_i, p_j) - \frac{2}{mn} \sum_{i,j} K(p_i, q_j) + \frac{1}{n^2} \sum_{i,j \in n} K(q_i, q_j)$$

[Dziugaite, Roy, Ghahramani, UAI 2015]
Now organized by Cell Type Rather than Patient
Cellular Manifolds of Dengue Patients
Cell Clusters: 180 Samples Combined
Patient Cluster Signatures

MMD distance can be derived between patient cluster proportions and embedded
Patient manifolds show coherent organization
**Postdocs:**
David van Dijk
Ryan Sowell

**Graduate Students**
Daniel Burkhardt
Jay Stanley
Matthew Amodio
Scott Gigante
Alexander Tong

**Medical Students**
William Chen
Mannik Kuchroo

**Undergraduate Students:**
Daniel Dagger
Ngan Vu

**Associate Researcher:**
Guy Wolf

**Prof on Sabbatical:**
Marc Hammarlund
Collaborations

Noonan Lab (Yale)
Kristina Yim
Rebecca Muhle
Martina Krenzer
Jim Noonan

Kevan Herold Lab
Kevan Herold
Ana Perdigito

Ruth Montgomery Lab
Ruth Mongomery
Yujiao (Helen) Zhao

Aaron Ring Lab
Diane Krause lab
Bluma Lesch Lab

Ivanova Lab (Yale)
Zheng Wang
Natalia Ivanova

Giraldez Lab (Yale)
Tonnie van den Elzen
Valerie Tornini

YCGA
Guilin Wang
Srikanth Mane
Kaya Bilguvar
Chris Castaldo
www.krishnaswamylab.org

- All Software on GITHUB: https://github.com/KrishnaswamyLab/
- All papers on bioarchive
- MAGIC in Cell
- Smita.krishnaswamy@yale.edu
Scaling PHATE Up

Original Points Original Points

Landmarks Landmarks

\[ P^m = (OL)^m = O(LO)^mL \]

Runtime Comparison - Subsampled EB Data

Runtime (sec)

Number of samples