Limitations of scRNA-seq Zero-Imputation Methods for Network Inference

NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES

Ankit Bhardwaj, Joshua Weiner, Preetha Balasubramanian, Lakshminarayanan Subramanian

Context

- Single Cell RNA Sequencing datasets are highly sparse.
- Many associate majority of zeros to imperfections in sequencing methodology [1].
- Zero imputation techniques aim to address these ‘dropouts’.
- Concerns include introducing artificial signals leading to incorrect biological interpretations [2,3].
- There is a lack of consensus in community on ‘gold-standard’ for treating single cell data.

Problem Introduction

- Zero imputation techniques are evaluated using two approaches.
- For synthetic data where the ‘ground truth’ is available, we can do direct comparison.
- For real data without ‘ground truth’, we measure performance on downstream tasks, like clustering, etc.
- **In prior work, we have not seen an evaluation on the task of gene regulatory network inference.**

Variability of Zero Imputation Methods

- Apply different imputation methods to sc-RNA seq dataset.
- Use GENIE3 algorithm [4] for GRN inference.
- Compare inferred networks using Jaccard similarity.
- Repeat across 5 datasets [5].
  - Datasets from Human Universal Single Cell Hub representing five unique tissues.
  - Random subset of 1,000 genes used for analysis from each.

GRN Inference Variations - 1

Different inferred gene regulatory networks for different imputation pipelines. Top 5% edges.

GRN Inference Variations - 2

	None	SAUCIE	scScope	DeepImpute	MAGIC	SCVI	KNN
None	1.000	0.046	0.088	0.522	0.064	0.058	0.032
SAUCIE	0.103	1.000	0.057	0.045	0.065	0.050	0.035
scScope	0.189	0.060	1.000	0.095	0.048	0.056	0.047
DeepImpute	0.513	0.109	0.153	1.000	0.059	0.046	0.044
MAGIC	0.008	0.026	0.016	0.013	1.000	0.059	0.034
SCVI	0.053	0.047	0.035	0.050	0.038	1.000	0.038
KNN	0.057	0.041	0.036	0.055	0.032	0.060	1.000

	None	SAUCIE	scScope	DeepImpute	MAGIC	SCVI	KNN
None	1.000	0.034	0.031	0.252	0.005	0.040	0.018
SAUCIE	0.041	1.000	0.062	0.048	0.046	0.043	0.056
scScope	0.126	0.033	1.000	0.045	0.036	0.035	0.028
DeepImpute	0.504	0.036	0.119	1.000	0.013	0.037	0.031
MAGIC	0.004	0.029	0.009	0.000	1.000	0.040	0.066
SCVI	0.047	0.032	0.031	0.046	0.027	1.000	0.050
KNN	0.028	0.031	0.043	0.035	0.022	0.024	1.000

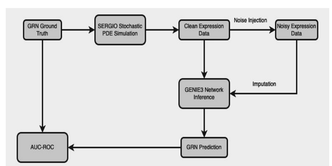
	None	SAUCIE	scScope	DeepImpute	MAGIC	SCVI	KNN
None	1.000	0.047	0.174	0.362	0.014	0.047	0.041
SAUCIE	-	1.000	0.042	0.059	0.030	0.037	0.025
scScope	-	-	1.000	0.151	0.014	0.034	0.038
DeepImpute	-	-	-	1.000	0.013	0.046	0.044
MAGIC	-	-	-	-	1.000	0.022	0.032
SCVI	-	-	-	-	-	1.000	0.029
KNN	-	-	-	-	-	-	1.000

Reference
HU_0379_Blood_GSE172495
HU_0243_Prostate_GSE134355
HU_0325_Esophagus_GSE159929
HU_0231_Pancreas_GSE134355
HU_0269_Uterus_GSE134355

Jaccard Similarity between different GRNs after applying different imputation techniques for 5 human datasets.

Benchmarking on Synthetic Data

- Use SERGIO simulator [6] to generate scRNA-seq data with known GRN.
- Apply noise and dropout to mimic real experimental data.
  - 3 synthetic datasets (DS1, DS2, DS3) with varying complexities, derived from subsets of experimentally validated *E. coli* and *S. cerevisiae* networks.
  - Clean and noisy versions to test imputation effectiveness.
- Benchmark imputation methods against ground truth.
  - Imputation methods evaluated: MAGIC [7], SAUCIE [8], scScope [9], DeepImpute [10], scVI [11], kNN-Smoothing [12].
- Evaluate using GENIE3 for inference and AUC-ROC metrics.



Pipeline for the synthetic benchmarking experiments.

Synthetic Benchmarking Results

Imputation Method	DS1	DS2	DS3
Clean (Dibaeinia & Sinha, 2020)	0.685 ± 0.005	0.806 ± 0.003	0.825 ± 0.003
Noisy (Dibaeinia & Sinha, 2020)	0.478 ± 0.003	0.444 ± 0.003	0.455 ± 0.003
MAGIC (van Dijk et al., 2018)	0.472 ± 0.006	0.489 ± 0.002	0.504 ± 0.003
SAUCIE (Amodio et al., 2019)	0.524 ± 0.022	0.439 ± 0.016	0.481 ± 0.013
scScope (Deng et al., 2019)	0.491 ± 0.051	0.464 ± 0.027	0.478 ± 0.024
DeepImpute (Arisdakessian et al., 2019)	0.530 ± 0.006	0.502 ± 0.005	0.411 ± 0.003
scVI (Lopez et al., 2018)	0.492 ± 0.020	0.505 ± 0.011	0.500 ± 0.007
kNN-Smoothing (Wagner et al., 2018)	0.513 ± 0.020	0.496 ± 0.005	0.480 ± 0.006

AUC-ROC on network inference task with GENIE3.

Takeaways

- High variability in network inference results with different imputation methods.
- Poor performance of imputation methods on synthetic data.
- This highlights the need for standardized pre-processing pipelines.
- We bring attention to the overlooked aspect of SC RNA-seq imputation data, recovering Gene Regulatory Networks, when performing imputation.

References

[1] Jiang, R., Sun, T., Song, D., and Li, J. J. Statistics or biology: the zero-inflation controversy about scrna-seq data. Genome biology, 23(1):31, 2022.
[2] Svensson, V. Droplet scrna-seq is not zero-inflated. Nature Biotechnology, 38(2):147–150, 2020.
[3] Qiu, P. Embracing the dropouts in single-cell rna-seq analysis. Nature communications, 11(1):1169, 2020.
[4] Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. Inferring regulatory networks from expression data using tree-based methods. PloS one, 5(9):e12776, 2010.
[5] HU_0379_ Blood_GSE172495, HU_0325_Esophagus_GSE159929, HU_0231_Pancreas_GSE134355, HU_0243_Prostate_GSE134355, HU_0269_Uterus_GSE134355. http://husch. comp-genomics.org/#/detail/dataset_id
[6] Dibaeinia, P. and Sinha, S. Sergio: A single-cell expression simulator guided by gene regulatory networks. Cell Systems, 11(3):252–271.e11, 2020. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2020.08. 003. URL https://www.sciencedirect.com/ science/article/pii/S2405471220302878.
[7] van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., Bierie, B., Mazutis, L., Wolf, G., Krishnaswamy, S., and Pe'er, D. Recovering gene interactions from single-cell data using data diffusion. Cell, 174(3):716–729.e27, 2018. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2018.05. 061. URL https://www.sciencedirect.com/ science/article/pii/S0092867418307244.
[8] Amodio, M., van Dijk, D., Srinivasan, K., Chen, W. S., Mohsen, H., Moon, K. R., Campbell, A., Zhao, Y., Wang, X., Venkataswamy, M., Desai, A., Ravi, V., Kumar, P., Montgomery, R., Wolf, G., and Krishnaswamy, S. Exploring single-cell data with deep multitasking neural networks. Nat. Methods, 16(11):1139–1145, November 2019.
[9] Deng, Y., Bao, F., Dai, Q., Wu, L. F., and Altschuler, S. J. Scalable analysis of cell-type composition from singlecell transcriptomics using deep recurrent learning. Nature Methods, 16(4):311–314, Apr 2019. ISSN 1548- 7105. doi: 10.1038/s41592-019-0353-7. URL https: //doi.org/10.1038/s41592-019-0353-7.
[10] Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., and Garmire, L. X. Deepimpute: an accurate, fast, and scalable deep neural network method to impute single-cell rna-seq data. Genome Biology, 20(1): 211, Oct 2019. ISSN 1474-760X. doi: 10.1186/ s13059-019-1837-6. URL https://doi.org/10. 1186/s13059-019-1837-6.
[11] Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. Nature, 15(12):1053–1058, 2018. ISSN 1548- 7105. doi: 10.1038/s41592-018-0229-2. URL https: //doi.org/10.1038/s41592-018-0229-2.
[12] Wagner, F., Yan, Y., and Yanai, I. K-nearest neighbor smoothing for high-throughput single-cell rna-seq data. bioRxiv, 2018. doi: 10.1101/ 217737. URL https://www.biorxiv.org/ content/early/2018/04/09/217737.