scFedVI: A Privacy-Preserving Approach to Mitigating Batch Effects in Single-Cell RNA-Sequencing Data

Parishad Mokhber[1 †], Alireza Gargoori Motlagh[2 †],
Babak H. Khalaj[2 *]

[1]Department of Computer Science, Sharif University of Technology, Tehran, Iran
[2]Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran

[†]These authors contributed equally to this work.
*Corresponding Author. Email: khalaj@sharif.edu

1 Abstract

The growing field of single-cell RNA sequencing (scRNA-seq) has revolutionized our understanding of cellular heterogeneity. Batch effects, arising from variances in cell processing such as different chips, sequencing lanes, or harvest times, significantly affect transcriptome measurements, creating discrepancies within and across experiments. In this study, we introduce **Single-Cell Federated Variational Inference (scFedVI)**, a novel federated learning-based method to address the challenge of batch effects in scRNA-seq data analysis. Our results, validated across diverse pancreatic and nervous system scRNA-seq datasets, illustrate that the scFedVI not only effectively corrects batch effects but also utilizes these variations to enhance overall data analysis. This is a significant advancement over conventional non-private batch correction methods, which typically aim to merely eliminate these effects. This method opens new avenues for collaborative research across different laboratories and institutions without compromising data privacy or integrity.

2 Methods

To mitigate these, we integrated deep neural networks, specifically modified Variational Autoencoders (VAEs) [1] with zero inflated negative binomial (ZINB) gene likelihoods, into our federated learning framework for sophisticated batch correction, enhancing biological insights while maintaining data integrity. Our approach utilizes the inherent differences in each client's dataset as a feature rather than a limitation, enabling more robust and generalizable models. By distributing the learning process across clients, each possessing their unique scRNA-seq dataset with distinct batch characteristics, we employed the Federated Averaging (Fed-Avg) [2] algorithm to aggregate the learned models. This approach, prioritizing data privacy, demonstrates enhanced effectiveness in batch effect correction compared to running single-cell variational inference individually on each client's data.

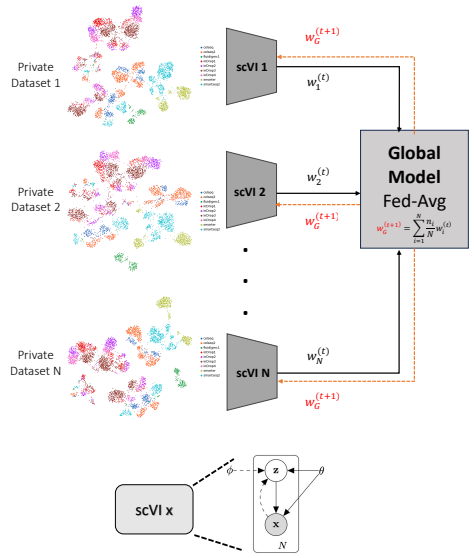

Figure 1: scFedVI Model

Algorithm 1 scFedVI Algorithm

Require: N clients, each with dataset D_i of size n_i, $i = 1, \ldots, N$
Ensure: Global VAE model parameters updated with Fed-Avg
for $round = 1$ to Fed_rounds **do**
for each client $i = 1$ to N in parallel **do**
Initialize VAE model with parameters $\theta_i^{(round)}$
Train VAE on dataset D_i minimizing ELBO loss:
$\mathcal{L}(\theta_i^{(round)}, D_i) = -E_{q_{\theta_i^{(round)}}(z|x)}[\log p_{\theta_i^{(round)}}(x|z)] + KL(q_{\theta_i^{(round)}}(z|x)||p(z))$
Store updated parameters $\theta_i^{(round+1)}$
end for
Aggregate parameters with Federated Averaging:
$\theta^{(round+1)} = \sum_{i=1}^{N} \frac{n_i}{N} \theta_i^{(round+1)}$
for each client $i = 1$ to N **do**
Update client model parameters: $\theta_i^{(round+1)} = \theta^{(round+1)}$
end for
end for=0

3 Results

Performance evaluation using the local inverse simpson's index (LISI) [3] for cell types and sequencing technologies (i.e. batches) confirms that scFedVI outperforms scVI [4], a current leading method in batch correction and integration for single-cell data. The scores indicate the effective number of different categories represented in the local neighborhood of each cell. Therefore, lower cell type LISI and higher batch LISI scores indicate a better performance. Furthermore, we establish the robustness of scFedVI by testing scenarios involving different numbers of clients, while there is sharp decline in the performance of scVI as the average number of samples for clients decrease.

Table 1: Average Batch LISI Scores

	2 Clients	3 Clients	5 Clients
scVI	1.89 ± 0.83	1.87 ± 0.80	1.79 ± 0.79
scFedVI	**2.40 ± 1.05**	**2.55 ± 1.09**	**2.30 ± 0.96**

Table 2: Average Cell Type LISI Scores

	2 Clients	3 Clients	5 Clients
scVI	1.10 ± 0.30	1.13 ± 0.35	1.22 ± 0.48
scFedVI	**1.08 ± 0.26**	**1.09 ± 0.28**	**1.12 ± 0.32**

References

[1] Kingma, D. P. & Welling, M. Auto-encoding variational bayes (2013). `arXiv:1312.6114`.

[2] McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data (2016). `arXiv:1602.05629`.

[3] Korsunsky, I., Millard, N., Fan, J. *et al.* Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods* **16**, 1289–1296 (2019).

[4] Lopez, R., Regier, J., Cole, M. *et al.* Deep generative modeling for single-cell transcriptomics. *Nature Methods* **15**, 1053–1058 (2018).