Joint identification of neuron types and type-specific activity-regulated genes with coupled autoencoders

04/12/2021

Yeganeh Marghi
Motivation

- From in-depth analysis of cells to understanding the brain
- Studying the whole brain at single-cell resolution by single-cell omics
- The potential to unravel the molecular programs underlying the cellular diversity

Motivation

• From in-depth analysis of cells to understanding the brain
• Studying the whole brain at single-cell resolution by single-cell omics
• The potential to unravel the molecular programs underlying the cellular diversity
• Measurement noise and biological variation cause significant challenges

Single-cell data: a mixture landscape

Mixture models: measurement is a function of two (random) variables.

$x = f(c, s)$

$x$: scRNA-seq data
$c$: cell type (discrete factor)
$s$: cell type-dependent variations (continuous factor)
Single-cell data: a mixture landscape

Mixture models: measurement is a function of two (random) variables.
Mixture representation learning: variational approach

\[ \mathcal{L}(\phi, \theta) = \mathbb{E}_{q_{\phi}(s,c|x)} [\log p_{\theta}(x|s,c)] - D_{KL} (q_{\phi}(s|x)\|p(s)) - D_{KL} (q_{\phi}(c|x)\|p(c)) \]

Mixture representation learning: variational approach

\[ x_a = f(c_a, s_a) \]

\[ f(c_a, s_a) = p(c_a)p(s_a | c_a) \]

- \( c_a \): cell type (discrete)
- \( s_a \): cell type-dependent variations (continuous)
Coupled mixture VAE framework (cpl-mixVAE)

Objective function:

\[
\max \sum_{a=1}^{A} (A - 1) \left( \mathbb{E}_{q(s_a, c_a | x_a)} [\log p(x_a | s_a, c_a)] - \mathbb{E}_{q(c_a | x_a)} [D_{KL} (q(s_a | c_a, x_a) \| p(s_a | c_a))] \right) \\
- \sum_{a<b} \mathbb{E}_{q(s_a | c_a, x_a)} \mathbb{E}_{q(s_b | c_b, x_b)} [D_{KL} (q(c_a | x_a)q(c_b | x_b) \| p(c_a, c_b))] \\
\text{s.t. } c_a = c_b \quad \forall a, b \in [1, A], \ a < b
\]
Coupled mixture VAE framework (cpl-mixVAE)

Objective function:

$$\max \sum_{a=1}^{A} \mathbb{E}_{q(s_a, c_a|x_a)} \left[ \log p(x_a|s_a, c_a) \right] - \mathbb{E}_{q(c_a|x_a)} \left[ D_{KL} \left( q(s_a|c_a, x_a) \parallel p(s_a|c_a) \right) \right] + H(c_a|x_a)$$

$$s.t. \mathbb{E}_{q(c_a|x_a)} \left[ d^2(c_a, c_0) \right] < \epsilon$$
**Proposition 1.** Consider the problem of mixture representation learning in a multi-arm VAE framework. For $A > B \geq 1$ and $\forall m$,

$$C^A_m(m) > C^B_m(m).$$

**Proposition 2.** In the $A$-arm VAE framework, there exists an $A$ such that $\forall m, n, m \neq n$,

$$C^A_m(m) > C^A_m(n),$$

independent of the relative abundances of categories.

$$C_m(k) = \mathbb{E}_{x|m}[\log p(c = k|x)]$$
Consensus assignment

Using Aitchison geometry:  \[ d(c_a, c_b) = D_A(c_a, c_b), \quad c_a, c_b \in S^K \]
Analogy in machine learning

The MNIST dataset
A-arm VAE framework

Network implementation

\[ x \xrightarrow{s} c \xrightarrow{\text{Gumb.}} \]

\[ n \sim \mathcal{N}(0, I) \]

\[ \ldots, x_a, x_b, \ldots \]

\[ \mathcal{C} \]

\[ \mathcal{D}_a \]

\[ \mathcal{D}_b \]

\[ \hat{x}_a \]

\[ \hat{x}_b \]
Benchmark dataset: interpretation of $c \& s$

Continuous factors

Discrete factors

Consensus among arms
Benchmark dataset: unknown |c|
scRNA-seq dataset (Tasic et al., 2018)

- Transcriptomic profiles for 22,365 cells
- 115 excitatory and inhibitory neuron types
- 5000 DE genes

Dissected areas

ALM  VISp

Taxonomy

Chance-level: ~ 6%
scRNA-seq dataset: transcriptomic identities
scRNA-seq dataset: transcriptomic identities

|c| = 115
scRNA-seq dataset: more than 2 arms
Identifying genes regulating continuous variability
Identifying genes ...

L5 NP ALM Trhr Nefl (n44)

L6 CT Nxph2 Sla (n48)
Identifying genes ...

Type dependent

L5 NP ALM Trhr Nefl

L6 CT Nxph2 Sla

L6 IT VIsP Car3

IEG

HKG(CM)

HKG(CC)
Robustness of type-dependent variabilities
Glutamatergic cells

Marker genes
Glutamatergic cells

House-keeping genes
GABAergic cells

Marker genes
GABAergic cells

House-keeping genes
Summary

• Introducing cpl-mixVAE as a general framework to apply the power of collective decision making in unsupervised joint learning of discrete and continuous generative factors.
• Determining the neuronal cell types in an unsupervised setting, while identifying the genes implicated in regulating biologically relevant neuronal states.
• Studying (differential) gene expression variabilities using the type-dependent continuous factor.

\[
\begin{align*}
\hat{x}_a &\quad \hat{x}_b \\
S_a, C_a &\quad S_b, C_b \\
\end{align*}
\]
Future studies

• Multi-modal datasets (Joint identification of cell types and states in different modalities)

• Trajectory-based differential expression analysis for single-cell sequencing data
THANK YOU

Team:
Uygar Sümbül
Rohan Gala
Olga Gliko
Fahimeh Baftizadeh
THANK YOU

We wish to thank the Allen Institute founder, Paul G. Allen, for his vision, encouragement, and support.

brain-map.org
Supplement
Single-cell generator
Single-cell generator

Characterization of Cell Identity
Single-cell generator

Characterization of Cell Identity
## All datasets: overall performance

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Chance-level</th>
<th></th>
<th></th>
<th>Method</th>
<th>ACC (%) $\uparrow$ (mean ± s.d.)</th>
<th>Computation $\uparrow$ (iteration/sec)</th>
<th>Disentanglement score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>10.0%</td>
<td>10</td>
<td>2</td>
<td>InfoGAN</td>
<td>77.87 ± 21.68</td>
<td>12.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>JointVAE</td>
<td>68.99 ± 11.76</td>
<td>74.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CascadeVAE</td>
<td>81.41 ± 09.54</td>
<td>23.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cpl-mixVAE</td>
<td>84.56 ± 06.47</td>
<td>17.5</td>
<td></td>
</tr>
<tr>
<td>dSprite</td>
<td>33.3%</td>
<td>3</td>
<td>6</td>
<td>JointVAE</td>
<td>44.79 ± 03.88</td>
<td>52.6</td>
<td>74.51 ± 05.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CascadeVAE</td>
<td>78.84 ± 15.65</td>
<td>15.4</td>
<td>90.49 ± 05.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cpl-mixVAE</td>
<td>96.30 ± 09.15</td>
<td>20.6</td>
<td>89.98 ± 04.09</td>
</tr>
<tr>
<td>scRNA-seq</td>
<td>06.3%</td>
<td>115</td>
<td>2</td>
<td>JointVAE</td>
<td>12.53 ± 01.83</td>
<td>28.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CascadeVAE</td>
<td>02.69 ± 00.05</td>
<td>03.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>cpl-mixVAE</td>
<td>38.78 ± 01.26</td>
<td>10.1</td>
<td></td>
</tr>
</tbody>
</table>
Consensus assignment

- $|c|=7$
- $|c|=10$
- $|c|=15$
- $|c|=30$