Geometry-based Data Exploration with Manifold Learning & Diffusion Geometry

Guy Wolf

Université de Montréal
Mila – Quebec AI Institute

guy.wolf@umontreal.ca

2020
Exploratory data analysis

One of the main challenges in modern data science is that:

- **Big high-dimensional data** are being produced everywhere
- **Limited numbers of domain scientists** have to process such data into useful knowledge

This challenge requires **exploratory data analysis** to produce human-interpretable data representations by

1. **Inferring structure** from collected data
2. Using this structure to **process data features** to become accessible for analysis

New frontier for data science & machine learning, beyond traditional predictive & generative tasks.
Exploratory data analysis

Example (high throughput single cell technologies)

**scRNA-seq:** cells × genes

**CyTOF:** cells × proteins

- Big volumes of data
- High dimensional feature space
- Nontrivial noise & collection artifacts
- Multiresolution structures & processes
- Exploration often targets sparse data regions
Descriptive exploration in genomics & proteomics

Single-cell data:
- Gene counts
- Protein counts

Clusters
- Vascular muscle cells
- Endothelium

Transitions
- Stem Cell
- Blastocysts
- Endothelial-Myeloid Progenitors
- Endothelium
- Muscle Precursors
- Vascular muscle cells

Guy Wolf (UdeM)
Visualizing progression & transitions in data

Progression & Transition Structures

High-dimensional Measurements

How can we reveal progression & transitions in data?
Visualizing progression & transitions in data

Progression & Transition Structures

How can we reveal progression & transitions in data?

High-dimensional Measurements

PHATE 1

PHATE 2
Question: is cellular development really a high-dimensional process?

Consider the following key properties:

1. Cells develop progressively via **small incremental steps** (e.g., differentiation and mutation)
2. Variations in each step have **limited degrees of freedom**

Conclusion: this progression can be modeled as a collection of **smoothly varying, locally low-dimensional, data patches**.

Such models are similar to the mathematical formulation of a manifold, and can be **inferred by manifold learning methods**.

Local affinities $g(x, y) \Rightarrow \text{transition probs. } \Pr[x \leadsto y] = \frac{g(x, y)}{||g(x, \cdot)||_1}$
Local affinities $g(x, y) \Rightarrow$ transition probs. $Pr[x \sim y] = \frac{g(x, y)}{\|g(x, \cdot)\|_1}$

Markov chain/process $\Rightarrow$ random walks on data manifold
Diffusion geometry
Random walks reveal intrinsic neighborhoods

\[ p^t(x, y) = \Pr[x^{\text{steps}} \rightarrow y] \]
Are geodesic distances sufficient for faithful intrinsic embedding?
Diffusion geometry
Geodesic vs. diffusion distances

Are geodesic distances sufficient for faithful intrinsic embedding?
Are geodesic distances sufficient for faithful intrinsic embedding?

Diffusion-based notions enable robust intrinsic data geometry.
Are geodesic distances sufficient for faithful intrinsic embedding?

Diffusion-based notions enable robust intrinsic data geometry.
Are geodesic distances sufficient for faithful intrinsic embedding?

Diffusion-based notions enable robust intrinsic data geometry.
Diffusion geometry

Diffusion & potential distances

DM (Coifman & Lafon):
\[ \| \Phi^t(x) - \Phi^t(y) \| \approx \left\| P^t_{(x, \cdot)} - P^t_{(y, \cdot)} \right\|_{L^2(\|q\|_1/q)} \]

PHATE (Moon et al.):
\[ \| \Phi^t(x) - \Phi^t(y) \| \approx \| \log P^t_{(x, \cdot)} - \log P^t_{(y, \cdot)} \| \]

\[ g(u, v) = \text{local affinity} \]
\[ q(u) = \| g(u, \cdot) \|_1 \]
\[ P_{(u, v)} = g(u, v) / q(u) \]
\[ P^t_{(u, v)} = \Pr[u \overset{t \text{ steps}}{\sim} v] \]

\[ \Phi^t : \text{data} \rightarrow \mathbb{R}^d \text{ (small } d) \]
Data visualization

PHATE (Moon et al., Nat. Biotech. 2019)

Overview

A. Data

B. Distances

C. Affinities

D. Diffusion Probabilities

E. Informational Distance

F. PHATE

Embed information distance with MDS

dist_{ij} = \| \text{Cell}_i - \text{Cell}_j \|_2

\text{Kernel}

\text{Dist}

\text{Aff}_{ij}

P' = [\text{Normalized Affinities}]^t
Affinities via adaptive $\alpha$-decaying kernel

$$\tilde{g}(x, y) = \exp\left[-\left(\frac{\|x - y\|}{\varepsilon_x}\right)^\alpha\right] \quad \Rightarrow \quad g(x, y) = \frac{\tilde{g}(x, y) + \tilde{g}(y, x)}{2}$$

Where

- $\varepsilon_x = \text{distance from } x \text{ to its } k\text{-th nearest neighbor}$
- $\alpha$ controls the decay rate of $\tilde{g}$ w.r.t $\frac{\|x - y\|}{\varepsilon_x}$
Diffusion time tuning with spectral entropy

Spectral entropy at time $t$:

$$- \sum_j \frac{\lambda_j^t}{\|\lambda^t\|_1} \log \frac{\lambda_j^t}{\|\lambda^t\|_1}$$

where $\lambda^t = \{\lambda_0^t, \lambda_1^t, \lambda_2^t, \ldots\}$ are the eigenvalues of $P^t$. 

Rapid decline corresponds to denoising

Flatter region corresponds to losing real signal
Information-geometry distance

\[ \text{dist}_{ij} = \| \Delta^{(+1)}_{(x_i, y_j)} \|_2 \text{ where} \]

\[ \Delta^{(\gamma)}_{(x, y)}(z) = -\int p_t^y(z) u - \frac{\gamma + 1}{2} du \]

\[ = \begin{cases} 
  p_t^x(z) - p_t^y(z) & \gamma = -1 \\
  \log p_t^x(z) - \log p_t^y(z) & \gamma = +1 \\
  \frac{2}{1-\gamma} \left[ \left( p_t^x(z) \right)^{ \frac{1-\gamma}{2} } - \left( p_t^y(z) \right)^{ \frac{1-\gamma}{2} } \right] & \text{otherwise}
\end{cases} \]
Data visualization

PHATE (Moon et al., Nat. Biotech. 2019)

Overview

A. Data

\[
\text{Cell}_i = \begin{bmatrix}
\text{Gene}_1 \\
\text{Gene}_2 \\
\vdots \\
\text{Gene}_n
\end{bmatrix}
\]

Molecular counts

B. Distances

\[
dist_{ij} = \|\text{Cell}_i - \text{Cell}_j\|_2
\]

C. Affinities

\[
dist_{ij} \rightarrow \text{Kernel} \rightarrow aff_{ij}
\]

D. Diffusion Probabilities

\[
P^\alpha = [\text{Normalized Affinities}]^\alpha
\]

E. Informational Distance

\[
dist_{ij} = \|\log P_i^\alpha - \log P_j^\alpha\|_2
\]

F. PHATE

Embed information distance with MDS

Guy Wolf (UdeM)  Geometry-based Data Exploration  2020 10 / 29
Example #1: artificial tree

- 40 dimensions, dense regions at branch- and end-points

PCA:
Data visualization

PHATE (Moon et al., Nat. Biotech. 2019)

Example #1: artificial tree

- 40 dimensions, dense regions at branch- and end-points

`tSNE:`
Data visualization

Example #1: artificial tree

- 40 dimensions, dense regions at branch- and end-points

PHATE (Moon et al., Nat. Biotech. 2019)
Data visualization

PHATE (Moon et al., Nat. Biotech. 2019)

Example #2: exploring differentiation trajectories in Embryoid Bodies

- New single-cell RNA-sequencing measured over 27-day timecourse
Data visualization

PHATE (Moon et al., Nat. Biotech. 2019)

Example #2: exploring differentiation trajectories in Embryoid Bodies

New single-cell RNA-sequencing measured over 27-day timecourse
Data visualization

PHATE (Moon et al., Nat. Biotech. 2019)

Example #2: exploring differentiation trajectories in Embryoid Bodies

- New single-cell RNA-sequencing measured over 27-day timecourse
Data visualization

PHATE (Moon et al., Nat. Biotech. 2019)

Example #2: exploring differentiation trajectories in Embryoid Bodies

New single-cell RNA-sequencing measured over 27-day timecourse
Data visualization

PHATE (Moon et al., Nat. Biotech. 2019)

Example #2: exploring differentiation trajectories in Embryoid Bodies

Time Samples

Clusters

Branches

Geometry-based Data Exploration

Guy Wolf (UdeM)
Data visualization

**PHATE** (Moon et al., Nat. Biotech. 2019)

Evaluation & comparisons

### A

- Splatter Groups
- Splatter Paths
- Noiseless simulation
- Geodesic distances
- Euclidean distances
- 2D embeddings
- PCA, PHATE, t-SNE, etc.

### B

- Dropout
- Biological Variation
- Subsampled cells
- Subsampled genes

- Spearman correlation
  - Noiseless simulation
  - Noisy simulation
  - Geodesic distances
  - Euclidean distances
  - 2D embeddings
  - PCA, PHATE, t-SNE, etc.

- Splatter Paths
- Splatter Groups

- Dropout probability
- Coefficient of Variation
- # of cells retained
- # of genes retained

Guy Wolf (UdeM)

Geometry-based Data Exploration

2020
Drug data exploration

PhEMD (Chen et al., Nat. Meth. 2020)

Embedding cell distributions with EMD-based diffusion maps (Coifman & Lafon, 2006)

Chronic treatment for 5 days

- **No treatment**
- **+TGF-β**
- **+TGF-β +Inhibitor**

60-well multiplexing

- **Negative control**
- **Positive control**
- **Perturbed condition**

Barcoding

Sample pooling

Antibody staining

**Antibody panel**

<table>
<thead>
<tr>
<th>Phenotypic Markers</th>
<th>Cell Metabolism</th>
<th>Survive</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-cadherin</td>
<td>p-AMPK</td>
<td>Cleaved Caspase 3</td>
</tr>
<tr>
<td>β-catenin</td>
<td>TGF-β/BMP pathway</td>
<td>c-myc</td>
</tr>
<tr>
<td>CD24</td>
<td>p-Smad2/p-Smad3</td>
<td>Survivin</td>
</tr>
<tr>
<td>Vimentin</td>
<td>p-Smad1/5</td>
<td>MAPK pathway</td>
</tr>
<tr>
<td>CD44</td>
<td>EMT TF</td>
<td>p-MEK1/2</td>
</tr>
<tr>
<td>Cell Cycle</td>
<td>Snail</td>
<td>p-Creb</td>
</tr>
<tr>
<td>Cyclin B1</td>
<td>Nanog</td>
<td>p-P38</td>
</tr>
<tr>
<td>p-Rb</td>
<td>AKT pathway</td>
<td>RTK</td>
</tr>
<tr>
<td>p-H3</td>
<td>p-AKT</td>
<td>p-PLCy2</td>
</tr>
<tr>
<td>p-MARCKS</td>
<td>p-S6</td>
<td>p-SHP2</td>
</tr>
<tr>
<td>JAK/STAT pathway</td>
<td>p-GSK3β</td>
<td>p-Tie2</td>
</tr>
<tr>
<td>p-STAT3</td>
<td>NFκB signaling</td>
<td>Integrin associated</td>
</tr>
<tr>
<td>p-STAT5</td>
<td>p-NFκB p65</td>
<td>p-FAK</td>
</tr>
</tbody>
</table>

Mass cytometry

Cell ionization-ICP TOF-MS

Cell 1

Cell 2
Drug data exploration

PhEMD (Chen et al., Nat. Meth. 2020)

Embedding cell distributions with EMD-based diffusion maps (Coifman & Lafon, 2006)

Merge all conditions

Sample 1 ($S_1$) + Sample 2 ($S_2$) = All samples

Compute sample distance matrix

$S_1$ $S_2$ $\cdots$ $S_n$

$S_1$ $\sim$ $S_2$ $\cdots$ $\sim$ $S_n$

0.54

Biological sample embedding

Identify cell subtypes

Cell subtypes

Global cell geometry

Determine the composition of each sample

Sample 1

Sample 2

Perform community detection

Guy Wolf (UdeM)
Earth-Mover’s Distances (EMD) between samples quantify the intrinsic difference in cell distribution over the data manifold.

Diffusion maps embedding of samples:

1. Pairwise EMD $\rightarrow$ sample neighborhoods $\rightarrow$ sample-wise diffusion
2. Eigendecomposition of $P^t$ $\rightarrow$ diffusion coordinates of samples
Drug data exploration  PhEHD (Chen et al., Nat. Meth. 2020)
Embedding cell distributions with EMD-based diffusion maps (Coifman & Lafon, 2006)

Earth-Mover's Distances (EMD) between samples quantify the intrinsic difference in cell distribution over the data manifold.

1. Pairwise EMD → sample neighborhoods → sample-wise diffusion
2. Eigendecomposition of $P^t$ → diffusion coordinates of samples

Diffusion coordinates

Eigenvals of $P$:

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

Eigenvecs of $P$:

$$\# \text{ samples} \begin{cases} \phi_0 & \phi_1 & \phi_2 & \ldots & \phi_\delta \\ 1 & 2 & 3 & \ldots & \delta \end{cases}$$

Diffusion Map at time $t$:

$$S \mapsto \Phi^t(S) \triangleq [\lambda_1^t \phi_1(S), \lambda_2^t \phi_2(S), \ldots, \lambda_\delta^t \phi_\delta(x)]^T$$
Drug data exploration  PhEMD (Chen et al., Nat. Meth. 2020)
Embedding cell distributions with EMD-based diffusion maps (Coifman & Lafon, 2006)
Generate random points:

Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

Walk toward the data manifold from randomly generated points
Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

Walk toward the data manifold from randomly generated points

Generate random points:

Walk towards the data manifold with diffusion: \( x \mapsto \sum_{y \in \text{data}} y \cdot p^t(x, y) \)
Generative models typically infer distribution from collected data, and sample it to generate more data.

- Biased by sampling density
- May miss rare populations
- Does not preserve the geometry
Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

New approach: geometry based data generation
Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

New approach: geometry based data generation
Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

New approach: geometry based data generation
Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

New approach: geometry based data generation
Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

New approach: geometry based data generation
Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

Correct density with MGC kernel (Bermanis et al., ACHA 2016)

Separate density/geometry with new kernel: 
\[ k(x,y) = \sum_{r \in \text{data}} \frac{g(x,r)g(y,r)}{\|g(r,\cdot)\|_1} \]

Use new diffusion process 
\[ p(x, y) = \frac{k(x,y)}{\|k(x,\cdot)\|_1} \] to walk to the manifold
Separate density/geometry with new kernel: 

\[ k(x, y) = \sum_{r \in \text{data}} \frac{g(x, r) g(y, r)}{\|g(r, \cdot)\|_1} \]

Use new diffusion process 

\[ p(x, y) = \frac{k(x, y)}{\|k(x, \cdot)\|_1} \]

to walk to the manifold
**Question:** How should we initialize new points to end up with uniform sampling from the data manifold?

**Answer:** For each \( x \in \text{data} \), initialize \( \hat{\ell}(x) \) points sampled from \( \mathcal{N}(x, \Sigma_x) \); set \( \hat{\ell} \) as the mid-point between the upper & lower bounds in the following proposition.

**Proposition**

The generation level \( \hat{\ell}(x) \) required to equalize density is bounded by

\[
\begin{align*}
\det \left( I + \frac{\Sigma_x}{2\sigma^2} \right)^{1/2} \frac{\max(\hat{d}(\cdot)) - \hat{d}(x)}{\hat{d}(x) + 1} - 1 &\leq \hat{\ell}(x) \leq \det \left( I + \frac{\Sigma_x}{2\sigma^2} \right)^{1/2} \left[ \max(\hat{d}(\cdot)) - \hat{d}(x) \right],
\end{align*}
\]

where \( \sigma \) is a scale used when defining Gaussian neighborhoods \( g(x, y) \) for the diffusion geometry, and \( \hat{d}(x) = \| g(x, \cdot) \|_1 \) estimates local density.
Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

Illuminate hypothetical cell types in single-cell data from Velten et al. (2017)

Recovering originally-undersampled lineage in early hematopoeisis:

B-cell maturation trajectory enhanced by SUGAR

SUGAR equalizes the total cell distribution
Data generation

SUGAR (Lindenbaum et al., NeurIPS 2018)

Recover gene-gene relationships in single-cell data from Velten et al. (2017)

SUGAR improves module correlation and MI identified by Velten et al.

![Graph showing the improvement of correlations and MI after SUGAR application.](image)

Velten et al., Nature Cell Biology 19 (2017)
Generated cells also follow canonical marker correlations

\[
\begin{align*}
\text{Original} & \quad \text{SUGAR} \\
\text{HOXA3} & \quad R = -0.323, \quad \text{MI} = 1.79 \\
& \quad R = -0.806, \quad \text{MI} = 2.78 \\
\text{CASP1} & \quad R = 0.364, \quad \text{MI} = 1.04 \\
& \quad R = 0.899, \quad \text{MI} = 2.62 \\
\text{EAF2} & \quad R = -0.054, \quad \text{MI} = 1.37 \\
& \quad R = -0.282, \quad \text{MI} = 2.08
\end{align*}
\]

Li et al., Nature communications 7 (2016)
Imputation & denoising

MAGIC (van Dijk et al., Cell 2018)

Overview

Gene Interactions

Archetypal Analysis

Population Analysis

Transcription Factor—Target Prediction

Zeb1  Snai1  Snai2  Twist1

Target1  Target2  Target3  Target4  Target5  Target6  Target7  Target8
Imputation & denoising

MAGIC (van Dijk et al., Cell 2018)

Recovering gene interactions in EMT data

Geometry-based Data Exploration

Guy Wolf (UdeM)
Imputation & denoising

MAGIC (van Dijk et al., Cell 2018)

Recovering gene interactions in EMT data

E-Cadherin vs Vimentin at t = 1

Snail vs Vimentin at t = 1
Understanding diffusion geometry
Harmonic analysis on data manifold / foundations of graph signal processing

The diffusion operator $P^t \rightarrow$ heat kernel $e^{t\Delta}$ when # data points $\rightarrow \infty$, neighborhood radius $\rightarrow 0$, up to density normalization.

- The eigenvectors / eigenfunctions of $P^t$ form generalized Fourier harmonics over the data geometry
- The eigenvalues of $P^t$ take the form of $e^{-t \cdot (\text{frequency})^2}$
- $f(x) \mapsto P^t f(x)$ acts as a lowpass filter

Harmonic analysis interpretation of presented methods:
- SUGAR / MAGIC – based on lowpass filtering of data features
- PHATE / DM – based on impulse responses of lowpass filters

Beyond lowpass - diffusion filters over intrinsic data geometry:
- $f(x) \mapsto (I - P^t)f(x)$ – highpass filter
- $f(x) \mapsto P^t(I - P^t)f(x)$ – bandpass filter / diffusion wavelet
Geometric scattering

(Gao et al., ICML 2019)

Deep diffusion-based filter bank for graph / manifold representation

- Provides whole-graph representation for graph data analysis
- Mathematical framework for geometric deep learning
  - Analogous to Euclidean scattering by Mallat (CPAM, 2012)
- New notion of deformation stability using rigid motions & distribution variations on manifolds (Perlmutter et al., NeurIPS DLT workshop 2018)
Geometric scattering

Feature extraction for graph data analysis

Scattering features embed graphs with signals over their vertices to a Euclidean feature space indexed by scattering paths (i.e., $j, j', q$)

$G = (V, E, W)$

$f : V \rightarrow \mathbb{R}$

Adjacency matrix: $A(v_i, v_j)$

Signal vector: $f(v_i)$

Diffusion wavelets:

$$\psi_j = p^{2j-1} - p^{2j}$$

$P = \frac{1}{2}(I + D^{-1}A)$

Scattering $f \mapsto Sf$

Traditional Euclidean algorithms (e.g., SVM/PCA)

- Multiple signals handled by concatenation of scattering features
Inferring EC exchange preferences in enzyme evolution:

Exchanged pref. inference

Compute pref(EC-i, EC-j) :=

\[ w_j \cdot \left[ \min \left\{ \frac{D(i,j)}{D(i,i)}, \frac{D(j,i)}{D(j,j)} \right\} \right]^{-1} \]

\[ w_j = \text{portion of enzymes in EC-j that choose another EC as their nearest subspace; } D(i,j) = \text{mean dist. of enzymes in EC-i from PCA (90% exp. var.) subspace of EC-j.} \]

Observed by Cuesta et al. (Biophysical Journal, 2015)

Inferred via geometric scattering features

- Geometric scattering features extracted from ENZYMES (Borgwardt et al., Bioinformatics 2005) containing 100 enzyme graphs from each EC.
- PCA over scattering: EC subspaces of 5–7 dims.; full space of 16 dims.
Conclusion

Exploratory data analysis, especially in genomics/proteomics, often requires to separate data geometry from distribution.

Diffusion geometry enables a multitude of tools highly suitable for geometry-based analysis:

- PHATE - data visualization with diffusion geometry
- PhEMD - learning drug perturbation manifold
- SUGAR - geometry-based data generation
- Geometric scattering - graph/manifold-level representations

Additional work includes, for example:

- MAGIC - data imputation & denoising (van Dijk et al., 2018)
- Data fusion with harmonic alignment (Stanley et al., 2019)
Acknowledgements

Yale Applied Math Program:
Ofir Lindenbaum
Jay Stanley
Raphy Coifman

Yale School of Medicine:
Smita Krishnaswamy
Scott Gigante
Will Chen
Dan Burkhardt
David van Dijk
Zheng Wang
Natalia Ivanova

Utah State University:
Kevin Moon

Michigan State University:
Matt Hirn
Fong Gao
Mike Perlmutter

University of Zurich:
Nevena Zivanovic
Bernd Bodenmiller

Memorial Sloan Kettering Cancer Center:
Dana Pe’er
Roshan Sharma

Funding:

Guy Wolf (UdeM)