Advances in Computer, Signals and Systems (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2025.090112
ISSN 2371-8838 Vol. 9 Num. 1

A Review of scRNA-seq Imputation Methods

Zhiqiang Zhang[*]

School of Information, Yunnan Normal, University, Kunming, China
zhangzq2312@163.com
**Corresponding author*

Keywords: Single-cell RNA Sequencing (scRNA-seq); Imputation; Dropout; Model-based Methods; Deep Learning; Autoencoder; Graph Neural Network; Cellular Heterogeneity; Gene Expression; Data Recovery

Abstract: Single-cell RNA sequencing (scRNA-seq) has emerged as a powerful tool for profiling gene expression at the individual cell level, enabling the discovery of cellular heterogeneity that traditional bulk RNA sequencing cannot capture. However, technical limitations such as low transcript capture efficiency, amplification biases, and limited sequencing depth have led to pervasive dropout events, where true gene expression is obscured by excessive zero counts. This review systematically examines and compares the principal imputation methods developed to address these challenges in scRNA-seq data analysis. We categorize these approaches into two broad groups: model-based methods and deep learning methods. Model-based techniques utilize probabilistic models or matrix factorization to exploit similarities among cells and genes—either independently or in combination—to predict and restore missing values. In contrast, deep learning methods leverage the capabilities of autoencoders, graph neural networks, and other innovative network architectures, including generative adversarial networks, to capture complex nonlinear relationships within high-dimensional, noisy data. While model-based approaches offer greater interpretability through explicit statistical assumptions, they are often limited by their sensitivity to noise and data sparsity. Deep learning strategies, although computationally intensive and less interpretable, excel in recovering intricate data structures in large-scale datasets. By providing a comprehensive overview of these imputation strategies, this review aims to guide researchers in selecting the most appropriate methods for their specific datasets and downstream analyses, and to suggest future directions for improving imputation accuracy and integrating multi-omics data.

1. Introduction

Single-cell RNA sequencing (scRNA-seq) technology has rapidly advanced in recent years, providing unprecedented resolution for analyzing gene expression at the cellular level.[1] Unlike traditional bulk RNA sequencing, which only captures the average expression level across a cell population, scRNA-seq enables the profiling of individual cells, thereby revealing subtle differences and complex heterogeneity among cells. This capability has enormous potential in fields such as tumor immunology, developmental biology, and neuroscience, facilitating the discovery of novel cell types, elucidating cell state transitions, and reconstructing cell developmental trajectories. However, due to technical limitations such as low transcript capture efficiency, reverse transcription and amplification biases, and insufficient sequencing depth, scRNA-seq data typically exhibit a high proportion of zero counts, known as “dropout” events.[2] These zeros often not only reflect the failure to capture true transcripts because of technical noise but may also partly represent genuine gene silencing in certain cells, making it a significant challenge to distinguish technical dropouts from biologically meaningful “true zeros.” Moreover, the intrinsic high sparsity, substantial noise, and high dimensionality of the data, compounded by variations in expression distributions and noise characteristics across different experimental platforms and sample types, make it particularly difficult to capture the complex relationships between cells and genes in high-dimensional space.

To address these issues, researchers have developed a variety of imputation methods that leverage the similarity among cells or genes to predict and recover the true gene expression values lost due to technical reasons, thereby enhancing data quality and the accuracy of downstream analyses.[3][4] Against this background and these challenges, this review systematically surveys and compares the principal imputation methods in the scRNA-seq field, discussing the fundamental principles, strengths, weaknesses, and applicable scenarios of each approach, with the aim of providing researchers with a detailed reference guide and offering insights for future methodological improvements and new technological developments.

2. Classification and Principles of Imputation Methods

2.1. Methods Based on Probabilistic Modeling

Model-based methods predict missing values by constructing statistical models or employing matrix factorization techniques that exploit the inherent structural information within the data. The core idea is to fully leverage the similarities among cells, among genes, or both, to provide a rational recovery strategy for missing data. Depending on the type of information utilized, these methods can be divided into three categories:

2.1.1. Cell-based Imputation Methods

These approaches assume that cells of the same type or in similar biological states exhibit similar gene expression patterns. By capturing the similarity among neighboring cells and sharing information among them, missing values are inferred. For instance, MAGIC[5] employs a diffusion mapping algorithm to construct a cell–cell similarity graph and smooth the expression profiles, thereby filling in dropouts; SAVER[6] uses a Bayesian model to integrate information from neighboring cells to estimate the true expression level of each gene, effectively reducing the impact of technical noise.

2.1.2. Gene-based Imputation Methods

These methods are based on the assumption that genes often exhibit co-expression or synergistic behavior within regulatory networks, implying a certain degree of correlation in their expression levels. By exploiting the correlations among genes, one can reconstruct missing data either by building correlation networks or through methods such as non-negative matrix factorization. For example, scNPF[7] integrates gene interaction network information and leverages the co-expression relationships among genes to infer missing values, while netNMF-sc[8] uses network-regularized non-negative matrix factorization to decompose the gene expression matrix into low-dimensional representations of genes and cells, with network constraints ensuring that interconnected genes remain similar in the reduced space, thus recovering the lost data.

2.1.3. Cell & Gene Integrated Imputation Methods

These methods consider both the similarities among cells and the co-expression information among genes by constructing a joint model that simultaneously captures features from both dimensions, thereby enhancing imputation accuracy. Typically, they first distinguish between technical dropouts and true zeros via clustering or preliminary assessments, and then estimate missing values based on local information. For example, scImpute[9] initially uses a statistical model to determine whether a zero count is due to a technical dropout and subsequently imputes the missing value based on the expression levels of similar cells; DrImpute[10] performs multiple rounds of cell clustering and averages the results to yield a robust estimate for missing data; VIPER[11] employs weighted regression to borrow information from a selected subset of similar cells; scTSSR[12] utilizes a sparse self-representation model for both cells and genes to jointly capture the necessary information for reconstructing the expression matrix; SCRABBLE[13] integrates single-cell and bulk RNA-seq data to improve imputation accuracy through external constraints; and SDImpute[14] builds a statistical model based on gene expression data unaffected by dropouts to predict missing values.

2.2. Methods Based on Deep Learning

2.2.1. Autoencoder-based Imputation Methods

These methods construct autoencoder neural networks that compress high-dimensional gene expression data into a low-dimensional latent space and then reconstruct the original data through a decoder. During training, the network automatically learns the nonlinear structure and underlying distribution of the data, effectively recovering missing values caused by dropouts during the reconstruction process. Representative methods include AutoImpute[15], DCA[16], scVI[17], and DeepImpute[18], all of which balance the recovery of subtle changes in expression levels with the preservation of overall structural information.

2.2.2. Graph Neural Network-based Imputation Methods

This category represents single-cell data as a graph where nodes correspond to individual cells and edges denote similarities or proximity between cells. Utilizing graph convolution and other graph neural network techniques, these methods perform information propagation and aggregation on the graph, using information from neighboring cells to jointly recover missing expression values. GraphSCI[19] and scGNN[20] are typical examples; they enhance the imputation effect by leveraging the relational structure among cells, which is especially useful for capturing complex intercellular interactions.

2.2.3. Other Deep Learning and Novel Network Architectures

Beyond autoencoders and graph neural networks, some approaches adopt generative adversarial networks (GANs) or other innovative network architectures to achieve data recovery. These methods typically build generators and discriminators, using adversarial training to enable the generator to produce imputed data that closely matches the true data distribution. For example, scIGANs[21] applies a GAN framework by treating the gene expression matrix as an image and reformulating the imputation task as an image restoration problem; TDimpute[22] utilizes transfer learning and other strategies under novel network architectures to achieve efficient recovery. Such methods demonstrate strong adaptability and recovery capability when handling high-dimensional and complex nonlinear data; scIDPMs[23] utilizes conditional diffusion probabilistic models to impute scRNA-seq data.

3. Advantages, Limitations, and Applicable Scenarios

Model-based methods offer strong interpretability by leveraging explicit statistical assumptions and prior knowledge to elucidate data, and their underlying principles, parameter settings, and model structures are relatively straightforward to understand and adjust. However, these methods require strict assumptions about data distributions, and if the actual data deviate from these assumptions, the imputation performance may suffer. Additionally, they tend to be sensitive to high levels of noise and sparsity and may struggle to capture complex nonlinear relationships. Consequently, they are best suited for scenarios with relatively low noise and high-quality data, such as certain Smart-Seq2 datasets. In contrast, deep learning methods are highly flexible, capable of automatically capturing intricate nonlinear relationships, and demonstrate robust performance in large-scale, high-noise, and high-dimensional datasets without relying on stringent statistical assumptions. Their drawbacks include high computational resource demands, longer training and parameter tuning times, and relatively lower interpretability due to their "black box" nature. These methods are more appropriate for situations involving large-scale data with high noise levels and complex expression patterns (e.g., 10× Chromium datasets) and for downstream tasks that require automatic extraction of underlying data structures, such as cell trajectory reconstruction and large-scale clustering analyses.

4. Conclusion

This review has provided a systematic overview of the principal imputation methods in the scRNA-seq field, categorizing them into model-based and deep learning-based approaches. Model-based methods, which rely on explicit statistical assumptions and the inherent similarities among cells and genes, offer good interpretability and effective data recovery under ideal conditions; however, they may be limited in scenarios with high noise and complex nonlinear structures. Conversely, deep learning methods, employing autoencoders, graph neural networks, and other innovative architectures, can capture complex patterns in large-scale, noisy datasets but come at the cost of higher computational resource consumption and reduced interpretability. In summary, each class of methods has its own strengths and limitations, and researchers should carefully consider the specific characteristics of their data and the requirements of their downstream analyses—such as clustering, differential expression detection, or cell trajectory reconstruction—when selecting an imputation strategy. Future research should focus on exploring hybrid approaches, enhancing model generalization, and addressing missing data recovery in the context of multi-omics integration, with the ultimate goal of advancing single-cell data analysis techniques and providing more precise computational tools for biological investigations.

Acknowledgements

This work was supported by the 2024 Graduate Research Innovation Fund of Yunnan Normal University (Grant No. YJSJJ23-B114).

References

[1] Stevenson K, Uversky V N. Single-cell RNA-Seq: a next generation sequencing tool for a high-resolution view of the individual cell[J]. Journal of Biomolecular Structure and Dynamics, 2020, 38(12): 3730-3735.

[2] Islam S, Zeisel A, Joost S, et al. Quantitative single-cell RNA-seq with unique molecular identifiers[J]. Nature methods, 2014, 11(2): 163-166.

[3] Cheng Y, Ma X, Yuan L, et al. Evaluating imputation methods for single-cell RNA-seq data[J]. BMC bioinformatics, 2023, 24(1): 302.

[4] Dai C, Jiang Y, Yin C, et al. scIMC: a platform for benchmarking comparison and visualization analysis of scRNA-seq data imputation methods[J]. Nucleic Acids Research, 2022, 50(9): 4877-4899.

[5] Van Dijk D, Sharma R, Nainys J, et al. Recovering gene interactions from single-cell data using data diffusion[J]. Cell, 2018, 174(3): 716-729. e27.

[6] Huang M, Wang J, Torre E, et al. SAVER: gene expression recovery for single-cell RNA sequencing[J]. Nature methods, 2018, 15(7): 539-542.

[7] Ye W, Ji G, Ye P, et al. scNPF: an integrative framework assisted by network propagation and network fusion for preprocessing of single-cell RNA-seq data[J]. Bmc Genomics, 2019, 20: 1-16.

[8] Elyanow R, Dumitrascu B, Engelhardt B E, et al. netNMF-sc: leveraging gene–gene interactions for imputation and dimensionality reduction in single-cell expression analysis[J]. Genome research, 2020, 30(2): 195-204.

[9] Li W V, Li J J. An accurate and robust imputation method scImpute for single-cell RNA-seq data[J]. Nature communications, 2018, 9(1): 997.

[10] Gong W, Kwak I Y, Pota P, et al. DrImpute: imputing dropout events in single cell RNA sequencing data[J]. BMC bioinformatics, 2018, 19: 1-10.

[11] Chen M, Zhou X. VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies[J]. Genome biology, 2018, 19(1): 196.

[12] Jin K, Ou-Yang L, Zhao X M, et al. scTSSR: gene expression recovery for single-cell RNA sequencing using two-side sparse self-representation[J]. Bioinformatics, 2020, 36(10): 3131-3138.

[13] Peng T, Zhu Q, Yin P, et al. SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data[J]. Genome biology, 2019, 20: 1-12.

[14] Qi J, Zhou Y, Zhao Z, et al. SDImpute: a statistical block imputation method based on cell-level and gene-level information for dropouts in single-cell RNA-seq data[J]. PLoS Computational Biology, 2021, 17(6): e1009118.

[15] Li W V, Li J J. An accurate and robust imputation method scImpute for single-cell RNA-seq data[J]. Nature communications, 2018, 9(1): 997.

[16] Eraslan G, Simon L M, Mircea M, et al. Single-cell RNA-seq denoising using a deep count autoencoder[J]. Nature communications, 2019, 10(1): 390.

[17] Lopez R, Regier J, Cole M B, et al. Deep generative modeling for single-cell transcriptomics[J]. Nature methods, 2018, 15(12): 1053-1058.

[18] Arisdakessian C, Poirion O, Yunits B, et al. DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data[J]. Genome biology, 2019, 20: 1-14.

[19] Rao J, Zhou X, Lu Y, et al. Imputing single-cell rna-seq data by combining graph convolution and autoencoder neural networks. iScience[J]. 2021.

[20] Wang J, Ma A, Chang Y, et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses[J]. Nature communications, 2021, 12(1): 1882.

[21] Xu Y, Zhang Z, You L, et al. scIGANs: single-cell RNA-seq imputation using generative adversarial networks[J]. Nucleic acids research, 2020, 48(15): e85-e85.

[22] Zhou X, Chai H, Zhao H, et al. Imputing missing RNA-sequencing data from DNA methylation by using a transfer learning–based neural network[J]. GigaScience, 2020, 9(7): giaa076.

[23] Zhang Z, Liu L. scIDPMs: Single-cell RNA-seq imputation using diffusion probabilistic models[J]. IEEE Journal of Biomedical and Health Informatics, 2024.