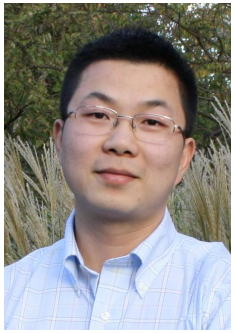
Carnegie Mellon University
Statistics & Data Science

High-dimensional statistical methods to model heterogeneity in genomic data

Kevin Lin

Advisors

Kathryn Roeder



Jing Lei

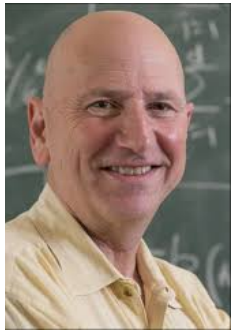
Committee Members

Max G'Sell

Han Liu

Anjali Mazumder

Ryan J. Tibshirani



Larry Wasserman

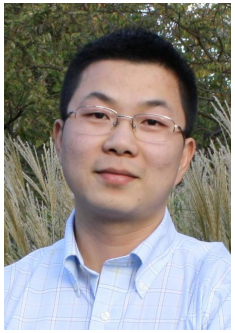
Carnegie Mellon University
Statistics & Data Science

High-dimensional statistical methods to model heterogeneity in genomic data

Kevin Lin

Advisors

Kathryn Roeder



Jing Lei

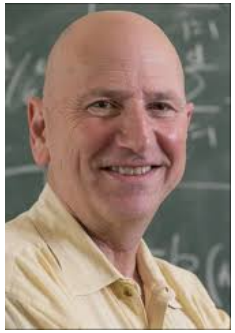
Committee Members

Max G'Sell

Han Liu

Anjali Mazumder

Ryan J. Tibshirani



Larry Wasserman

Biologists model the heterogeneity in genomic data in order to investigate different biological processes.

Chapter 2: Studying risk genes for autism, by finding a subset of homogeneous samples within a heterogeneous dataset via multiple covariance testing

(Lin, Liu, Roeder: JASA, 2020)

Biologists model the heterogeneity in genomic data in order to investigate different biological processes.

Chapter 2: Studying risk genes for autism, by finding a subset of homogeneous samples within a heterogeneous dataset via multiple covariance testing

(Lin, Liu, Roeder: JASA, 2020)

Chapter 3-4: Studying copy number variation, by proving properties of changepoint estimation and inference

(Lin, Sharpnack, Rinaldo, Tibshriani: NeurIPS, 2017)
(Hyun, Lin, G'Sell, Tibshriani: Biometrics, In revision)

Biologists model the heterogeneity in genomic data in order to investigate different biological processes.

Chapter 2: Studying risk genes for autism, by finding a subset of homogeneous samples within a heterogeneous dataset via multiple covariance testing

(Lin, Liu, Roeder: JASA, 2020)

Chapter 3-4: Studying copy number variation, by proving properties of changepoint estimation and inference

(Lin, Sharpnack, Rinaldo, Tibshriani: NeurIPS, 2017)
(Hyun, Lin, G'Sell, Tibshriani: Biometrics, In revision)

Chapter 5: Studying developmental trajectories of oligodendrocytes, by developing non-linear embedding based on exponential-family distributions

(Lin, Lei, Roeder: JASA, In revision)

Focus of this talk

Biologists model the heterogeneity in genomic data in order to investigate different biological processes.

Chapter 5: Studying developmental trajectories of oligodendrocytes, by developing non-linear embedding based on exponential-family distributions

(Lin, Lei, Roeder: JASA, In revision)

Focus of this talk

Biologists model the heterogeneity in genomic data in order to investigate different biological processes.

Chapter 5: Studying developmental trajectories of oligodendrocytes, by developing non-linear embedding based on exponential-family distributions

(Lin, Lei, Roeder: JASA, In revision)

Focus of this talk

Punch line (Statistics): eSVD performs a non-linear embedding by estimating a low-rank matrix of natural parameters via alternating minimization based on a random dot product model.

Biologists model the heterogeneity in genomic data in order to investigate different biological processes.

Chapter 5: Studying developmental trajectories of oligodendrocytes, by developing non-linear embedding based on exponential-family distributions

(Lin, Lei, Roeder: JASA, In revision)

Focus of this talk

Punch line (Statistics): eSVD performs a non-linear embedding by estimating a low-rank matrix of natural parameters via alternating minimization based on a random dot product model.

Biologists model the heterogeneity in genomic data in order to investigate different biological processes.

Chapter 5: Studying developmental trajectories of oligodendrocytes, by developing non-linear embedding based on exponential-family distributions

(Lin, Lei, Roeder: JASA, In revision)

Focus of this talk

Punch line (Statistics): eSVD performs a non-linear embedding by estimating a low-rank matrix of natural parameters via alternating minimization based on a random dot product model.

Biologists model the heterogeneity in genomic data in order to investigate different biological processes.

Chapter 5: Studying developmental trajectories of oligodendrocytes, by developing non-linear embedding based on exponential-family distributions

(Lin, Lei, Roeder: JASA, In revision)

Focus of this talk

Punch line (Statistics): eSVD performs a non-linear embedding by estimating a low-rank matrix of natural parameters via alternating minimization based on a random dot product model.

Biologists model the heterogeneity in genomic data in order to investigate different biological processes.

Chapter 5: Studying developmental trajectories of oligodendrocytes, by developing non-linear embedding based on exponential-family distributions

(Lin, Lei, Roeder: JASA, In revision)

Focus of this talk

Punch line (Statistics): eSVD performs a non-linear embedding by estimating a low-rank matrix of natural parameters via alternating minimization based on a random dot product model.

Oligodendrocytes are cells in the nervous system that produce myelin, and understanding their development can lead to better treatment for diseases such as MS.

- Oligodendrocytes
- Developmental trajectory
- Single-cell RNA-seq data
- Analysis pipeline

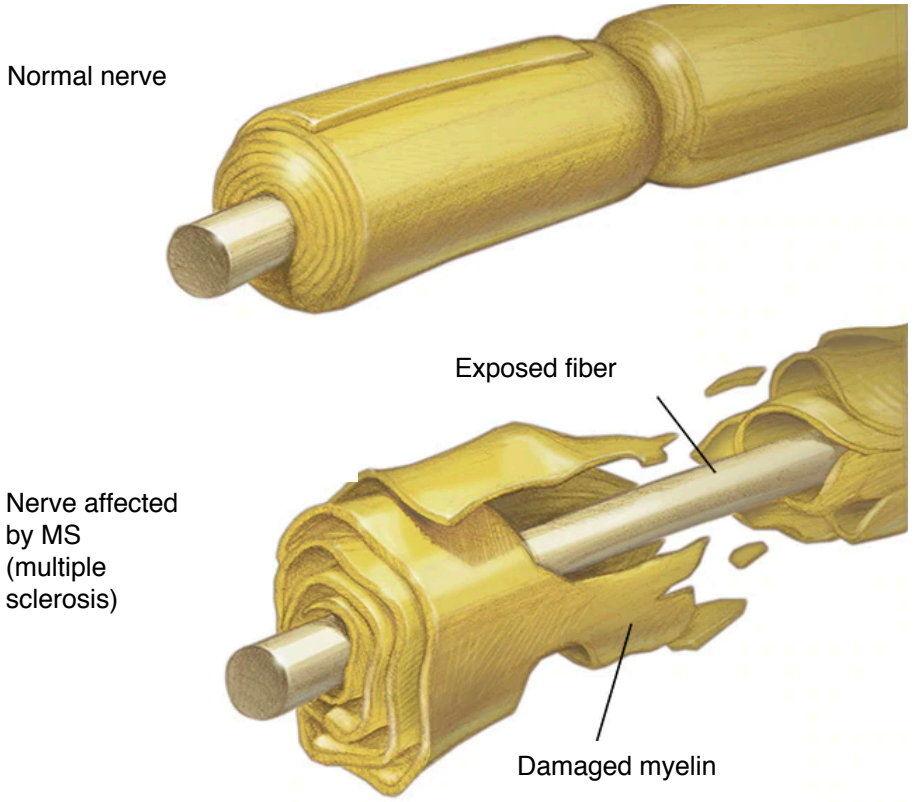
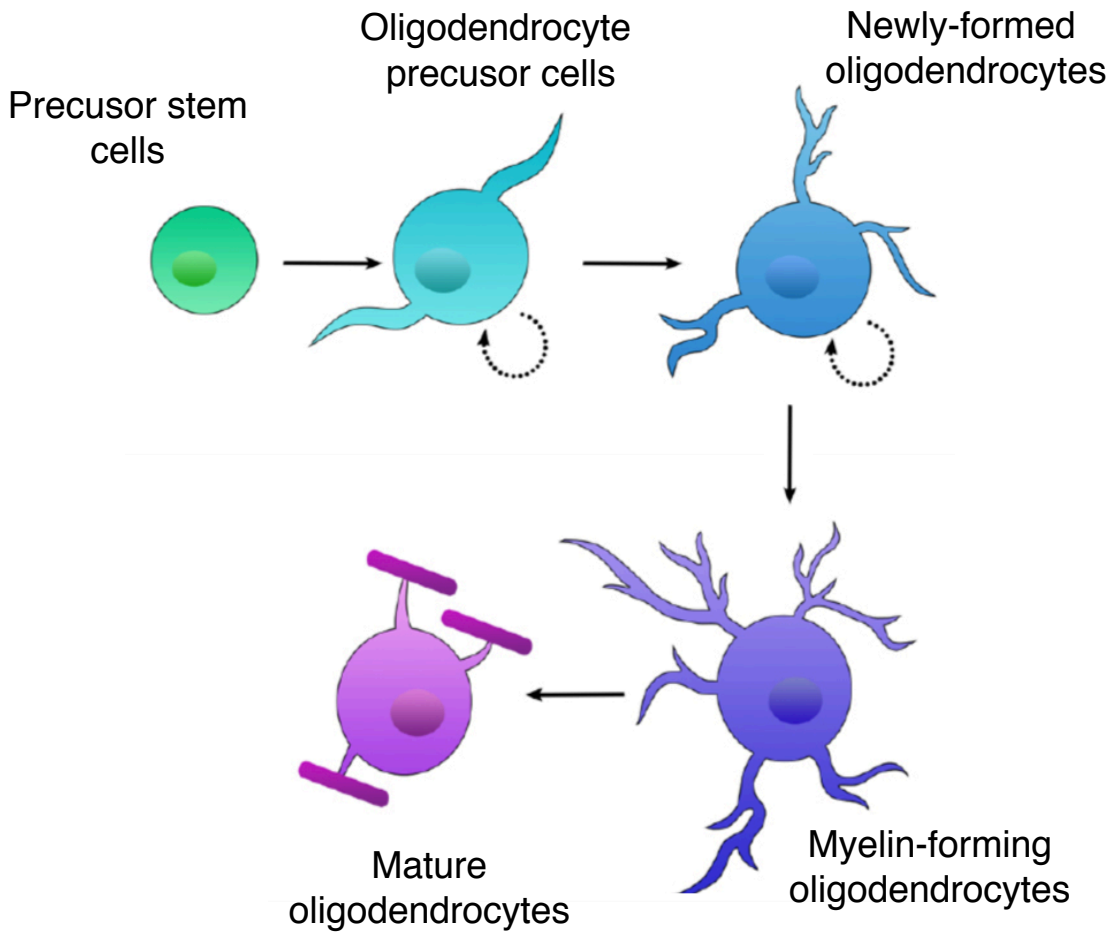


Illustration from Mayo clinic

Oligodendrocytes are cells in the nervous system that produce myelin, and understanding their development can lead to better treatment for diseases such as MS.

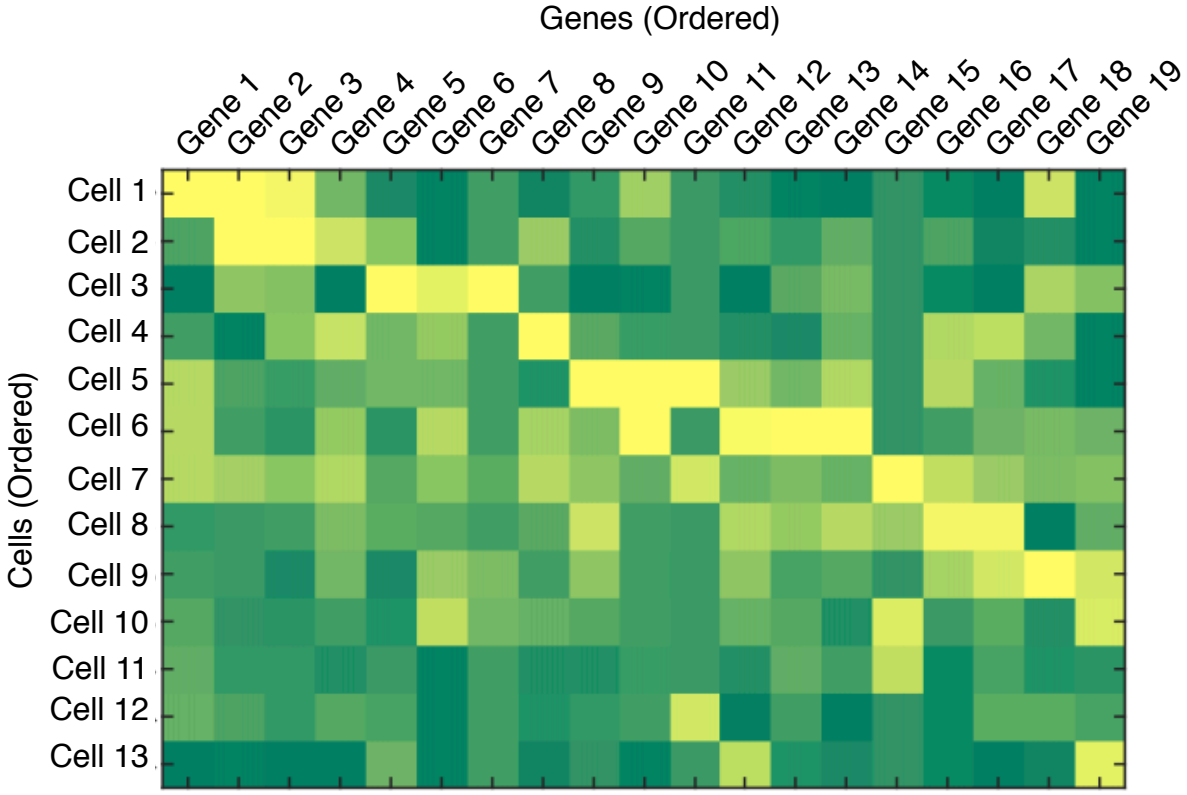
- Oligodendrocytes
- Developmental trajectory
- Single-cell RNA-seq data
- Analysis pipeline



Cartoon from Newville et. al (2017)

Oligodendrocytes are cells in the nervous system that produce myelin, and understanding their development can lead to better treatment for diseases such as MS.

- Oligodendrocytes
- Developmental trajectory
- Single-cell RNA-seq data
- Analysis pipeline



Altered figure from Marques et. al (2016)

Oligodendrocytes are cells in the nervous system that produce myelin, and understanding their development can lead to better treatment for diseases such as MS.

Preprocessing → Embedding → Known cell types + Trajectory estimation

- Oligodendrocytes
- Developmental trajectory
- Single-cell RNA-seq data
- Analysis pipeline

Oligodendrocytes are cells in the nervous system that produce myelin, and understanding their development can lead to better treatment for diseases such as MS.

Preprocessing → Embedding → Known cell types + Trajectory estimation

- Oligodendrocytes
- Developmental trajectory
- Single-cell RNA-seq data
- Analysis pipeline

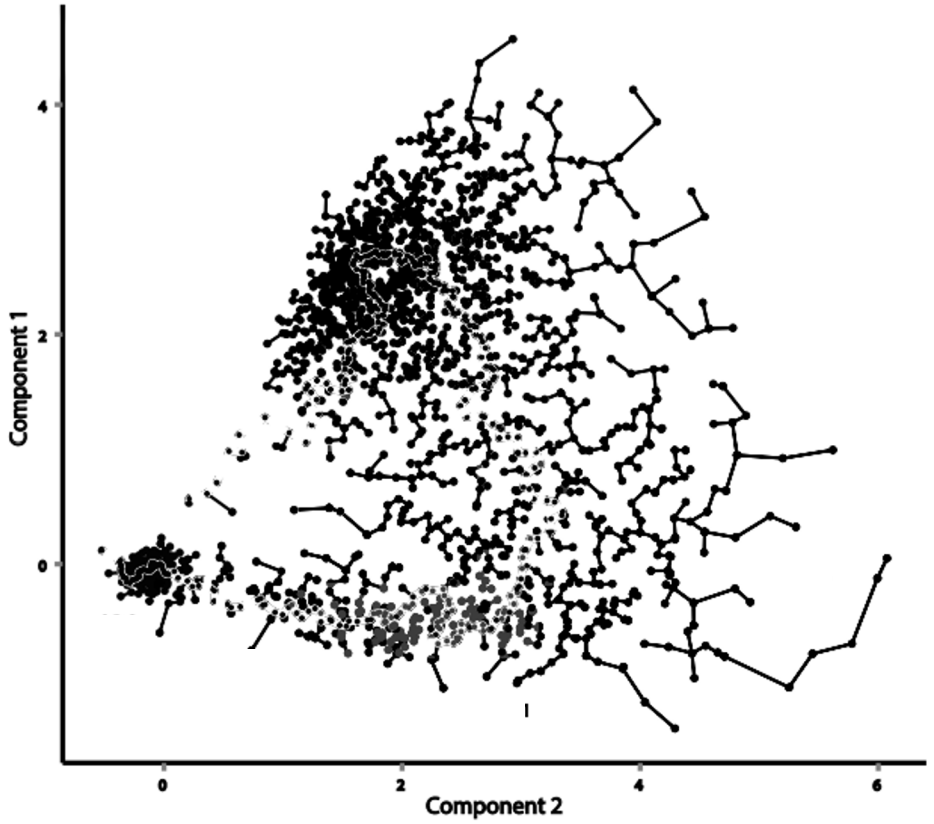


Figure from Marques et. al (2016)

Oligodendrocytes are cells in the nervous system that produce myelin, and understanding their development can lead to better treatment for diseases such as MS.

Preprocessing → Embedding → Known cell types + Trajectory estimation

- Oligodendrocytes
- Developmental trajectory
- Single-cell RNA-seq data
- Analysis pipeline

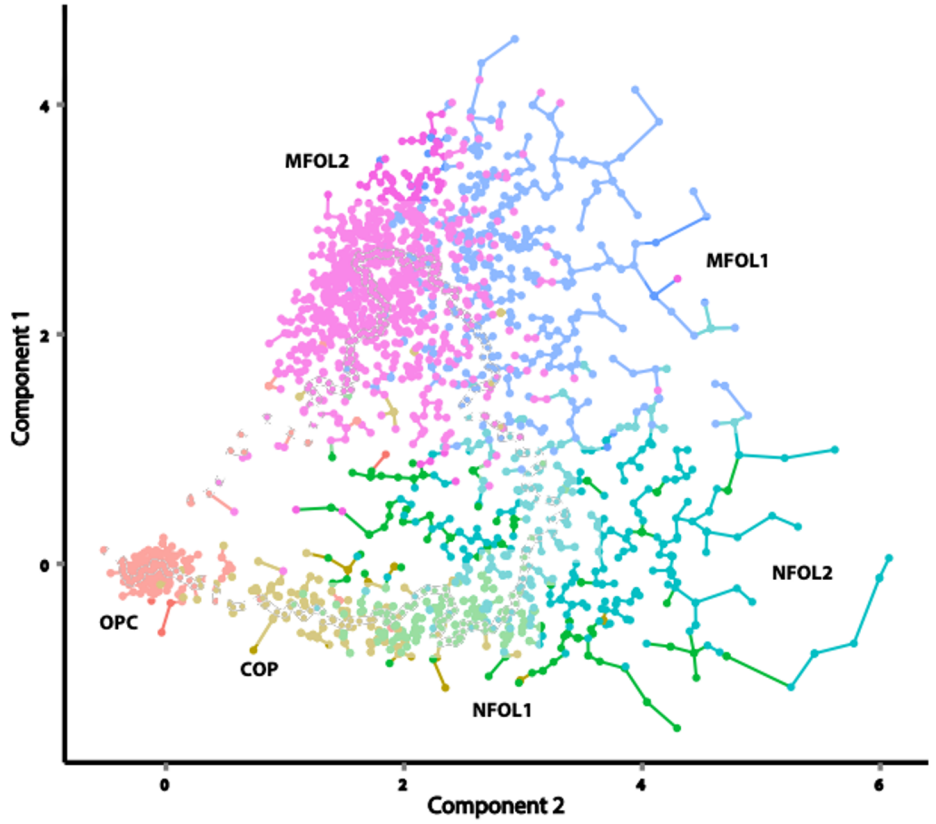


Figure from Marques et. al (2016)

Oligodendrocytes are cells in the nervous system that produce myelin, and understanding their development can lead to better treatment for diseases such as MS.

Preprocessing → Embedding → Known cell types + Trajectory estimation

- Oligodendrocytes
- Developmental trajectory
- Single-cell RNA-seq data
- Analysis pipeline

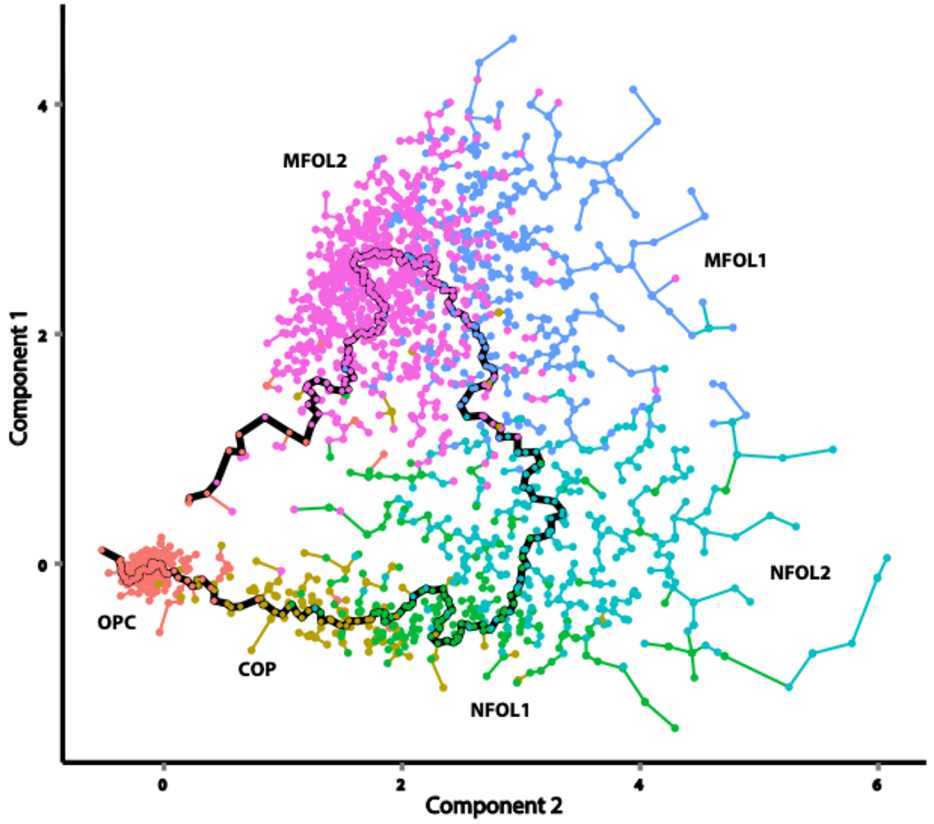


Figure from Marques et. al (2016)

Oligodendrocytes are cells in the nervous system that produce myelin, and understanding their development can lead to better treatment for diseases such as MS.

- Oligodendrocytes
- Developmental trajectory
- Single-cell RNA-seq data
- Analysis pipeline

Preprocessing → Embedding → Known cell types + Trajectory estimation

Focus of our paper

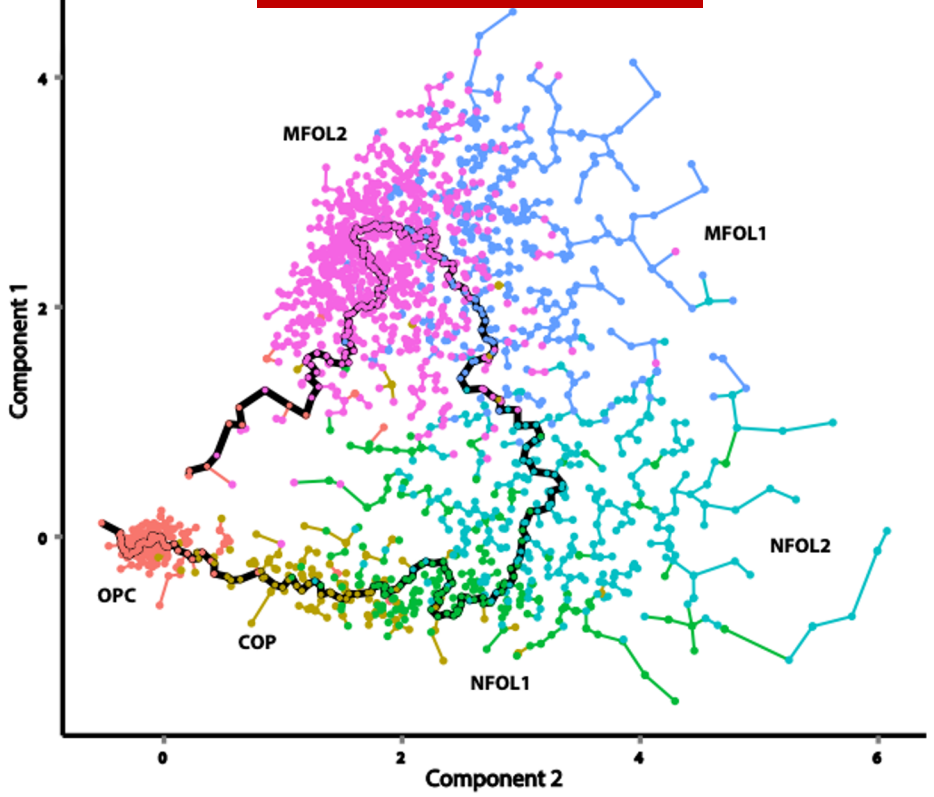


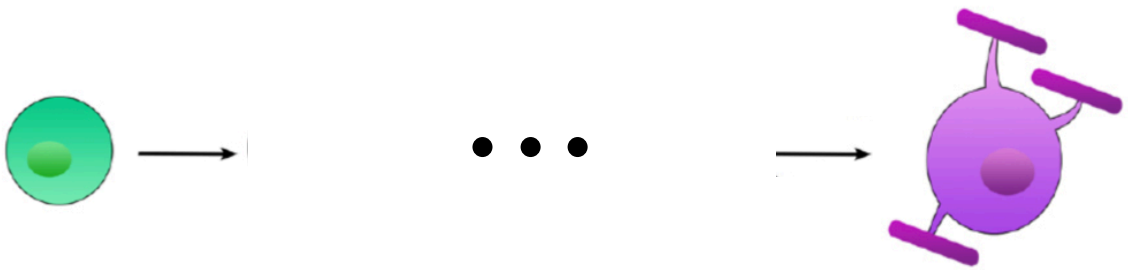
Figure from Marques et. al (2016)

The most common embedding, based on the SVD, is not desirable to model our single-cell RNA-seq dataset of oligodendrocytes from Marques et. al (2016).

$\boldsymbol{A} \in \mathbb{R}^{n \times p}$: n cells and p genes

Preprocessed single-cell RNA-seq data

$n = 5000+$ cells over 6 cell types, $p = 900+$ genes



Precusor stem cells

Mature oligodendrocytes

The most common embedding, based on the SVD, is not desirable to model our single-cell RNA-seq dataset of oligodendrocytes from Marques et. al (2016).

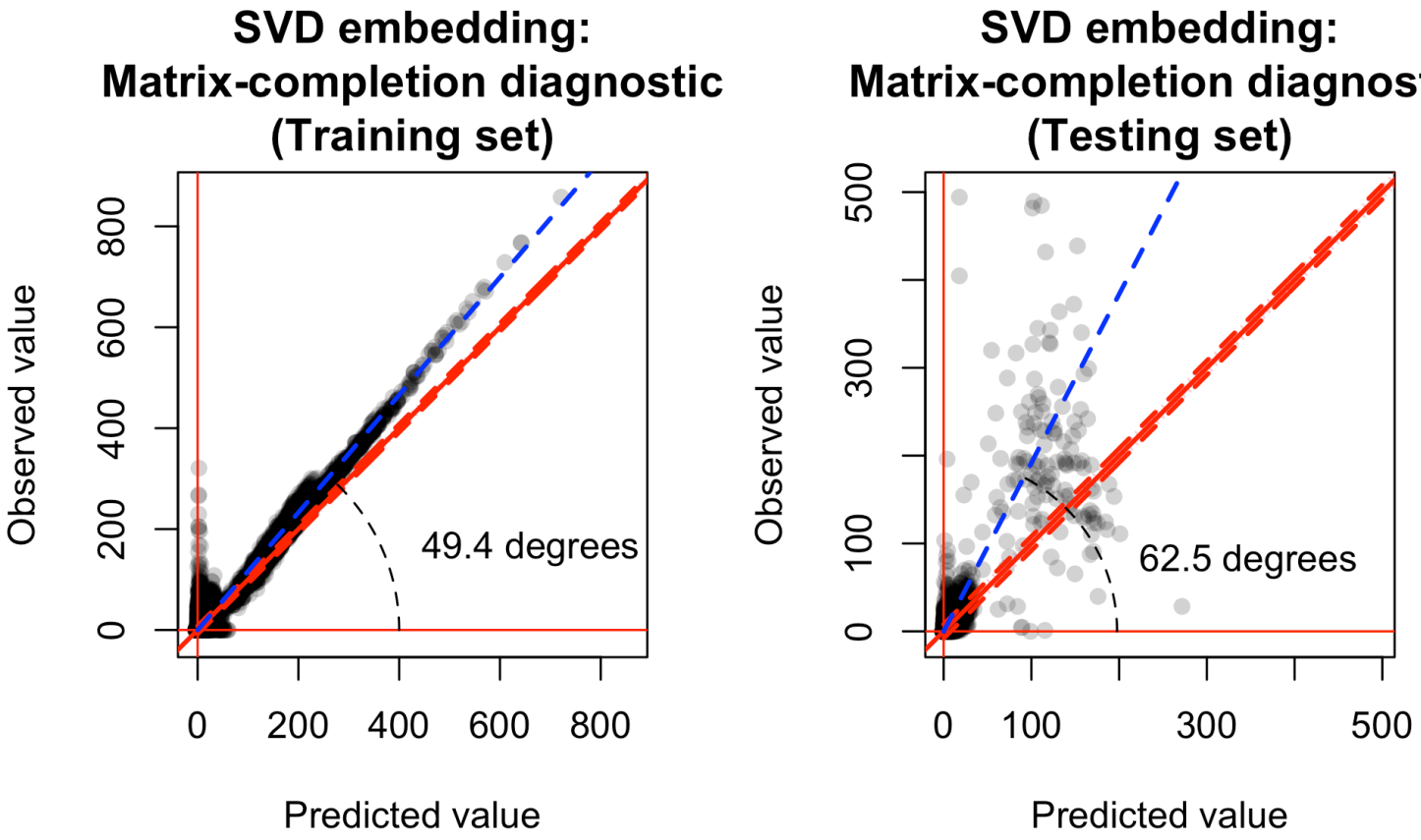
$$\boldsymbol{A} \in \mathbb{R}^{n \times p} \; : \; n \text{ cells and } p \text{ genes}$$

$$\boldsymbol{A} = \hat{\boldsymbol{U}} \hat{\boldsymbol{D}} \hat{\boldsymbol{V}}^{\top} \} \text{ SVD}$$

$$\underbrace{\hat{\boldsymbol{X}}}_{\mathbb{R}^{n \times k}} = \left(\frac{n}{p}\right)^{1/4} \cdot \hat{\boldsymbol{U}}_{1:k} \sqrt{\hat{\boldsymbol{D}}_{1:k}}$$

$\mathbb{R}^{n \times k}$ matrix that encodes the lower-dimension embedding, based on a low-rank approximation of $\boldsymbol{A}$

The most common embedding, based on the SVD, is not desirable to model our single-cell RNA-seq dataset of oligodendrocytes from Marques et. al (2016).



Using softImpute,
matrix-completion variant of SVD
(Mazumder et. al 2010)

The SVD embedding is actually a particular instance of a random dot product model, using the Gaussian distribution with constant variance.

Low-dim. embedding for cells

Low-dim. embedding for genes

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{i.i.d.}{\sim} G, \quad \text{and} \quad \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p \overset{i.i.d.}{\sim} H$$

$$A_{ij} \sim F\left(\theta_{ij} = \boldsymbol{X}_i^\top \boldsymbol{Y}_j\right)$$

The SVD embedding is actually a particular instance of a random dot product model, using the Gaussian distribution with constant variance.

Low-dim. embedding for cells

Low-dim. embedding for genes

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{i.i.d.}{\sim} G, \quad \text{and} \quad \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p \overset{i.i.d.}{\sim} H$$

$$A_{ij} \sim F\left(\theta_{ij} = \boldsymbol{X}_i^\top \boldsymbol{Y}_j\right)$$

Preprocessed single-cell RNA-seq data

A single-parameter distribution family

The SVD embedding is actually a particular instance of a random dot product model, using the Gaussian distribution with constant variance.

Low-dim. embedding for cells

Low-dim. embedding for genes

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{i.i.d.}{\sim} G, \quad \text{and} \quad \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p \overset{i.i.d.}{\sim} H$$

$$A_{ij} \sim F\left(\theta_{ij} = \boldsymbol{X}_i^\top \boldsymbol{Y}_j\right)$$

Preprocessed single-cell RNA-seq data

Dot product between latent vectors

The SVD embedding is actually a particular instance of a random dot product model, using the Gaussian distribution with constant variance.

Low-dim. embedding for cells

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{i.i.d.}{\sim} G, \quad \text{and} \quad \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p \overset{i.i.d.}{\sim} H$$

$$A_{ij} \sim F\left(\theta_{ij} = \boldsymbol{X}_i^\top \boldsymbol{Y}_j\right)$$

Exponential-family distribution

$$p(A_{ij} \mid \theta_{ij}) \propto \exp\left(A_{ij} \cdot \theta_{ij} - g(\theta_{ij})\right)$$

Natural parameter

The SVD embedding is actually a particular instance of a random dot product model, using the Gaussian distribution with constant variance.

Low-dim. embedding for cells

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{i.i.d.}{\sim} G, \quad \text{and} \quad \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p \overset{i.i.d.}{\sim} H$$

$$A_{ij} \sim F\left(\theta_{ij} = \boldsymbol{X}_i^\top \boldsymbol{Y}_j\right)$$

Exponential-family distribution

Intuitive idea: For a given one-parameter exponential-family distribution F, estimate the embedding $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ by minimize the negative log-likelihood using $\boldsymbol{A}$ (i.e., “MLE”)

The SVD embedding is actually a particular instance of a random dot product model, using the Gaussian distribution with constant variance.

Low-dim. embedding for cells

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{i.i.d.}{\sim} G, \quad \text{and} \quad \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p \overset{i.i.d.}{\sim} H$$

$$A_{ij} \sim F\left(\theta_{ij} = \boldsymbol{X}_i^\top \boldsymbol{Y}_j\right)$$

$$\underbrace{\mathcal{L}(\boldsymbol{X}, \boldsymbol{Y})}_{\text{Loss function to minimize}} = \frac{1}{np} \sum_{(i,j)} \underbrace{\left[g(\boldsymbol{X}^\top \boldsymbol{Y}_j) - A_{ij} \cdot (\boldsymbol{X}_i^\top \boldsymbol{Y}_j) \right]}_{\text{Negative log-likelihood}}$$

The SVD embedding is actually a particular instance of a random dot product model, using the Gaussian distribution with constant variance.

Low-dim. embedding for cells

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{i.i.d.}{\sim} G, \quad \text{and} \quad \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p \overset{i.i.d.}{\sim} H$$

$$A_{ij} \sim F\left(\theta_{ij} = \boldsymbol{X}_i^\top \boldsymbol{Y}_j\right)$$

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{np} \sum_{(i,j)} \left[g(\boldsymbol{X}^\top \boldsymbol{Y}_j) - A_{ij} \cdot (\boldsymbol{X}_i^\top \boldsymbol{Y}_j) \right]$$

Plugging in Gaussian w/ constant variance

$$\propto \frac{1}{np} \sum_{(i,j)} \left[\left(A_{ij} - (\boldsymbol{X}_i^\top \boldsymbol{Y}_j) \right)^2 \right]$$

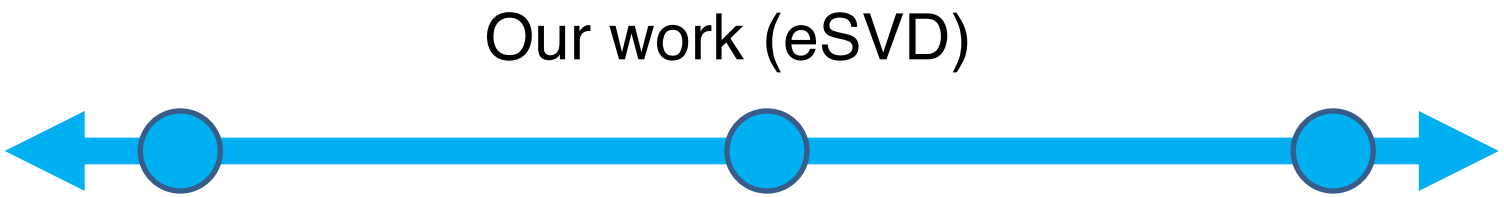
(Solved via SVD)

Many previous papers that use similar formulations either lack modeling flexibility or the desired theoretical properties.

Theoretical statistics literature:
(Zhao et. al 2015, Chen et. al 2019)
(Matrix factorization)

Biostatistics literature:
(Witten 2011, Pierson et. al 2015)
(Fancy dot product models)

Many previous papers that use similar formulations either lack modeling flexibility or the desired theoretical properties.



Goals: Computational efficiency, theoretical foundation, modeling flexibility, retainment of random dot product model, amendable for a tuning procedure

eSVD performs a suitable initialization then uses alternating minimization based on the negative log-likelihood.

$$\underbrace{\arg\min_{\boldsymbol{X},\boldsymbol{Y}}}_{\boldsymbol{X}\in\mathbb{R}^{n\times k},\boldsymbol{Y}\in\mathbb{R}^{p\times k}\quad(\text{where } \boldsymbol{X}\boldsymbol{Y}^\top=\boldsymbol{\Theta})} \mathcal{L}(\boldsymbol{X},\boldsymbol{Y}) = \arg\min_{\boldsymbol{X},\boldsymbol{Y}} \frac{1}{np}\sum_{(i,j)}\left[g\left(\boldsymbol{X}_i^\top\boldsymbol{Y}_j\right) - A_{ij}\cdot\left(\boldsymbol{X}_i^\top\boldsymbol{Y}_j\right)\right]$$

eSVD performs a suitable initialization then uses alternating minimization based on the negative log-likelihood.

$$\underset{\boldsymbol{X},\boldsymbol{Y}}{\arg\min}\, \mathcal{L}(\boldsymbol{X},\boldsymbol{Y}) = \underset{\boldsymbol{X},\boldsymbol{Y}}{\arg\min}\, \frac{1}{np} \sum_{(i,j)} \left[g\left(\boldsymbol{X}_i^\top \boldsymbol{Y}_j\right) - A_{ij} \cdot \left(\boldsymbol{X}_i^\top \boldsymbol{Y}_j\right) \right]$$

- Stage 1: Initialization
  - Initialize the estimate based on SVD of some transformation of $\boldsymbol{A}$

eSVD performs a suitable initialization then uses alternating minimization based on the negative log-likelihood.

$$\arg\min_{\boldsymbol{X},\boldsymbol{Y}} \mathcal{L}(\boldsymbol{X},\boldsymbol{Y}) = \arg\min_{\boldsymbol{X},\boldsymbol{Y}} \frac{1}{np} \sum_{(i,j)} \left[g\left(\boldsymbol{X}_i^\top \boldsymbol{Y}_j\right) - A_{ij} \cdot \left(\boldsymbol{X}_i^\top \boldsymbol{Y}_j\right) \right]$$

- Stage 1: Initialization
  - Initialize the estimate based on SVD of some transformation of $\boldsymbol{A}$
- Stage 2: Alternating minimization
  - Update the estimate by re-estimating $\boldsymbol{X}$, keeping $\boldsymbol{Y}$ fixed, or re-estimating $\boldsymbol{Y}$, keeping $\boldsymbol{X}$ fixed

eSVD performs a suitable initialization then uses alternating minimization based on the negative log-likelihood.

$$\arg\min_{\boldsymbol{X},\boldsymbol{Y}} \mathcal{L}(\boldsymbol{X},\boldsymbol{Y}) = \arg\min_{\boldsymbol{X},\boldsymbol{Y}} \frac{1}{np} \sum_{(i,j)} \left[g\left(\boldsymbol{X}_i^\top \boldsymbol{Y}_j\right) - A_{ij} \cdot \left(\boldsymbol{X}_i^\top \boldsymbol{Y}_j\right) \right]$$

- Stage 1: Initialization
  - Initialize the estimate based on SVD of some transformation of $\boldsymbol{A}$
- Stage 2: Alternating minimization
  - Update the estimate by re-estimating $\boldsymbol{X}$, keeping $\boldsymbol{Y}$ fixed, or re-estimating $\boldsymbol{Y}$, keeping $\boldsymbol{X}$ fixed
- Stage 3: Reparameterization
  - Ensure identifiability conditions are met; $\left(\frac{1}{n}\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{X}} = \frac{1}{p}\hat{\boldsymbol{Y}}^\top \hat{\boldsymbol{Y}}\right)$

eSVD performs a suitable initialization then uses alternating minimization based on the negative log-likelihood.

- Stage 2: Alternating minimization

eSVD performs a suitable initialization then uses alternating minimization based on the negative log-likelihood.

- Stage 2: Alternating minimization

$$
\begin{aligned}
\boldsymbol{X}^{(t)} &= \underset{\boldsymbol{X}\in\mathbb{R}^{n\times k}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{X}, \overline{\boldsymbol{Y}}^{(t)}) \;:\; \boldsymbol{X}_i^\top \overline{\boldsymbol{Y}}_j^{(t)} \in \mathcal{R}, \quad \forall (i,j), \\
\overline{\boldsymbol{X}}^{(t+1)} &= \sqrt{n} \cdot \mathrm{LeftSVD}(\boldsymbol{X}^{(t)}), \\
\boldsymbol{Y}^{(t+1)} &= \underset{\boldsymbol{Y}\in\mathbb{R}^{p\times k}}{\operatorname{argmin}} \mathcal{L}(\overline{\boldsymbol{X}}^{(t+1)}, \boldsymbol{Y}) \;:\; (\overline{\boldsymbol{X}}_i^{(t+1)})^\top \boldsymbol{Y}_j \in \mathcal{R}, \quad \forall (i,j), \\
\overline{\boldsymbol{Y}}^{(t+1)} &= \sqrt{p} \cdot \mathrm{LeftSVD}(\boldsymbol{Y}^{(t+1)})
\end{aligned}
$$

eSVD performs a suitable initialization then uses alternating minimization based on the negative log-likelihood.

- Stage 2: Alternating minimization

$$\boldsymbol{X}^{(t)} = \underset{\boldsymbol{X} \in \mathbb{R}^{n \times k}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{X}, \overline{\boldsymbol{Y}}^{(t)})$$

$$\boldsymbol{X}^{(t+1)} = \sqrt{n} \cdot \operatorname{LeftSVD}(\boldsymbol{X}^{(t)}),$$

$$\boldsymbol{Y}^{(t+1)} = \underset{\boldsymbol{Y} \in \mathbb{R}^{p \times k}}{\operatorname{argmin}} \mathcal{L}(\overline{\boldsymbol{X}}^{(t+1)}, \boldsymbol{Y})$$

$$\overline{\boldsymbol{Y}}^{(t+1)} = \sqrt{p} \cdot \operatorname{LeftSVD}(\boldsymbol{Y}^{(t+1)})$$

Minimizing one of the factors based on the log-likelihood (Convex problem)

eSVD performs a suitable initialization then uses alternating minimization based on the negative log-likelihood.

- Stage 2: Alternating minimization

Need to ensure the inner product still yields a valid natural parameter

We use Frank-Wolfe to do this constrained optimization

$$\boldsymbol{X}_i^\top \overline{\boldsymbol{Y}}_j^{(t)} \in \mathcal{R}, \quad \forall (i,j),$$

eSVD performs a suitable initialization then uses alternating minimization based on the negative log-likelihood.

- Stage 2: Alternating minimization

$$\boldsymbol{X}^{(t)} = \underset{\boldsymbol{X} \in \mathbb{R}^{n \times k}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{X}, \overline{\boldsymbol{Y}}^{(t)}) \ : \ \boldsymbol{X}_i^\top \overline{\boldsymbol{Y}}_j^{(t)} \in \mathcal{R}, \quad \forall (i,j),$$

$$\overline{\boldsymbol{X}}^{(t+1)} = \sqrt{n} \cdot \operatorname{LeftSVD}(\boldsymbol{X}^{(t)}),$$

$$\boldsymbol{Y}^{(t+1)} = \underset{\boldsymbol{Y} \in \mathbb{R}^{p \times k}}{\operatorname{argmin}} \mathcal{L}(\overline{\boldsymbol{X}}^{(t+1)}, \boldsymbol{Y}) \quad \forall (i,j),$$

$$\overline{\boldsymbol{Y}}^{(t+1)} = \sqrt{p} \cdot \operatorname{LeftSVD}(\boldsymbol{Y}^{(t+1)})$$

We reparameterize each estimate by its left singular vectors for identifiability reasons

eSVD performs a suitable initialization then uses alternating minimization based on the negative log-likelihood.

- Stage 2: Alternating minimization

$$\boldsymbol{X}^{(t)} = \underset{\boldsymbol{X} \in \mathbb{R}^{n \times k}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{X}, \overline{\boldsymbol{Y}}^{(t)}) \ : \ \boldsymbol{X}_i^\top \overline{\boldsymbol{Y}}_j^{(t)} \in \mathcal{R}, \quad \forall (i,j),$$

$$\overline{\boldsymbol{X}}^{(t+1)} = \sqrt{n} \cdot \operatorname{LeftSVD}(\boldsymbol{X}^{(t)}),$$

$$\boldsymbol{Y}^{(t+1)} = \underset{\boldsymbol{Y} \in \mathbb{R}^{p \times k}}{\operatorname{argmin}} \mathcal{L}(\overline{\boldsymbol{X}}^{(t+1)}, \boldsymbol{Y}) \ : \ (\ldots$$

$$\overline{\boldsymbol{Y}}^{(t+1)} = \sqrt{p} \cdot \operatorname{LeftSVD}(\boldsymbol{Y}^{(t+1)})$$

We then move to the next factor. We keep alternating between the two factors until convergence.

We prove convergence rates using theoretical techniques from nonconvex optimization and network literature.

- Estimating low-rank matrix of natural parameters:
  - Requires understanding the sufficient conditions for initialization as well as considering the minimization on a population level (Zhao et al. 2015, Balakrishnan et al. 2017)

We prove convergence rates using theoretical techniques from nonconvex optimization and network literature.

- Estimating low-rank matrix of natural parameters:
  - Requires understanding the sufficient conditions for initialization as well as considering the minimization on a population level (Zhao et al. 2015, Balakrishnan et al. 2017)
- Handling the extra source of randomness due to the hierarchical model:
  - Requires careful study of identifiability conditions and perturbation bounds (Lei 2019)

We prove convergence rates using theoretical techniques from nonconvex optimization and network literature.

- Estimating low-rank matrix of natural parameters:
  - Requires understanding the sufficient conditions for initialization as well as considering the minimization on a population level (Zhao et al. 2015, Balakrishnan et al. 2017)
- Handling the extra source of randomness due to the hierarchical model:
  - Requires careful study of identifiability conditions and perturbation bounds (Lei 2019)
- Applying these generic bounds to specific distribution families:
  - Requires deploying empirical process and concentration bounds (Candes and Sur 2019)

eSVD with a curved Gaussian distribution models the oligodendrocyte dataset much more appropriately.

$$A_{ij} \sim F = \text{Normal}(\mu_{ij},\ \mu_{ij}^2/4)$$

so the natural parameter becomes:

$$\boldsymbol{X}_i^\top \boldsymbol{Y}_j = \theta_{ij} = -1/\mu_{ij}$$

- Most of the distribution's mass is positive
- Overall, allows for more variability compared to the negative binomial or Poisson distribution

eSVD with a curved Gaussian distribution models the oligodendrocyte dataset much more appropriately.

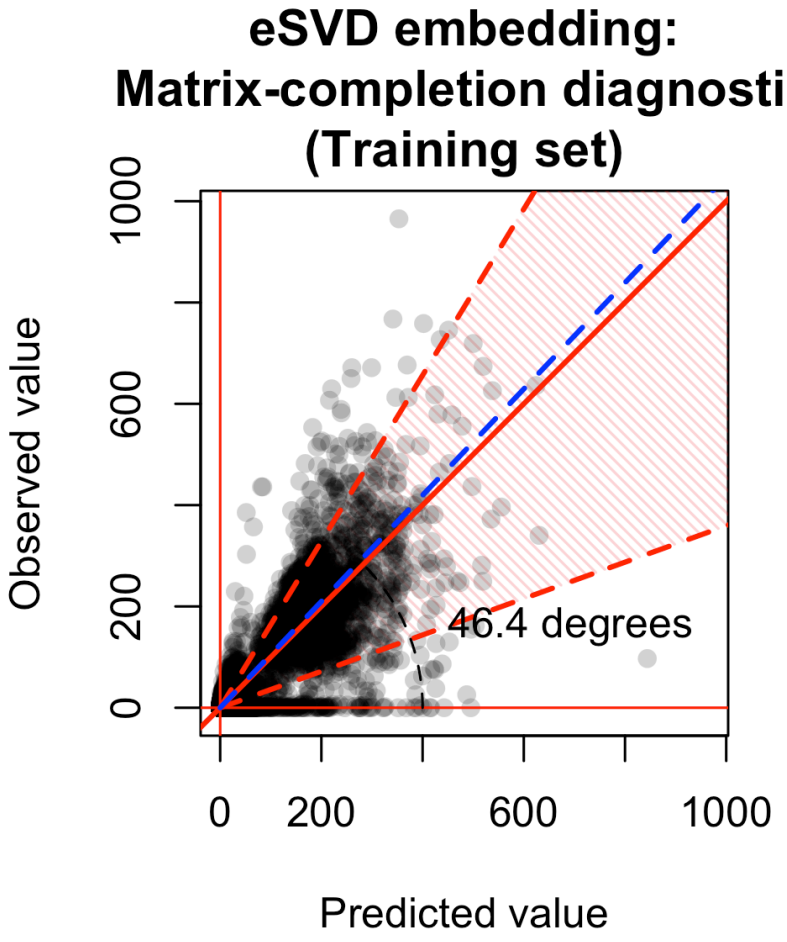
$$A_{ij} \sim F = \text{Normal}(\mu_{ij},\ \mu_{ij}^2/4)$$

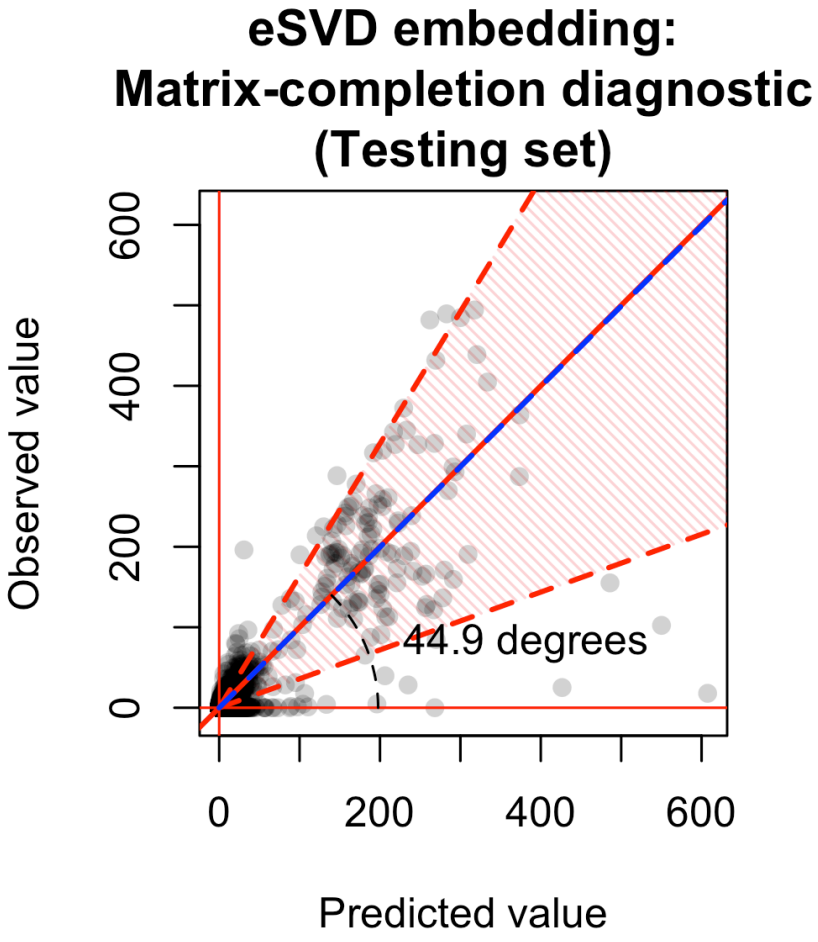
so the natural parameter becomes:

$$\boldsymbol{X}_i^\top \boldsymbol{Y}_j = \theta_{ij} = -1/\mu_{ij}$$

- Most of the distribution's mass is positive
- Overall, allows for more variability compared to the negative binomial or Poisson distribution
- The constant 4 or the dimensionality of the latent space k can be tuned via the matrix-completion diagnostic

eSVD with a curved Gaussian distribution models the oligodendrocyte dataset much more appropriately.





The eSVD embedding enables us to discover two branching trajectories in downstream analyses, whereas the SVD embedding only leads to one trajectory.

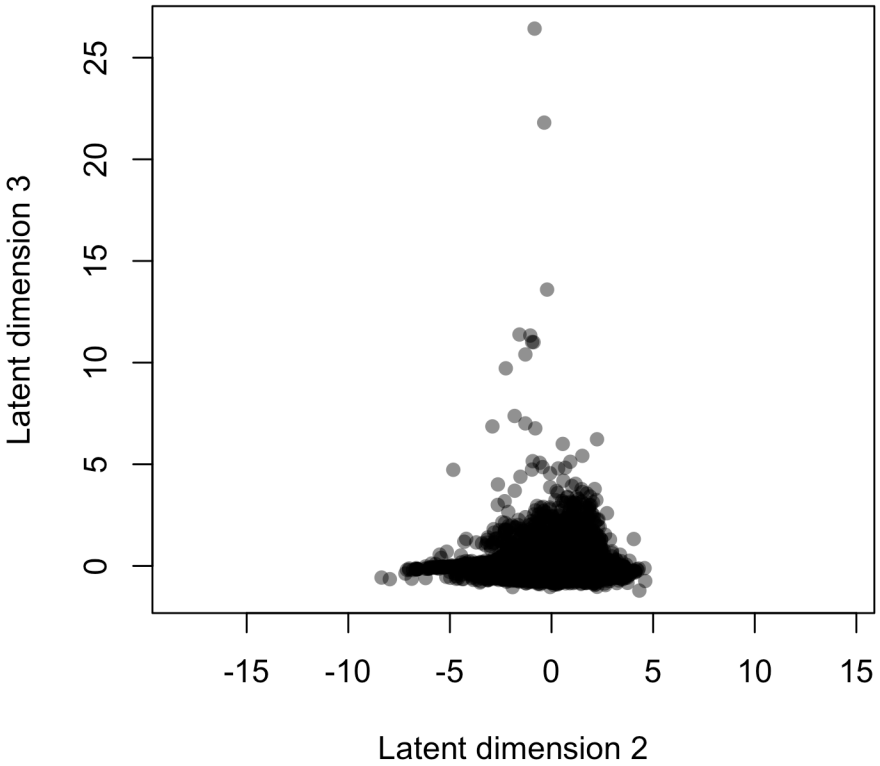
Color coding in upcoming plots	Major Cell Type (6 total)	Number of cell sub-types (13 total)
	Pdgfra+ precursor	1
	Oligodendrocyte precursor cell	1
	Differentiation-committed oligodendrocyte precursors	1
	Newly formed oligodendrocytes	2
	Myelin-forming oligodendrocytes	2
	Mature oligodendrocyte	6

Reminder: $n = 5000+$ cells in total and $p = 900+$ genes

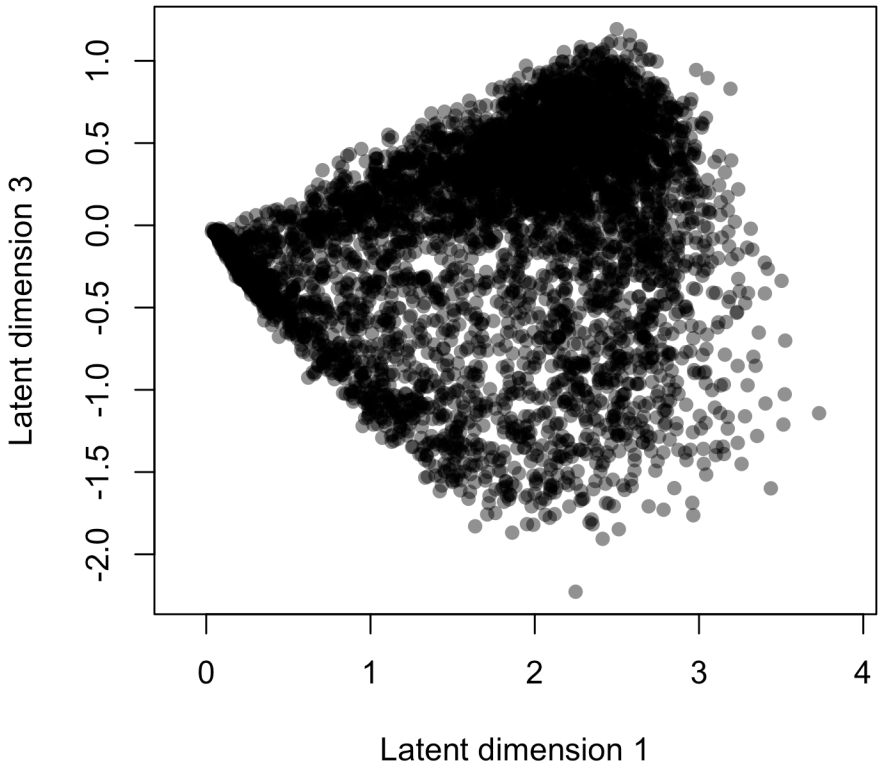
Data from Marques et. al (2016)

The eSVD embedding enables us to discover two branching trajectories in downstream analyses, whereas the SVD embedding only leads to one trajectory.

SVD embedding
(Constant-variance Gaussian)



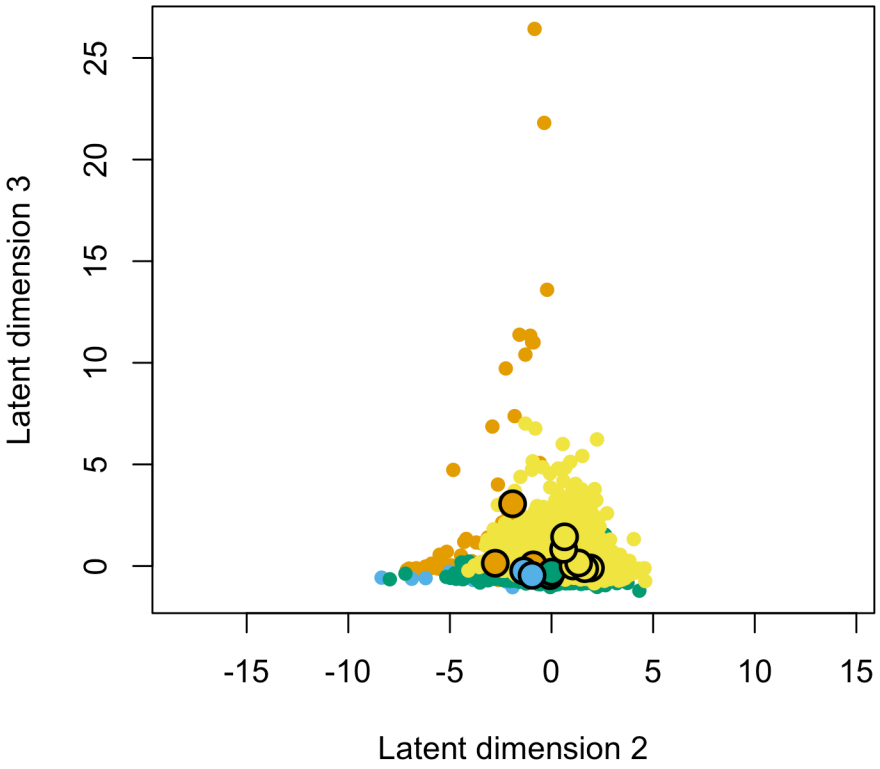
eSVD embedding
(Curved Gaussian)



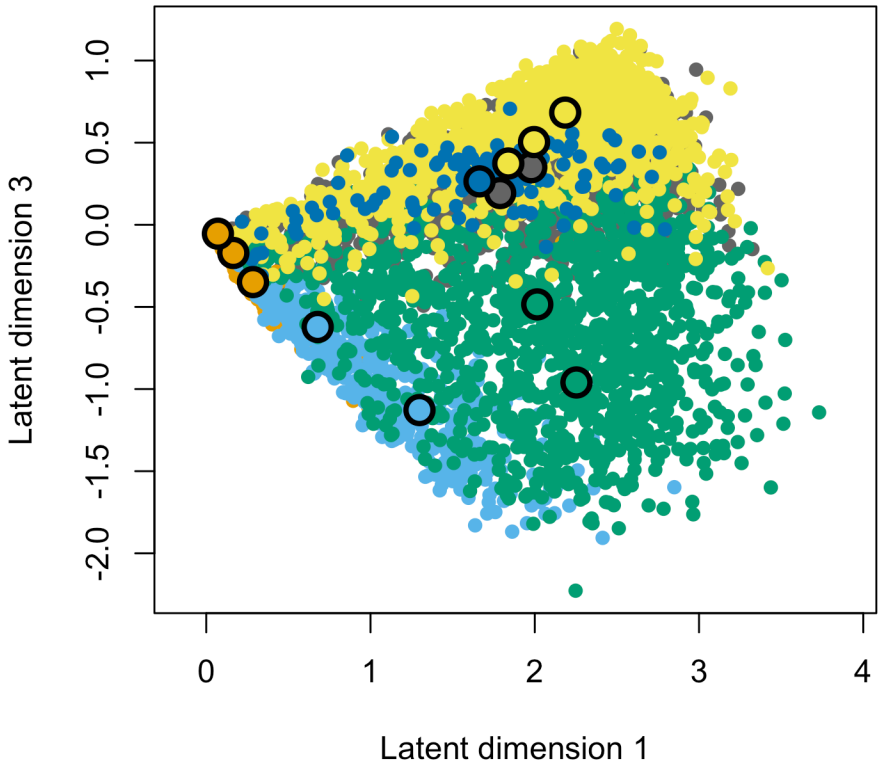
Preprocessing → **Embedding** → Known cell types + Trajectory estimation

The eSVD embedding enables us to discover two branching trajectories in downstream analyses, whereas the SVD embedding only leads to one trajectory.

SVD embedding
(Constant-variance Gaussian)



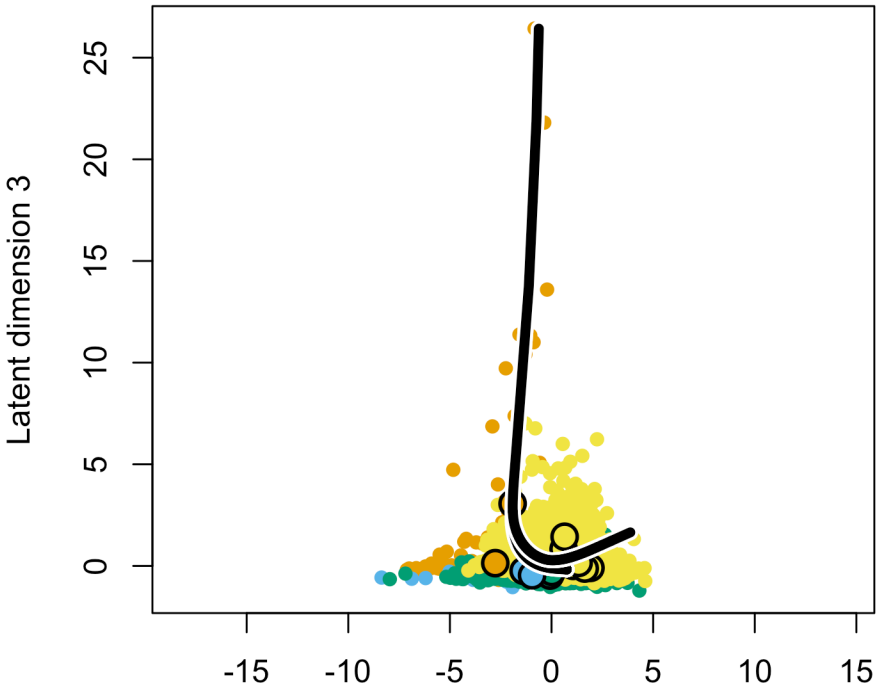
eSVD embedding
(Curved Gaussian)



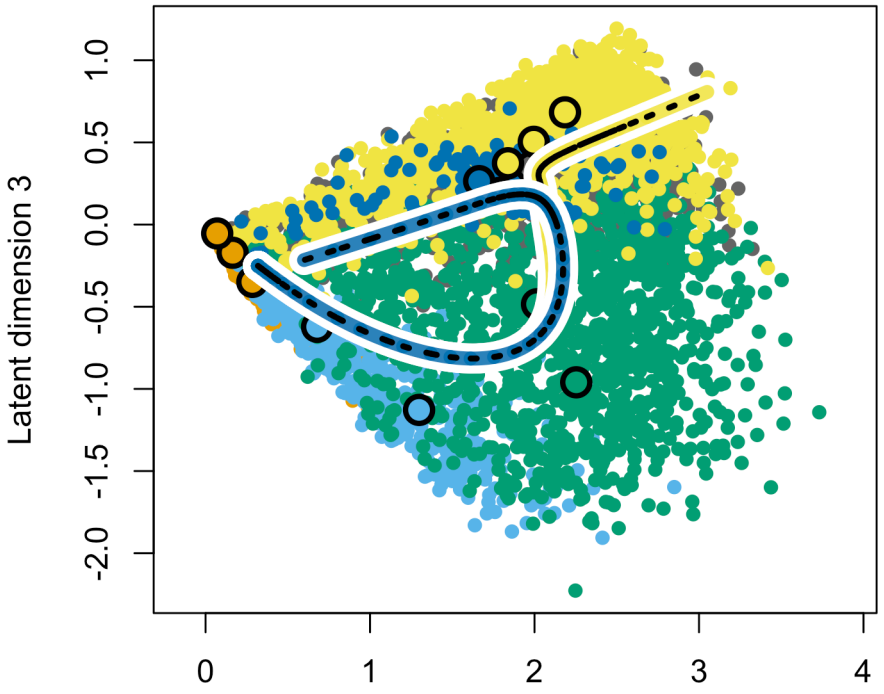
Preprocessing → **Embedding** → **Known cell types** + Trajectory estimation

The eSVD embedding enables us to discover two branching trajectories in downstream analyses, whereas the SVD embedding only leads to one trajectory.

SVD embedding and trajectory
(Constant-variance Gaussian)



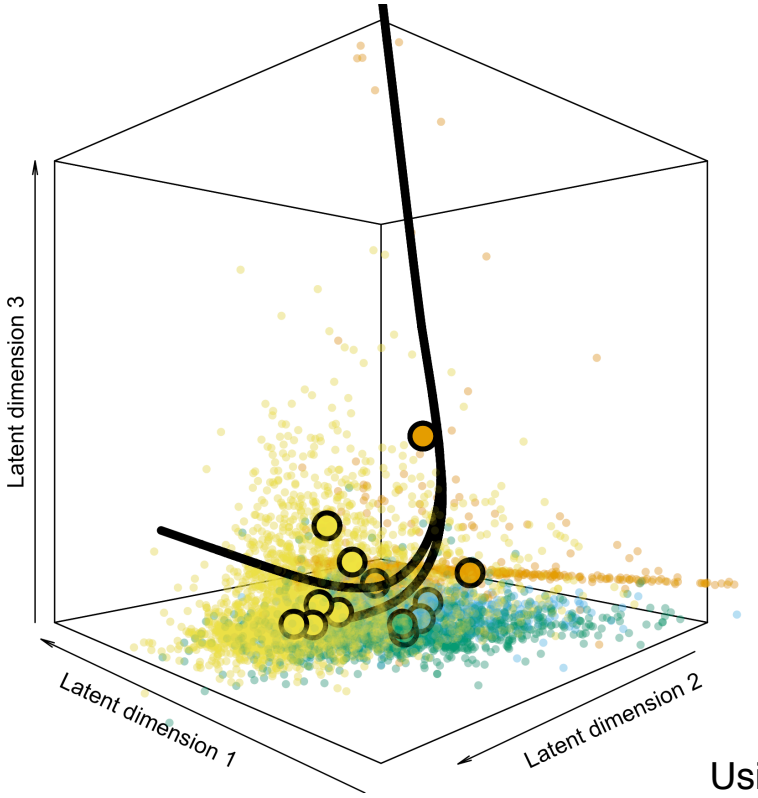
eSVD embedding and trajectory
(Curved Gaussian)



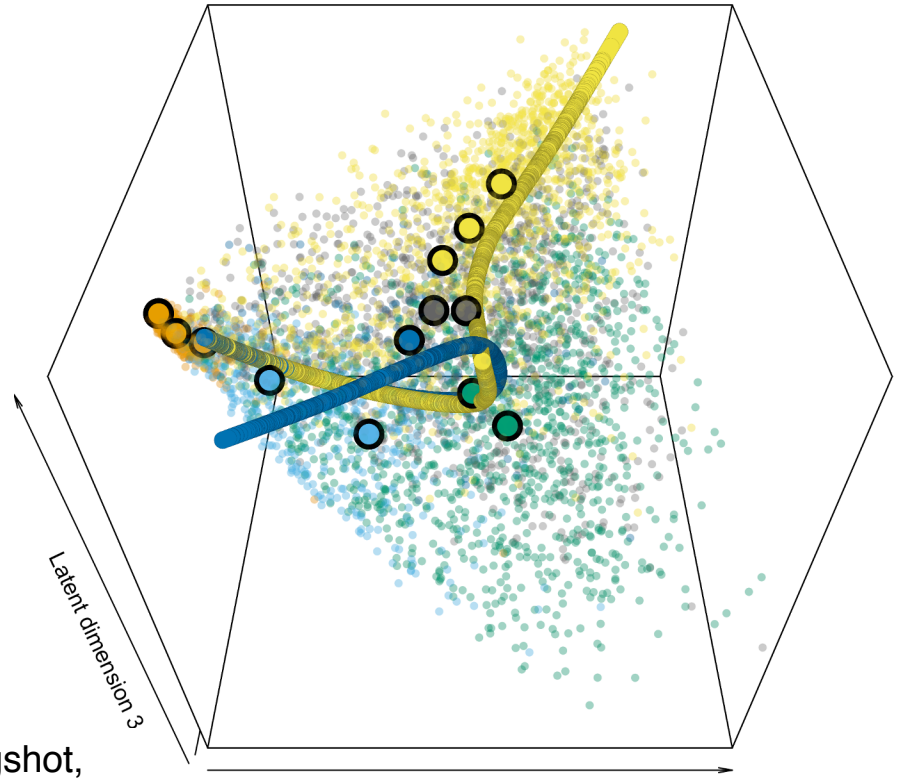
Using Slingshot,
a trajectory estimator
(Street et. al 2018)

The eSVD embedding enables us to discover two branching trajectories in downstream analyses, whereas the SVD embedding only leads to one trajectory.

SVD embedding and trajectory
(Constant-variance Gaussian)



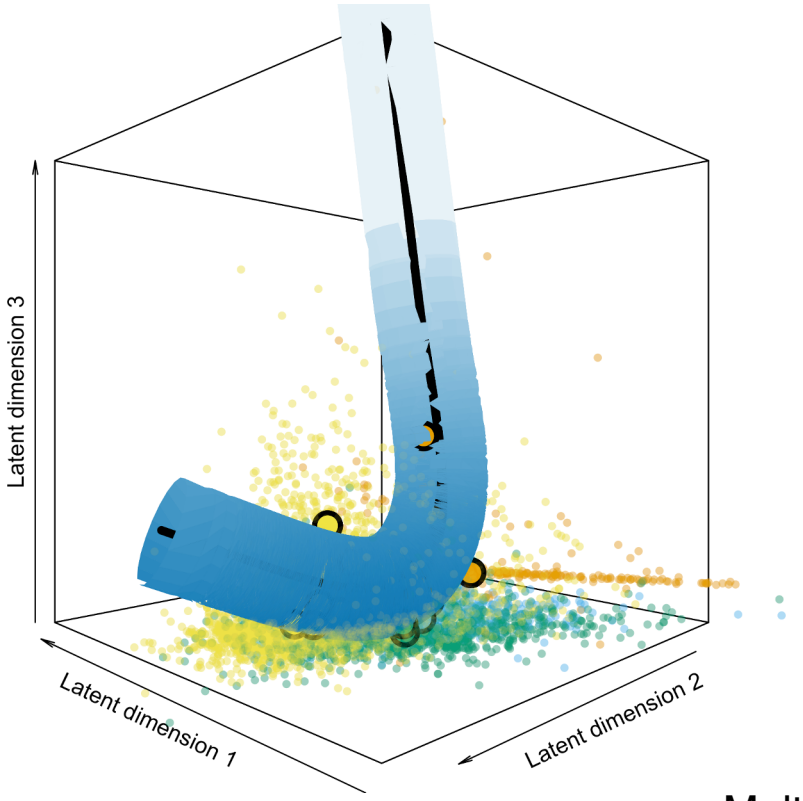
eSVD embedding and trajectory
(Curved Gaussian)



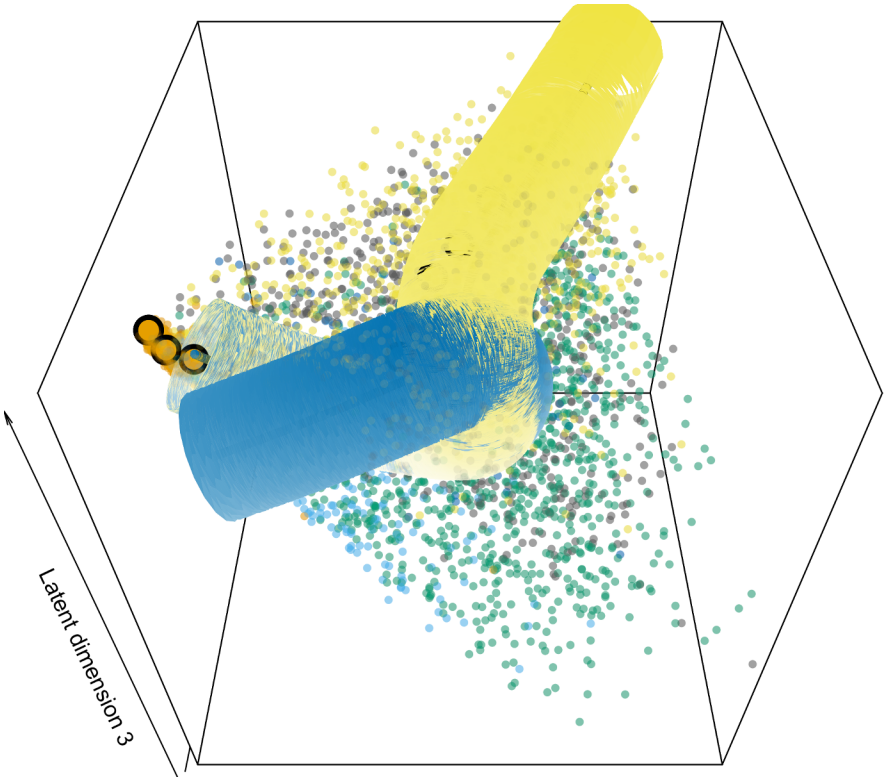
Using Slingshot,
a trajectory estimator
(Street et. al 2018)

The eSVD embedding enables us to discover two branching trajectories in downstream analyses, whereas the SVD embedding only leads to one trajectory.

SVD embedding and uncertainty tube
(Constant-variance Gaussian)



eSVD embedding and uncertainty tube
(Curved Gaussian)



Multiple trajectories in agreement with ongoing research
(Marques et. al 2018, Jakel et. al 2019)

Since embeddings offer many more downstream applications, future work can strengthen eSVD's theoretical properties as well as modeling flexibility.

Other potential downstream applications:

- Visualization
- Imputation
- Clustering (or bi-clustering)

Summary: eSVD is a non-linear embedding method based on any one-parameter exponential-family distribution, and we prove theoretical foundation and apply it to study oligodendrocytes' developmental trajectories.

Thank you!

linnylin92


Kathryn Roeder

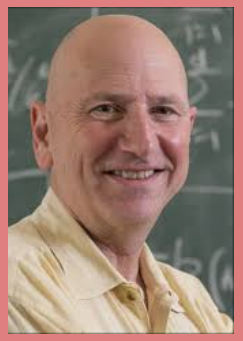

Jing Lei


Max G'Sell


Han Liu


Anjali Mazumder


Ryan J. Tibshirani

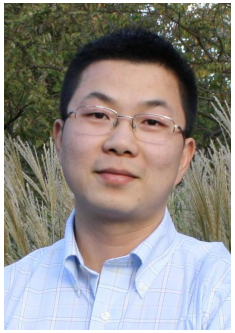

Larry Wasserman

eSVD is a non-linear embedding method based on any one-parameter exponential-families, and we prove theoretical foundation and apply it to study oligodendrocytes' developmental trajectories.

Possible questions to ask:

- More details about oligodendrocyte data or known scientific findings
- Identifiability condition
- Assumptions for convergence
- Theorems for convergence
- Details on this missing-value diagnostic/tuning procedure
- Results from simulation studies
- Details about how Slingshot works (i.e., how to estimate developmental trajectories)
- Relation to other non-linear embedding methods or matrix factorization work
- Thoughts about how to improve eSVD in future work
- Brief overview of the other chapters in thesis


Kathryn Roeder

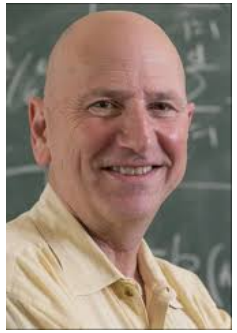

Jing Lei


Max G'Sell


Han Liu


Anjali Mazumder


Ryan J. Tibshirani


Larry Wasserman