Semisoft Clustering of Single-cell Data

Mohamad Haghir Ebrahimabadi
Siddharth Kothari
Uma B Kota

What is SOUP?

- A method of semisoft clustering for classifying pure and transitional cells based on gene expression..

Semisoft clustering is a combination of:

1. Hard clustering is where each data point belongs to one single cluster
2. In Soft clustering , a data point can belongs to multiple clusters

Growth in human body involves differentiation of cell from progenitors into more specialized descendants. This process comprise of various cells of recognizable types (pure cells) and cells in transitional phase between two pure cells (transitional single cells).

SOUP Model

$$\Theta \in \mathbb{R}_{+}^{n \times K}$$

Θ : Non-negative membership matrix
Represents the membership of n cells to K clusters

$$\Theta_i := (\theta_{i1}, \ldots, \theta_{iK})$$

Each row of the membership matrix showing portion of cell i in k clusters. Pure cells will have 1 in a clusters and 0 in k-1 clusters

$$C \in \mathbb{R}^{p \times K}$$

Cluster Centers, expected gene expression for each pure cell type

$$X = \Theta C^T + E$$

Simple probability model for gene expression matrix, where E is a zero - mean noise matrix

$$X \in \mathbb{R}^{n \times p}$$

X : Expression matrix of p genes in n cells

SOUP Model

$$A := \mathbb{E}\left[XX^T\right] = \Theta Z \Theta^T + \sigma^2 I$$

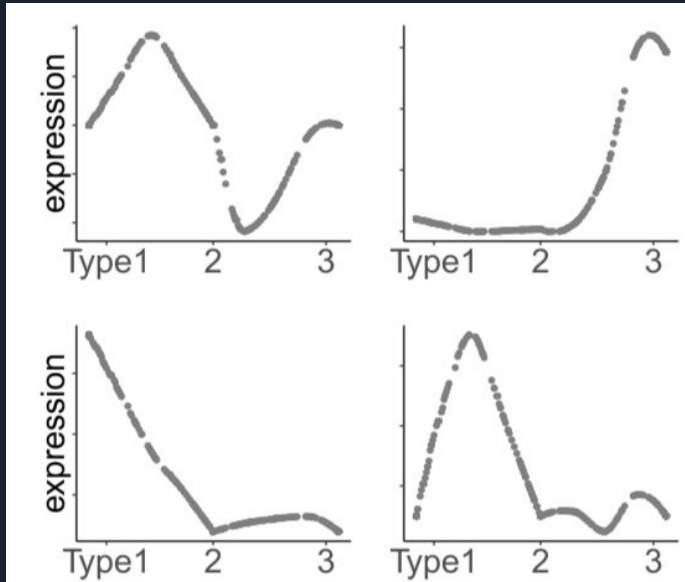
- Gramian Matrix: each entry (i,j) of matrix A represents the correlation between row i and j in matrix X.
- Cell-cell similarity matrix

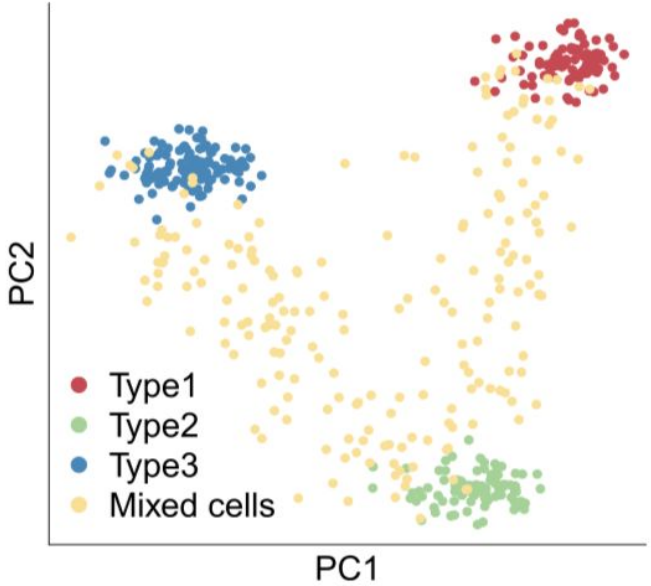
$$Z = C^T C \in \mathbb{R}^{K \times K}$$

Represents the association among different cell types

Example of a SOUP model

Type 1 -> Type 2 -> Type 3





Four differentially expressed genes along the developmental trajectory, with potentially nonlinear differentiation patterns

Simulation of 300 pure cells and 200 mixed cells, visualized in the leading principal component space.

SOUP Algorithm

Finding the set of pure cells and then cluster transitional cells

Pure cells play a critical role in this problem. They provide valuable information for recovering cluster centers which further guides the estimation of Θ for the mixed cells.

So the algorithm is broken into two parts:

Theorem 1: Identifiability - Finding pure cells and then estimating Θ
Theorem 2: SOUP clustering - Soft clustering of transitional cells based on Θ

SOUP Algorithm

Theorem 1: Identifiability

Assumptions:

1. Θ is a membership matrix
2. There exist at least two pure cells per cluster
3. Z is full rank.

Method:

First we calculate purity score for cell i by

$$\hat{S}_i = \{j \neq i : \text{the top } \epsilon \text{ percent with the largest } |\hat{A}_{ij}|\},$$

$$\hat{p}_i = \frac{1}{|\hat{S}_i|} \sum_{j \in \hat{S}_i} \frac{|\hat{A}_{ij}|}{\hat{m}_j}, \text{ where } \hat{m}_i = \max_{j \neq i} |\hat{A}_{ij}|,$$

Then we estimate set of pure cells i by taking top γ percent of cells

$$\hat{\mathscr{P}} = \{i : \text{the top}\gamma \text{ percent with the largest } \hat{p}_i\}$$

Then we cluster these cells into K clusters using K means

SOUP Algorithm

To recover Θ, consider the top K eigenvectors of the similarity matrix A, denoted as $V \in \mathbb{R}^{n \times K}$.

There exists a matrix $Q^* \in \mathbb{R}^{K \times K}$ such that $\Theta = VQ^*$

We have identified the set of pure cells I and their clusters, so we essentially know their memberships $\Theta_{\mathcal{I}\cdot}$.

Then it is easy to get Q* from the submatrix $\Theta_{\mathcal{I}\cdot} = V_{\mathcal{I}\cdot}Q^*$

Theorem 2: SOUP clustering

Optimization of the equation $\min_{Q \in \mathbb{R}^{K \times K}} \|\Theta_{\mathcal{I}\cdot} - V_{\mathcal{I}\cdot}Q\|_F^2$

has a unique solution Q* such that Θ=VQ*.

SOUP Algorithm

Parameter Tuning

We need to tune two parameters:

1. γ: The quantile γ should be an estimate of the proportion of pure cells in the data. However, authors found that the results has not varied a lot with change in γ and used value of 0.5

2. ϵ: It is the the smallest ratio of pure cell across all cell types $\epsilon \leq \min_k |\mathcal{I}_k|/n,$. Authors have used the value 0.1 for 1000 cells, 0.05 for 1000-2000 cells and 0.03 for larger number of cells

Gene Selection

Not all genes provide useful information like housekeeping genes which are unlikely to defer across cell types. So to select genes two algorithms are used:

1. The DESCEND algorithm based on the Gini index

2. The Sparse PCA (SPCA) algorithm

Performance Evaluation

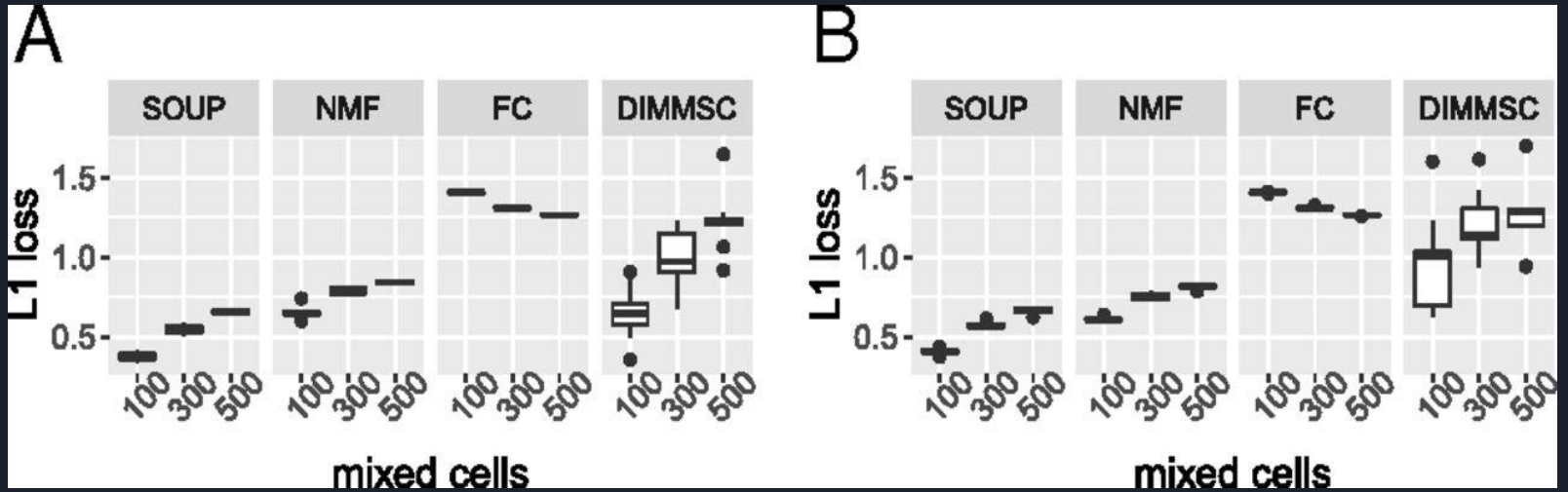
Due to lack of semi-soft clustering algorithms, SOUP has been compared with following three algorithms:

1. NMF, where we use the standard algorithm to solve for nonnegative $(\Theta, \hat{C})$
2. Fuzzy C-Means (FC), a generic soft-clustering algorithm
3. DIMMSC (Dirichlet mixture model for clustering droplet-based single cell) , a probabilistic clustering algorithm for single-cell data based on Dirichlet mixture models

Splatter is used to create synthetic data in R. Total 500 genes, 300 pure cells and 100-500 transtational cells were also stimulated. Splatter is a single-cell simulation framework that generates synthetic scRNA-seq data with hyperparameters estimated from a real dataset.

Algorithms are compared based on L1 norm (least absolute errors). Average L1 loss per cell was calculated $\frac{1}{n}\left\|\hat{\Theta} - \Theta\right\|_1$

Performance Evaluation



Without Dropout

With Dropout

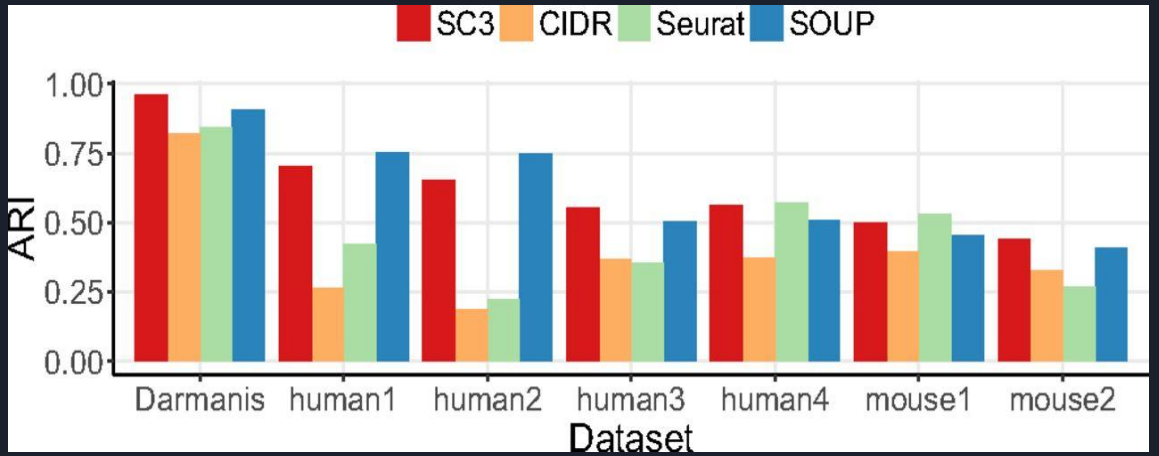
Three simulations with 100, 300 and 500 mixed cells was done. γ = 0.5 was used in all three simulations but SOUP remained stable across.

We can also observe that SOUP outperforms everything else even with dropout which is one of the biggest challenge in single cell RNA - seq data.

Performance Evaluation

SOUP can be used as hard clustering algorithm and it was compared with three popular hard clustering algorithms:

1. SC3, or single-cell consensus clustering
2. CIDR, or clustering through imputation and dimensionality reduction
3. Seurat



Case Studies

Fetal Brain Cells I:

- SOUP applied to a fetal brain scRNA-seq dataset.
- Dataset : 220 developing fetal brain cells between 12 and 13 gestational weeks
- With help of marker genes single cells are labeled with seven types:
  - 2 subtypes of apical progenitors (AP1, AP2)
  - 2 subtypes of basal progenitors (BP1, BP2)
  - 3 subtypes of neurons (N1, N2, N3)
- At this age many cells are still transitioning between different types.Therefore, SOUP can be used to recover the fine-scaled soft-clustering structure.

Hard SOUP:

- SOUP is run with K = 2, 3, . . . , 7 on the log-transformed transcript counts and the clusters of cells are examined.
- Initially treated as a hard-clustering problem ,focusing on the dominating type for each cell.
- For K = 6 and 7, some clusters have no cells assigned to them, which is indicative of a misspecified K.
- For K = 5, the algorithm identifies cell types that correspond to A1, A2, B1, N2, and N3 in Camp's nomenclature
- when cells are in various developmental stages, hard clustering appears to overfit the data.

Case Studies

Soft SOUP:

- Next, they examine the soft assignments.
- For each cluster k, is labeled by an anchor gene, which is the marker gene that has the largest anchor score.

Anchor Score:$[C_{gk} - \max\{C_{g,(-k)}\}]/sd(C_{g,(-k)})$

- where $C_{g,(-k)}$ represents the center values of gene g on the (K-1) clusters other than k.
- The expression levels of the five anchor genes along the SOUP trajectory vary smoothly over developmental time

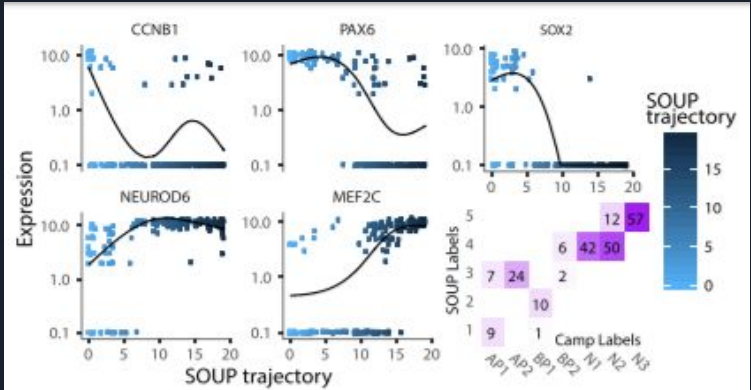


Fig. 4. Expression levels of five anchor genes, visualized in log scale, where the 220 fetal brain cells are ordered by a SOUP unilineal developmental trajectory.

Case Studies

- To model the developmental trajectories the cluster centers determined directly by softSOUP and by hardSOUP are plotted.
- Fitting an MST to the cluster centers, softSOUP identifies two lineages, AP-BP-N and AP-N capturing the true transition of cell types leading to neurons by accounting for the soft membership structures.
- hardSOUP identifies less intuitive BP-AP-N and AP-N lineages. Using Slingshot to fit smooth branching curves to these lineages via simultaneous principal curves, hardSOUP recovers AP-N and BP-N transitions, and the artificial BP1–AP2 transition in the initial MST fit is dropped

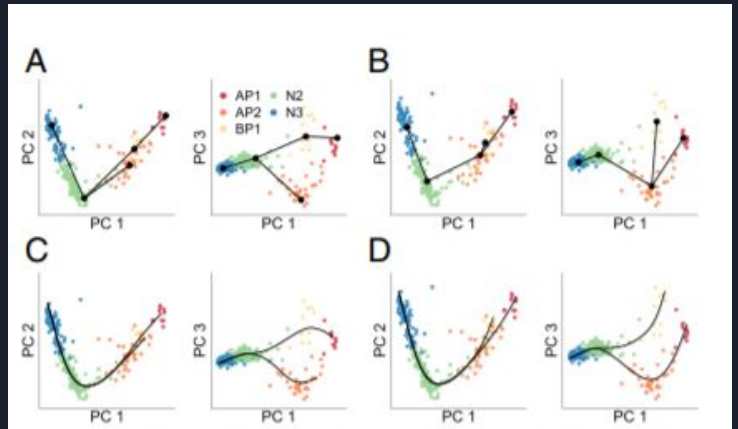


Fig. 5. Two hundred twenty fetal brain cells, cluster centers, lineages, and branching curves in the top three PCs space. Cells are colored according to their SOUP major types, but annotated using Camp labels based on the largest overlap (Fig. 4). (*A* and *B*) MST of softSOUP and hardSOUP cluster centers. (*C* and *D*) Smooth branching curves fitted by Slingshot based on MST in *A* and *B*, respectively.

Case Studies

Fetal Brain Cells II:

- SOUP applied to a richer dataset with 2,309 single cells from human embryonic prefrontal cortex (PFC) from 8 GW to 26 GW .
- six major clusters: neural progenitor cells (NPC), excitatory neurons (EN), interneurons (IN), astrocytes (AST), oligodendrocyte progenitor cells (OPC) and microglia (MIC),
- The objective is to evaluate the developmental trajectories of the major cell types, after excluding IN and MIC, which are known to originate elsewhere and migrate to the PFC

- 1,503 cells remain , after removing IN and MIC and they cluster into K = 7 types and the new types correspond fairly well with the labels
- But, many cells have low majority membership probabilities and do not strongly favor a particular cluster

Case Studies

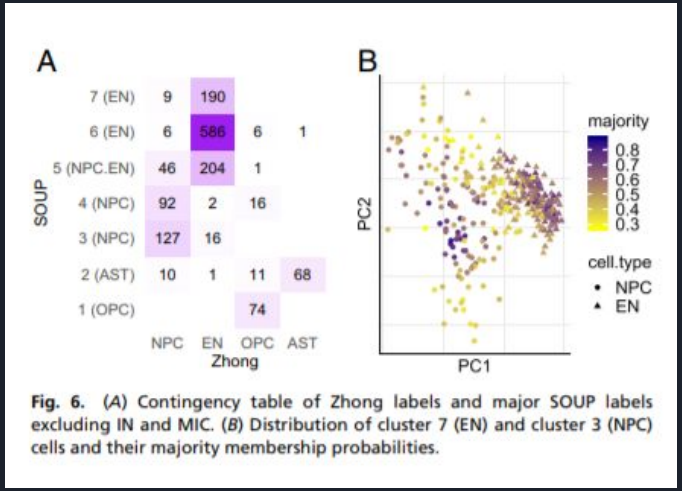


Fig. 6. (*A*) Contingency table of Zhong labels and major SOUP labels excluding IN and MIC. (*B*) Distribution of cluster 7 (EN) and cluster 3 (NPC) cells and their majority membership probabilities.

- In Figure B, cells assigned to clusters 3 (NPC) and 7 (EN), color coded by the majority membership probability
- The two clusters divide the PC space evenly, with the pure cells identifying the cluster centers, while many nonpure cells can be best described as transitioning between clusters.

Case Studies

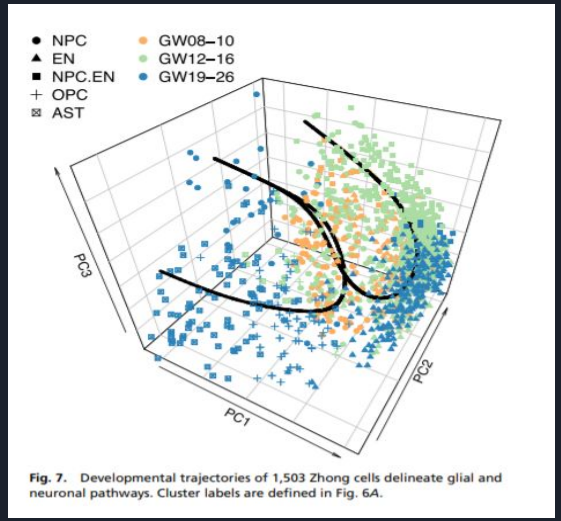


Fig. 7. Developmental trajectories of 1,503 Zhong cells delineate glial and neuronal pathways. Cluster labels are defined in Fig. 6A.

- two developmental paths are revealed by SOUP with help of soft clustering:
  - A neuronal lineage showing NPCs evolving to ENs (clusters: 4 → 3 → 7 → 6 → 5)
  - A glial lineage showing NPCs evolving to OPCs and then to ASTs.
- For the NPCs evolve to OPCs and then to ASTs (clusters: 4 → 3 → 1 → 2) path, The latter transitional step, which differs from the published analysis, is consistent with the literature .
- Therefore, cluster 5, which consists of a mixture of cells labeled as EN and NPC, is placed at the end of the neuronal lineage, hinting some NPC labels are incorrect and that this cluster constitutes a distinct class of ENs.

Discussion

- SOUP is a semisoft single cell clustering algorithm for clustering pure and transitional cells.
- Using SOUP. some cells can be reliably assigned to a cluster and these cells constitute pure types.
- Other cells are transitioning and estimated membership will fall within two, or even more, cell types.
- SOUP estimates membership matrix for each cell using similarity matrix.
- Comparable performance to other algorithms even in case of using as hard clustering algorithm!
- Case studies shows that SOUP can provide valuable information about developmental trajectories.