Quantitative Biology 2023, 11(3): 297–305
https://doi.org/10.15302/J-QB-022-0324

RESEARCH ARTICLE

Tuning hyperparameters of doublet-detection methods for single-cell RNA sequencing data

Nan Miles Xi*, Angelos Vasilopoulos

Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL 60660, USA
* Correspondence: mxi1@luc.edu

Received October 29, 2022; Revised January 13, 2023; Accepted February 5, 2023

***Background*: The existence of doublets in single-cell RNA sequencing (scRNA-seq) data poses a great challenge in downstream data analysis. Computational doublet-detection methods have been developed to remove doublets from scRNA-seq data. Yet, the default hyperparameter settings of those methods may not provide optimal performance.**
***Methods*: We propose a strategy to tune hyperparameters for a cutting-edge doublet-detection method. We utilize a full factorial design to explore the relationship between hyperparameters and detection accuracy on 16 real scRNA-seq datasets. The optimal hyperparameters are obtained by a response surface model and convex optimization.**
***Results*: We show that the optimal hyperparameters provide top performance across scRNA-seq datasets under various biological conditions. Our tuning strategy can be applied to other computational doublet-detection methods. It also offers insights into hyperparameter tuning for broader computational methods in scRNA-seq data analysis.**
***Conclusions*: The hyperparameter configuration significantly impacts the performance of computational doublet-detection methods. Our study is the first attempt to systematically explore the optimal hyperparameters under various biological conditions and optimization objectives. Our study provides much-needed guidance for hyperparameter tuning in computational doublet-detection methods.**

Keywords: scRNA-seq; doublet detection; hyperparameter tuning; experimental design; response surface model

Author summary: Doublet is a major confounder in single-cell RNA sequencing data analysis. Computational doublet-detection methods aim to remove doublets from scRNA-seq data. The performance of those methods relies on the appropriate setting of their hyperparameters. In this study, we explore the optimal hyperparameters for scDblFinder, a cutting-edge doublet-detection method. Our optimization utilizes a full factorial design, a response surface model, and 16 real scRNA-seq datasets. The optimal hyperparameters achieve top doublet-detection performance under a wide range of biological conditions. Our methodology is applicable to broader computational methods in scRNA-seq data analysis.

INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) is a cutting-edge sequencing technology that can quantify genome-wide gene expression levels in a large number of cells [1,2]. Since its debut, scRNA-seq has been widely applied in various fields, including precision medicine [3], drug discovery [4], cancer therapy [5], and vaccine development [6]. The successful application of scRNA-seq relies on separating and labeling mRNA molecules from different cells. However, results may be confounded by the formation of doublets — when two cells are captured in one reaction volume by chance [7]. Because doublets appear as but are not real cells, they potentially bias downstream scRNA-seq data analysis. For example, doublets may be falsely identified as new cell types in cell clustering analysis [8]. To tackle this issue, the scRNA-seq community has developed computational methods to detect doublets from

scRNA-seq data [7–11]. These methods utilize statistical and machine learning models, each with a set of default hyperparameters. Despite rapid development in methodology, one critical question remains untouched: whether default hyperparameters offer the best doublet-detection performance, especially for scRNA-seq datasets generated under various biological conditions.

Here, we systematically explore the optimal hyperparameters of scDblFinder [9], one cutting-edge computational doublet-detection method. We collect detection accuracy data for various hyperparameter combinations from 16 real scRNA-seq datasets with experimentally annotated doublets. Then, we fit a second-degree polynomial regression model with first-order, second-order, and interaction terms of three key hyperparameters. Convex optimization is performed to find the hyperparameters that maximize average detection accuracy across 16 datasets. We show that our optimal hyperparameters significantly improve doublet-detection accuracy over the method's default settings on most datasets. The detection accuracy of our optimal hyperparameters also ranks close to the best performance obtained by exhaustive searches in many datasets. We also apply our tuning strategy to scRNA-seq datasets under various biological conditions using different double-detection measurements. We find similar benefits from hyperparameter tuning, and the optimal hyperparameters vary depending on the biological conditions and accuracy measurements. Our exploratory strategy can be easily extended to other computational doublet-detection methods and provides hyperparameter tuning insights for broader computational methods in scRNA-seq data analysis.

RESULTS

Overall optimal hyperparameter evaluation

We consider three key hyperparameters of scDblFinder, *i.e.*, the number of top features, the number of top principal components, and the maximum depths of decision trees. We refer to them as nf, pc, and $depth$ moving forward, respectively. The optimal hyperparameters are obtained by maximizing the average area under the precision-recall curve (AUPRC) of doublet detection across 16 scRNA-seq datasets (see methods). To examine if these parameters can improve doublet-detection accuracy on individual datasets, we execute scDblFinder with optimal nf and pc on all 16 scRNA-seq datasets. Since $depth$ is not significant, we set it to the default value in the execution. Table 1 compares the AUPRCs of the optimal hyperparameters, the method's default hyperparameters, and the maximal AUPRCs achieved by one of 125 hyperparameter combinations. Our optimal hyperparameters outperform the method's performance with default settings on 12 out of 16 scRNA-seq datasets. Figure 1 summarizes the AUPRC improvement by hyperparameters tuning over the method's default settings. The most significant improvement is over 5% on dataset pbmc-1B-dm. There are

Table 1 The AUPRC of doublet detection under optimal and default hyperparameters

Dataset	Optimum	Default	Maximum
cline-ch	0.4280	0.4202	0.4369
HEK-HMEC-MULTI	0.4723	0.4830	0.4966
HEK-orig-MULTI	0.4911	0.4873	0.5054
hm-12k	0.9281	0.9506	0.9850
hm-6k	0.9737	0.9896	0.9972
HMEC-rep-MULTI	0.6010	0.5964	0.6020
J293t-dm	0.2052	0.1999	0.2525
mkidney-ch	0.6125	0.6080	0.6183
nuc-MULTI	0.4600	0.4430	0.4704
pbmc-1A-dm	0.5454	0.5441	0.5693
pbmc-1B-dm	0.4375	0.4145	0.4818
pbmc-1C-dm	0.5953	0.5744	0.6082
pbmc-2ctrl-dm	0.6980	0.6749	0.7088
pbmc-2stim-dm	0.7003	0.6763	0.7124
pbmc-ch	0.6405	0.6472	0.6520
pdx-MULTI	0.4457	0.4268	0.4477

The last column shows the highest AUPRC achieved by one of the 125 hyperparameter combinations. The highest AUPRC between the optimum and default of each dataset is underscored.

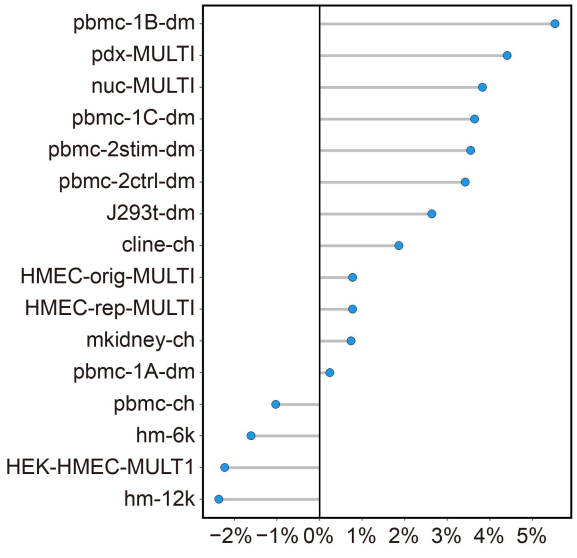

Figure 1. AUPRC improvement by hyperparameter optimization over the method's default settings on 16 scRNA-seq datasets.

eight datasets on which the improvement is over 2%. Figure 2 shows each dataset's AUPRC ranking under optimal hyperparameters among 125 hyperparameter combinations. We can see that the AUPRCs of optimal hyperparameters rank at or higher than the 20th percentile on ten datasets. The highest ranking is 3rd on dataset pdx-MULTI. The optimal hyperparameters also achieve the 50th percentile or higher on all 16 datasets.

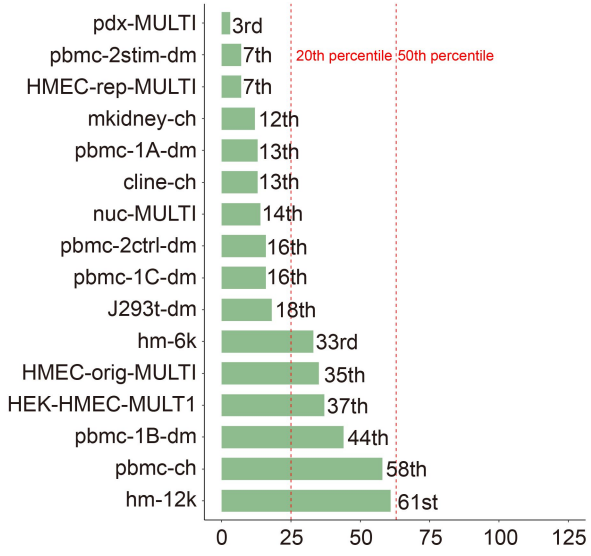

Figure 2. Rankings of optimal hyperparameter performances by AUPRC among 125 hyperparameter combinations on 16 datasets.

Tuning hyperparameters for various biological conditions

The previous analysis presents the optimal hyperparameters based on the average of 16 scRNA-seq datasets. In practice, users mainly conduct doublet detection on datasets generated under specific biological or technical conditions. Those datasets need unique hyperparameter settings to achieve optimal performance. To demonstrate the generability of our tuning strategy to those applications, we replicate the hyperparameter optimization on two subsets of 16 datasets.

The first is the pbmc-related subset, including six datasets: pbmc-1A-dm, pbmc-1B-dm, pbmc-1C-dm, pbmc-2ctrl-dm, pbmc-2stim-dm, and pbmc-ch. We find that the optimal *pc* is 18 (after rounding to the nearest integer), with the other two hyperparameters insignificant (Supplementary Table S1). Supplementary Table S2 compares the AUPRCs of the optimal hyperparameters, the method's default hyperparameters, and the maximal AUPRCs achieved by one of 125 hyperparameter combinations on pbmc-related datasets. Supplementary Fig. S1A shows each dataset's AUPRC improvement by hyperparameters optimization over the method's default settings, and AUPRC ranking under optimal hyperparameters among 125 hyperparameter combinations. Compared with the optimization on all 16 datasets (Figs. 1 and 2), the AUPRC improvement is greater with hyperparameters specifically tuned for pbmc-related datasets.

The second subset includes the three HMEC-related datasets: HMEC-orig-MULTI, HMEC-rep-MULTI, and HEK-HMEC-MULTI. We find that the optimal nf is 1520 (after rounding to the nearest integer), with the other two hyperparameters insignificant (Supplementary Table S1). Supplementary Table S3 compares the AUPRCs of the optimal hyperparameters, the method's default hyperparameters, and the maximal AUPRCs achieved by one of 125 hyperparameter combinations on HMEC-related datasets. Supplementary Fig. S1B shows each dataset's AUPRC improvement by hyperparameters optimization over the method's default settings, and AUPRC ranking under optimal hyperparameters among 125 hyperparameter combinations. Compared with the optimization on all 16 datasets (Figs. 1 and 2), the AUPRC improvement is greater with hyperparameters specifically tuned for HMEC-related datasets.

The two analyses provide guidance for choosing appropriate hyperparameters for specific biological conditions. Future studies can easily expand our tuning strategies to other cell types or platforms if datasets with doublet labels under more diverse biological and technical conditions are available. For example, users can optimize hyperparameters for different sequencing protocols (Smart-seq2, Drop-seq, Chromium, etc.) or the combinations of biological and technical conditions (pbmc and Drop-seq, HMEC and Smart-seq2, etc.).

Tuning hyperparameters for various measurements

The previous analyses use AUPRC, an overall accuracy measurement of doublet detection, as the optimization objective. In practice, users may also be interested in the method's capacity to identify doublets or singlets, *i.e.*, the true positive or negative rate. Different from AUPRC, the calculation of true positive/negative rate requires a user-specified doublet rate. The true doublet rate is typically unknown to the users and needs to be estimated based on the sequencing platform, sequencing throughput, and prior knowledge [12,13]. Because optimization relies on the doublet rate, it is infeasible to find universal optimal hyperparameters for the true positive/negative rate.

To provide hyperparameter guidance under this scenario, we set the doublet rates to their true values for each dataset (Table 2) and calculate the corresponding

Table 2 The 16 scRNA-seq datasets with experimentally annotated doublets in this study

Dataset	Cell type	Droplet #	Gene #	Doublet rate	Doublet annotation technique
pbmc-ch	pbmc	15,272	21,639	16.66%	Cell hashing [14]
cline-ch	HEK293T, K562, KG1, THP1	7954	25,221	18.42%	Cell hashing
mkidney-ch	mouse kidney	21,179	18,940	37.31%	Cell hashing
hm-12k	HEK293T, NIH3T3	12,820	15,106	5.69%	Species mixture [15]
hm-6k	HEK293T, NIH3T3	6806	15,080	2.51%	Species mixture
pbmc-1A-dm	pbmc	3298	15,170	3.64%	Demuxlet [16]
pbmc-1B-dm	pbmc	3790	15,143	3.43%	Demuxlet
pbmc-1C-dm	pbmc	5270	15,865	6.00%	Demuxlet
pbmc-2ctrl-dm	pbmc	13,913	17,584	11.49%	Demuxlet
pbmc-2stim-dm	pbmc	13,916	17,315	11.72%	Demuxlet
J293t-dm	jurkat, HEK293T	500	16,374	8.40%	Demuxlet
pdx-MULTI	human breast cancer, mouse immune	10,296	14,025	12.79%	MULTI-seq [17]
HMEC-orig-MULTI	HMEC	26,426	24,199	13.50%	MULTI-seq
HMEC-rep-MULTI	HMEC	10,580	17,473	31.02%	MULTI-seq
HEK-HMEC-MULTI	HEK293T, HMEC	10,641	23,982	4.60%	MULTI-seq
nuc-MULTI	nuclei (HEK293T, MEF, Jurkat)	5578	21,490	8.52%	MULTI-seq

true positive/negative rates for 125 hyperparameter combinations. Then we replicate our tuning strategy using these two measurements as objectives. We find that the optimal maximum depth of decision trees (*depth*) is 5 for both measurements (after rounding to the nearest integer), with the other two hyperparameters insignificant (Supplementary Table S1). Supplementary Fig. S2A and Table S4 show each dataset's true positive rate improvement by hyperparameters optimization over the method's default settings and true positive rate ranking under optimal hyperparameters among 125 hyperparameter combinations. Most datasets exhibit similar improvement as AUPRC, except for J293t-dm, with a 29% increase, significantly larger than others. Such a difference indicates this dataset's unique biological and technical characteristics, which require stronger hyperparameter tuning efforts.

Supplementary Fig. S2B and Table S5 show each dataset's true negative rate improvement and ranking. Although most datasets still benefit from hyperparameter tuning, the improvement of the true negative rate is milder (below 1%) compared to other metrics. One reason is that the true negative rates under default hyperparameters are already high on many datasets (above 0.95, Supplementary Table S5), limiting the improvement space by hyperparameter optimization. It is worth noting that the optimal hyperparameters and corresponding true positive/negative rates are obtained using the true doublet rates. If users choose different doublet rates, the optimization results will be different. It is straightforward to generalize our tuning strategy in those cases.

DISCUSSION

The existence of doublets is a key confounder in scRNA-seq data analysis. With the wide application of scRNA-seq technologies, much effort has been invested in developing computational doublet-detection methods. Such methods are primarily based on statistical and machine learning algorithms and are sensitive to hyperparameter configurations [18]. Although most methods provide a set of default hyperparameters, they cannot guarantee the best doublet-detection performance universally, especially when scRNA-seq data are generated under various biological conditions [19,20].

In this study, we utilize a full factorial design to build a model of hyperparameters and overall doublet-detection accuracy based on a leading method, scDblFinder, and 16 scRNA-seq datasets. The optimal hyperparameter combination obtained by convex optimization not only surpasses the default setting but also offers close-to-best detection accuracy on many datasets. We expand our optimization strategy to two subgroups of 16 datasets separately, providing optimal parameters for various biological conditions. We show that our method can also be applied to optimize different measurements of doublet-detection accuracy.

The improved doublet-detection performance by hyperparameter tuning presents several insights regarding the data generalization and doublet annotation mechanisms. First, there are two datasets, hm-6k and hm-12k, whose doublets are annotated by species mixture [15]. Both have lower AUPRCs using optimal hyperparameters than default hyperparameters (Fig. 1). In contrast, most

datasets generated by the other three doublet annotation techniques, *i.e.*, cell hashing [14], demuxlet [16], and MULTI-seq [17], benefit from hyperparameter tuning. One possible reason is due to their different doublet-annotation mechanisms. While species mixture can only annotate doublets from two species, the other three techniques utilize oligo-tagged antibody, SNP, or lipid-tagged index to label doublets from much broader sources. Consequently, the true doublets in hm-6k and hm-12k are likely undercounted, causing their inconsistent optimization results.

Second, the hyperparameter tuning fails to improve the AUPRC for dataset pbmc-ch, even if the hyperparameters are optimized specifically for pbmc-related datasets (Supplementary Fig. S1). In contrast, optimal hyperparameters consistently improve AUPRC for the other five pbmc-related datasets, and the improvements are greater with specifically tuned hyperparameters (Fig. 1 and Supplementary Fig. S1). Such discrepancy is potentially due to the different doublet annotation techniques (cell hashing vs. demuxlet) and batch effects among those datasets. Further investigations, especially from the experimental perspective, are needed to reveal the impacts of these two factors on doublet detection.

Third, the optimal hyperparameters vary depending on the biological conditions and optimization objectives. There are no universal hyperparameters adaptive to all scenarios. The significant hyperparameters when optimizing AUPRC across all 16 datasets are *pc* and *nf*, with optimal values as 19 and 1252, respectively (Supplementary Table S1). If optimized on pbmc-related datasets, *nf* is no longer significant, and the optimal value of pc changes to 18. If optimized on HMEC-related datasets, *pc* is no longer significant, and the optimal value of *nf* changes to 1520. *Depth* is the only significant hyperparameter when optimizing the true positive and negative rate on all 16 datasets, with optimal values as 5 in both cases. This result indicates that existing and future doublet-detection methods need to fine-tune hyperparameters for a variety of biological conditions and accuracy measurements.

CONCLUSION

In summary, doublet detection is one essential step in the quality control of scRNA-seq data analysis. The hyperparameter configuration significantly impacts the performance of computational doublet-detection methods. Our study is the first attempt to systematically explore the optimal hyperparameters under various biological conditions and optimization objectives. Our study provides much-needed guidance for hyperparameter tuning in computational doublet-detection methods. Future directions of our study include increasing the exploratory space by utilizing advanced experimental design strategies, *e.g.*, space-filling design [21,22] and fractional factorial design [23,24]. Another direction is to optimize hyperparameters for other doublet-detection methods by our tuning strategy. More scRNA-seq datasets with experimentally annotated doublets could also be incorporated into the tuning process to enhance the generalizability of optimal hyperparameters.

METHODS

Datasets

In this study, we utilize 16 public scRNA-seq datasets collected in a previous benchmark study [12]. All datasets contain ground-truth doublet labels annotated by experimental techniques. They are so far the most comprehensive scRNA-seq data collection with ground-truth doublet labels. The datasets cover a wide range of cell types, droplet numbers, gene numbers, and doublet rates, representing various difficulty levels in detecting doublets from scRNA-seq data. Table 2 summarizes the key characteristics of the 16 datasets used in this study. In scRNA-seq experiments, droplets refer to the reaction volumes that encapsulate the cell suspension. While most droplets contain one cell (singlets) as expected, others accidentally encapsulate two cells (doublets). Therefore, we will use “droplet” instead of “cell” to denote one data point in scRNA-seq datasets in the following text.

Hyperparameter setting

We choose scDblFinder as the target method for exploring optimal hyperparameter settings. The design of scDblFinder can be summarized in the following steps. First, it generates artificial doublets by combining gene expression profiles of two randomly selected droplets in the dataset. Second, it conducts PCA dimension reduction on the union of artificial doublets and true droplets using top highly variable genes. Third, scDblFinder constructs a nearest neighbor network on top of the low-dimensional representations from the dimension reduction. Fourth, it sets different neighborhood sizes to create multiple predictors that will be used for binary classification. Finally, it performs cross-validation to assign a doublet score to each droplet. In each iteration of cross-validation, it trains a gradient boosting model to distinguish true droplets and artificial doublets in the training set, and then assigns each droplet in the test set a doublet probability (doublet score). The design of scDblFinder helps to reduce the

impact of batch effects on doublet detection: since cross-validation randomly assigns droplets from different batches to training and test sets, the batch effects will not cause the domain shift problem [25] in the final classification step.

scDblFinder has shown superior performance in previous benchmark studies [12,26]. We consider three key hyperparameters of scDblFinder, including the number of top features (nf), the number of top principal components (pc), and the maximum depths of decision trees ($depth$). These three hyperparameters are discrete numerical variables, and we set each of them to five different levels (Table 3). Therefore, there are $5\times5\times5=125$ hyperparameter combinations in total. In experimental design literature, this is a 3^5 full factorial design [27]. It allows investigation of the effects of individual hyperparameters, as well as the effects of interactions between different hyperparameters on the performance of doublet detection.

Table 3 The three hyperparameters and their default and exploratory values in this study

Hyperparameter	Default values	Exploratory values
nf	1000	500, 1000, 1500, 2000, 2500
pc	10	5, 10, 15, 20, 25
depth	4	2, 3, 4, 5, 6

We choose the five values for each hyperparameter according to the following rule. First, we start with the default values of each hyperparameter (Table 3). In general, the default values are selected by the developers based on extensive numerical experiments and thus are likely close to a local optimum in the hyperparameter space. Second, with the default value as the center, we increase or decrease each hyperparameter by one or several fixed step sizes, generating four extra alternative values. We determine the step size and the search space boundaries for each hyperparameter based on the common practice in scRNA-seq data analysis.

The hyperparameter nf refers to the number of highly variable genes used in the downstream analysis. Its value is often set from several hundred to several thousand in many applications. For example, scDblFinder uses 1000 as the default value and Seurat [28], a popular R package for scRNA-seq data analysis, chooses 2000. After including these two values in our search space, we insert 1500, the median value between 1000 and 2000, as the third search value, resulting in a step size of 500. We further expand the search space downward and upward by one step size separately. The final search space for nf is 500, 1000, 1500, 2000, and 2500.

The hyperparameter pc is the number of principal components used in the downstream analysis after performing PCA dimension reduction on highly variable genes. Its value is often set from single digits to several dozen in practice. For example, the Seurat tutorial suggests exploring between 5 to 20 for various scenarios. We start with the default value of 10 and include 5 (the lower bound suggested by Seurat) in the search space, using a step size of 5. We further expand the search space by three step sizes up to 25. The final search space for pc is 5, 10, 15, 20, and 25.

The hyperparameter $depth$ is the maximum depth of decision trees in the gradient boosting model used in scDblFinder. The larger values indicate more complex gradient boosting models in binary classification (singlet vs. doublet). This hyperparameter is often set to below ten in ordinary classification tasks to avoid overfitting. For example, XGBoost [29], a generic gradient boosting package, chooses 6 as the default value. We use scDblFinder's default value of 4 as the center of the search space. With 6 as the maximum and 1 as the step size, we create a final search space for $depth$ including 2, 3, 4, 5, and 6.

Doublet detection

We use the R package DoubletCollection [26] to execute scDblFinder on 16 real datasets with the 125 hyperparameter combinations listed in Table 3. Since doublet detection is essentially a binary classification task, we use AUPRC to measure the overall doublet-detection accuracy. After execution, each dataset results in a 125×4 data matrix, in which the first three columns are *nf*, *pc*, and *depth*, and the last column is AUPRC. Each row in the data matrix represents one combination of three hyperparameters and corresponding AUPRC. The 16 scRNA-seq datasets generate 16 such data matrices. Finally, we merge the 16 data matrices by averaging their AUPRCs for each hyperparameter combination. The final data matrix is 125×4, which contains the relationship between hyperparameters and overall doublet-detection accuracy. We refer to this data matrix as detection accuracy data moving forward.

Model setup and optimization

We build a second-degree polynomial regression model to examine the relationship between hyperparameters and doublet-detection accuracy. Specifically, we set average AUPRC as the response variable and the first order of the three hyperparameters, the second order of the three hyperparameters, and their interactions as the independent variables. Model (1) shows the complete setup of this second-degree polynomial regression.

$$AUPRC = \beta_0 + \beta_1 nf + \beta_2 pc + \beta_3 depth + \beta_4 nf^2 + \beta_5 pc^2 + \beta_6 depth^2 + \beta_7 nf\,pc + \beta_8 nf\,depth + \beta_9 pc\,depth + \epsilon \quad (1)$$

where $\beta_i, i = 0, 1, ..., 9$, are the unknown model parameters, and is the random error.

Second-degree polynomial regression is one classic model in the response surface methodology (RSM) [30]. It is commonly used to explore the relationship between several independent variables (hyperparameters) and one response variable (AUPRC) based on a full factorial design [31]. It can obtain an optimal response by estimating hyperparameters' main and quadratic effects and interactions between them. The second-degree polynomial regression balances model complexity and interpretation, while higher-degree models may cause overfitting and are harder to interpret.

We fit this model by least square estimation using detection accuracy data. We perform a t-test to assess the significance of estimated parameters $\widehat{\beta_i}$ and set 0.01 as the p-value cutoff. We find that the first and second orders of nf and pc are significant. Equation (2) shows the estimated model (1) with significant independent variables (including the intercept).

$$AUPRC = 5.444 \times 10^{-1} + 1.016 \times 10^{-5} nf + 1.336 \times 10^{-3} pc - 3.760 \times 10^{-9} nf^2 - 3.484 \times 10^{-5} pc^2 \quad (2)$$

To obtain the nf and pc that maximize AUPRC, we take the partial derivative of AUPRC in respect of nf and pc in Eq. (2) and let the derivatives equal zero simultaneously.

$$\begin{cases} \dfrac{\partial AUPRC}{\partial nf} = 1.016 \times 10^{-5} - 7.520 \times 10^{-9} nf = 0 \\ \dfrac{\partial AUPRC}{\partial pc} = 1.336 \times 10^{-3} - 6.968 \times 10^{-5} pc = 0 \end{cases} \quad (3)$$

Solving Eq. (3) gives the optimal nf and pc (after rounding to the nearest integers).

$$\begin{cases} nf = 1352 \\ pc = 19 \end{cases} \quad (4)$$

Model diagnostics

The 16 scRNA-seq datasets are generated by different sequencing protocols using various doublet annotation techniques. The error terms in model (1) may have non-constant variance, causing the heterogeneity issue. We conduct model diagnostics to examine the existence and severity of heterogeneity. First, we plot the residue against the fitted value of model (1). Supplementary Fig. S3A shows that most residues have constant variance with no obvious patterns. The only concern is on the left, where the six residues may have a smaller variance. Second, we perform a Breusch-Pagan test [32] for heterogeneity. With the p-value as 0.451, we fail to reject the null hypothesis that constant variance is present.

Additionally, we perform a sensitivity analysis to examine the robustness of model (1). We conduct a natural log transformation and a square root transformation on the response variable AUPRC in the detection accuracy data. We then fit model (1) on the two transformed datasets and obtain the optimal hyperparameters using the same optimization method in Eq. (3). We find that the significant hyperparameters in model (1) and their optimal values are identical (after rounding to the nearest integers) to the results without transformation (Supplementary Table S6). The patterns of residue plots (Supplementary Figs. S3B and C) are also similar to those before transformation (Supplementary Fig. S3A). Log transformation and square root transformation on the response variable are common remedies for heterogeneity. If heterogeneity exists, then these two transformations would significantly change the model fitting, optimization, and residue plot. Similar results before and after transformations indicate that the heterogeneity is very mild, if any. We suspect that the heterogeneity is largely reduced or removed by averaging the AUPRCs of 16 datasets when creating the detection accuracy data (on which we fit model (1)).

Data availability

The 16 scRNA-seq datasets used in this study are available at Zenodo repository (DOI: 4562782)

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at https://doi.org/10.15302/J-QB-022-0324.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to Dr. Lin Wang at Purdue University Department of Statistics for generously sharing her expert insights and knowledge regarding statistical analysis.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Nan Miles Xi and Angelos Vasilopoulos declare that they have no conflict of interest or financial conflicts to disclose.

This article does not contain any studies with human or animal materials performed by any of the authors

REFERENCES

1. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. and Teichmann, S. A. (2015) The technology and biology of single-cell RNA sequencing. Mol. Cell, 58, 610−620
2. Saliba, A.-E., Westermann, A. J., Gorski, S. A. and Vogel, J. (2014) Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res., 42, 8845−8860
3. Wiedmeier, J. E., Noel, P., Lin, W., Von Hoff, D. D. and Han, H. (2019) Single-cell sequencing in precision medicine. Cancer Treat. Res., 178, 237−252
4. Aissa, A. F., Islam, A. B. M. M. K., Ariss, M. M., Go, C. C., Rader, A. E., Conrardy, R. D., Gajda, A. M., Rubio-Perez, C., Valyi-Nagy, K., Pasquinelli, M., *et al*. (2021) Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. Nat. Commun., 12, 1628
5. Sun, G., Li, Z., Rong, D., Zhang, H., Shi, X., Yang, W., Zheng, W., Sun, G., Wu, F., Cao, H., *et al*. (2021) Single-cell RNA sequencing in cancer: applications, advances, and emerging challenges. Mol. Ther. Oncolytics, 21, 183−206
6. Noé, A., Cargill, T. N., Nielsen, C. M., Russell, A. J. C. and Barnes, E. (2020) The application of single-cell RNA sequencing in vaccinology. J. Immunol. Res., 2020, 8624963
7. Wolock, S. L., Lopez, R. and Klein, A. M. (2019) Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Cell Syst., 8, 281−291.e9
8. McGinnis, C. S., Murrow, L. M. and Gartner, Z. J. (2019) DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst., 8, 329−337.e4
9. Germain, P.-L., Lun, A., Garcia Meixide, C., Macnair, W. and Robinson, M. D. (2021) Doublet identification in single-cell sequencing data using *scDblFinder*. F1000 Res., 10, 979
10. Bais, A. S. and Kostka, D. (2019) scds: computational annotation of doublets in single-cell RNA sequencing data. Bioinformatics, 15, 1150−1158
11. Bernstein, N. J., Fong, N. L., Lam, I., Roy, M. A., Hendrickson, D. G. and Kelley, D. R. (2020) Solo: doublet identification in single-cell RNA-seq via semi-supervised deep learning. Cell Syst., 11, 95−101.e5
12. Xi, N. M. and Li, J. J. (2021) Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. Cell Syst., 12, 176−194.e6
13. Luecken, M. D. and Theis, F. J. (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol., 15, e8746
14. Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B. Z., Mauck, W. M. 3rd, Smibert, P. and Satija, R. (2018) Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol., 19, 224
15. Alles, J., Karaiskos, N., Praktiknjo, S. D., Grosswendt, S., Wahle, P., Ruffault, P.-L., Ayoub, S., Schreyer, L., Boltengagen, A., Birchmeier, C., *et al*. (2017) Cell fixation and preservation for droplet-based single-cell transcriptomics. BMC Biol., 15, 44
16. Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C. M., *et al*. (2018) Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat. Biotechnol., 36, 89−94
17. McGinnis, C. S., Patterson, D. M., Winkler, J., Conrad, D. N., Hein, M. Y., Srivastava, V., Hu, J. L., Murrow, L. M., Weissman, J. S., Werb, Z., *et al*. (2019) MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat. Methods, 16, 619−626
18. Probst, P., Boulesteix, A. L. and Bischl, B. (2019) Tunability: importance of hyperparameters of machine learning algorithms. arXiv,1802.09596
19. Hu, Q. and Greene, C. S. (2019) Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. Pac. Symp. Biocomput., 24, 362−373
20. Raimundo, F., Vallot, C. and Vert, J.-P. (2020) Tuning parameters of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biol., 21, 212
21. Wang, L., Xiao, Q. and Xu, H. (2018) Optimal maximin L_1-distance Latin hypercube designs based on good lattice point designs. Ann Stat., Ann. Statist. 46, 3741–3766
22. Wang, L., Sun, F., Lin, D. K. J. and Liu, M.-Q. (2018) Construction of orthogonal symmetric Latin hypercube designs. Stat. Sin., 28, 1503−1520
23. Wang, L., Xu, H. and Liu, M.-Q. (2022) Fractional factorial designs for Fourier-cosine models. Metrika, 86, 373−390
24. Wang, L. and Xu, H. (2022) A class of multilevel nonregular designs for studying quantitative factors. Stat. Sin., 32, 825−845
25. Redko, I., Morvant, E., Habrard, A., Sebban, M. and Bennani, Y. (2019) Advances in Domain Adaptation Theory. Amsterdam: Elsevier
26. Xi, N. M. and Li, J. J. (2021) Protocol for executing and benchmarking eight computational doublet-detection methods in single-cell RNA sequencing data analysis. STAR Protoc., 2, 100699
27. Steinberg, D. M. and Hunter, W. G. (1984) Experimental design: review and comment. Technometrics, 26, 71−97

28. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M. 3rd, Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., *et al*. (2021) Integrated analysis of multimodal single-cell data. Cell, 184, 3573−3587.e29
29. Chen, T. and Guestrin, C. (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794
30. Box, G. E. P. and Draper, N. R. (1987) Empirical Model-building and Response Surfaces. New York: John Wiley & Sons
31. Box, G. E. P. and Wilson, K. B. (1951) On the experimental attainment of optimum conditions. J. R. Stat. Soc. Series B Stat. Methodol., 13, 1−45
32. Breusch, T. and Pagan, A. (1979) A simple test for heteroscedasticity and random coefficient variation. Econometrica, 47, 1287−1294

