Methodological contributions to the challenges and opportunities of high dimensional clustering in the context of single-cell data

Cornelia Sigrid Fütterer

München 2022

Methodological contributions to the challenges and opportunities of high dimensional clustering in the context of single-cell data

Cornelia Sigrid Fütterer

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität München

vorgelegt von
Cornelia Sigrid Fütterer
aus München

München, den 01. April 2022

Erstgutachter: Prof. Dr. Thomas Augustin
Zweitgutachterin: Prof. Donna Ankerst, Ph.D.
Drittgutachterin: Univ. Prof. Dr. rer. nat. HDR Anne-Laure Boulesteix
Tag der Einreichung: 01. April 2022
Tag der Disputation: 28. Juni 2022

Acknowledgement

This thesis has been developed during my time as a PhD student at the working group "Foundations of Statistics and their Applications" at the Institute of Statistics of the Ludwig-Maximilians-Universität München. My special thanks go to ...

- *... Thomas Augustin, who gave me the opportunity to find my own research topic and the freedom to develop methods in the subjects I was burning for. He actively supported my projects and was always ready to critically discuss my approaches, no matter what day or time. With my creativity and his talent for generalization, I highly appreciated the work on our joint research projects. Thank you very much, Thomas, for your trust, great supervision, collaboration, proofreading, and guidance, which has greatly contributed to my further development.*
- *... Christiane Fuchs, to whom I owe my interest in the research field of single-cell data. She supervised my master thesis and introduced me to current research projects and researchers at the Helmholtz-Zentrum München. I really enjoyed getting an insight into Christiane's perspectives on research, as well as her encouragement to write a thesis. I appreciate a lot that it was possible to do a joint research project with her and Thomas, thus having two critical experts on board, which resulted in focused work.*
- *... Donna Ankerst and Anne-Laure Boulesteix for their interest in my research, as well as for their willingness to serve as my external reviewers and Christian Heumann and Volker Schmid for being part of the examination committee.*
- *... my coauthors Thomas Augustin, Christiane Fuchs, Malte Nalenz, and Georg Schollmeyer for the nice collaboration in research as well as in teaching.*
- *... the members of Society for Imprecise Probabilities, especially to Thomas Augustin, Cassio de Campos, Enrique Miranda, Ignacio Montes, Teddy Seidenfeld, and Matthias Troffaes, who made me feel immediately at home at the summer school in Oviedo and the conference in Ghent.*
- *... Jan Baumbach and Dominik Heider for the stimulating discussions at the European Conference on Data Analysis in Bayreuth.*
- *... Francesca Biagini and Markus Heydenreich, my mentors in the LMUMentoring program. Thank you for the connection with experienced researchers and the individual conversations with me as well as the financial support that gave me the opportunity to present the work of myself and my co-authors at conferences and summer schools. The support during Covid 19 with technical equipment and personal development workshops allowed for straightforward exchanges from home.*
- *... Gerhard for his interest in my research progress. I appreciate a lot his attitude of listening to new ideas without judging or rejecting them from the first moment on, no matter how magical they sometimes sounded. As a good statistician after having built his opinion he asked critical questions to challenge my ideas. It was really nice having discussions with such an open-minded biostatistician for whom it is also always important to keep the application of methods in mind. If we have learned one thing, then it is that there are still many things to be discovered in this world.*

... *all former and current colleagues of the working group for the great discussions about research topics, teaching as well as students projects. Especially the common lunch time and coffee breaks made my time at the institute very enjoyable! Not only the events and conferences with the working group have been unique, but also the uncomplicated cooperation with all members of the institute, who were always willing to participate in events, such as Eisbach sessions, poker, skiing or squash.*

... *to my former colleague Julia and now best neighbor, as well as my dance teacher and my dancing crew, who always kept me in good spirits especially during the home office time because of Covid 19. Thanks to all the fun walks and continuing dancing throughout the lockdown in my living room that allowed me to stay positive in my mind.*

... *the support of my friends who mean a lot to me. They were always there for me and suggested or agreed to a vacation at the right moment. Thanks a lot for the common and stimulating moments! I would also like to thank my friends in France, especially David for his trust and encouragement in my projects. Merci beaucoup!*

... *my grandparents, my parents and my brother to whom I owe my ability to persevere. Thank you very much for all the encouraging comments and your trust in me that I will not only make it but also do it right.*

Zusammenfassung

Mit der Sequenzierung von Einzelzellen ist es möglich, die Genexpression jeder einzelnen Zelle zu messen, im Gegensatz zur Massensequenzierung, die nur eine Messung der durchschnittlichen Genexpression ermöglicht. Eine Kenntnis der Genexpression der einzelnen Zellen ermöglicht, dass darauf basierend Methoden aufgebaut werden können, um eine automatisierte Zuteilung von Einzelzellen zu Zelltypen vorzunehmen. Die Bestimmung von Zelltypen ist entscheidend für die Analyse von Krankheiten und für das Verständnis der menschlichen Gesundheit basierend auf dem genetischen Profil einzelner Zellen. Üblicherweise werden Zelltypen mithilfe von Clustering-Verfahren zugeordnet, die speziell für Einzelzelldaten entwickelt worden sind. Zu diesem Zweck wird beispielsweise das *single-cell consensus clustering (SC3)* verwendet, welches von Kiselev et al. (Nat Methods 14(5):483-486, 2017) vorgeschlagen wurde. Dieses gehört zu den führenden Clustering-Verfahren von Einzelzelldaten und wird auch für die folgenden Beiträge von Bedeutung sein.

Diese kumulative Dissertation zielt auf die Entwicklung geeigneter Analysetechniken für das Clustering hochdimensionaler Einzelzelldaten und deren zuverlässige Validierung ab. Außerdem wird ein Simulationsrahmen für die Untersuchung des Einflusses verzerrter Messungen von Einzelzellen auf die Clustering Performance bereitgestellt. Darüber hinaus werden Clusterindizes als informative Gewichte in die regularisierte Regression einbezogen, was als weicher Genfilter betrachtet werden kann.

Beitrag 1 verbessert die Anpassung des ursprünglichen SC3s an den zugrundeliegenden biologischen Prozess der Übergänge von Einzelzellen. Wir schlagen das unüberwachte *'adapted single-cell consensus clustering (adaSC3)"* vor, das die Hauptkomponentenanalyse des ursprünglichen SC3s durch sogenannte Diffusion Maps ersetzt, die den Übergang einzelner Zellen berücksichtigen. Daher respektiert adaSC3 nicht nur methodisch den biologischen Prozess der einzelnen Zellen, sondern verbessert auch die Genauigkeit. Wir evaluieren die Genauigkeit von SC3, adaSC3 und einigen konkurrierenden Methoden sowohl auf Einzelzell-RNA-Sequenzierungsdaten als auch auf Simulationsdaten, welche alle denkbaren Kombinationen von Partitionen bei gleicher Stichprobenanzahl enthalten. Für alle durchgeführten Studien können wir eine überzeugende Leistungsfähigkeit von adaSC3 feststellen.

Beitrag 2 schlägt eine *Assoziations-Genauigkeits-Heuristik* vor, die eine interne Qualitätsbewertung des unüberwachten Clusterings ermöglicht. Da sich die interne Validierung nicht auf externe Informationen stützen kann, sind Heuristiken das Beste, worauf man hoffen kann. Unser Beitrag motiviert eine Analogie zur Entscheidungstheorie, in der eine hohe Homogenität der Meinungen unter Experten ein starker Indikator dafür ist die richtige Entscheidung getroffen zu haben. Wir betrachten die Einteilung von Einzelzelldaten, die mit verschiedenen Methoden erlangt werden können als Analogie zu sogenannten Expertenmeinungen. Für die Beurteilung assoziierter Clustering Ergebnisse entwickeln wir χ^2-basierte Assoziationsmaße. Die Evaluierung unserer Heuristik erfolgt auf denselben Benchmark-Daten, die in Beitrag 1 benutzt wurden, sowie auf Simulationsdaten mit unterschiedlichen Abhängigkeitsgraden. Dazu werden die unterliegenden Kodierungen ex-post in diese Analyse inkludiert. Dabei stellen wir fest, dass hoch assoziierte Clusterings zu einer insgesamt hohen Genauigkeit führen und daher als vertrauenswürdig gelten. Bei geringer Assoziation besteht ein höheres Risiko, ein Clustering mit schlechter Performance zu wählen.

Beitrag 3 verwendet Techniken der Theorie von verallgemeinerter Intervallwahrscheinlichkeit, um die Unsicherheit bei der Messung von Einzelzell-RNA-Sequenzierungsdaten zu quantifizieren. Wir analysieren mehrdimensionale Simulationsdaten, die auch Abhängigkeitsstrukturen unter Verwendung von Copulas einbeziehen. Dabei simulieren wir drei verschiedene Szenarien, die ein homogenes, ein heterogenes und ein intermediäres Szenario repräsentieren und alle der gleichen Abhängigkeitsstruktur folgen. Basierend auf diesen Szenarien konstruieren wir obere und untere verzerrte Messungen mithilfe von unteren und oberen Verteilungsfunktionen, die zu zwei weiteren Simulationsszenarien führen. Unter Einbeziehung eines Goldstandards ist es möglich, dass wir Hinweise für eine mögliche Kalibrierung von Messinstrumenten geben können.

Beitrag 4 kombiniert Clustering-Techniken mit regularisierter Regression. Wir schlagen das *"Discriminative Power Lasso (DP-Lasso)"* vor, das ein weiches Genfiltern ermöglicht. Für unseren Ansatz betrachten wir verschiedene Clustering-Validierungsmetriken basierend auf jeder einzelnen Kovariablen. Je besser eine Variable univariat die entsprechende kategoriale Zielvariable in verschiedene Cluster zerlegen kann, desto höher wird die Variable im Rahmen der adaptiven Lasso-Regularisierung gewichtet. Dieser Vorschlag ist durch den zugrundeliegenden genetischen Hintergrund motiviert, wobei davon ausgegangen wird, dass die entscheidenden Gene in jeder Zielgruppe unterschiedlich exprimiert sind und eine hohe Trennschärfe aufweisen. Nach der Skalierung integrieren wir die erhaltenen Gewichte in das adaptive Lasso. Da wir vor der Durchführung der regularisierten Regression keine Kovariable ausschließen, kann unser Ansatz in der Tat als eine weiche Filtermethode angesehen werden. Wenn eine Kovariable weniger wichtig erscheint, wird sie höher bestraft, während der Strafwert bei wichtigen Kovariablen niedrig ist. Die Anwendung von DP-Lasso im Kontext von Einzelzelldatensätzen und generierten Simulationsdaten zeigt, dass DP-Lasso im Vergleich zum klassischen Lasso sowohl eine variablensparsamere Lösung als auch eine genauere Vorhersage liefert.

Summary

With the sequencing of single cells it is possible to measure gene expression of each single-cell in contrast to bulk sequencing which enables only average gene expression. This procedure provides access to read counts for each single cell and allows the development of methods such that single cells are automatically allocated to cell types. The determination of cell types is decisive for the analysis of diseases and to understand human health based on the genetic profile of single cells. It is of common use that cell types are allocated using clustering procedures that have been developed explicitly for single-cell data. For that purpose the *single-cell consensus clustering (SC3)*, proposed by Kiselev et al. (Nat Methods 14(5):483-486, 2017) is part of the leading clustering methods in this context and is also of relevance for the following contributions.

This PhD thesis aims at the development of appropriate analysis techniques for the clustering of high-dimensional single-cell data and their reliable validation. It also provides a simulation framework for the investigation of the influence of distorted measurements of single cells towards clustering performance. We further incorporate cluster indices as informative weights into the regularized regression, which allows a soft filtering of variables.

Contribution 1 improves the matching of the original SC3 to the underlying biological process of transitions of single cells. We propose the unsupervised *"adapted single-cell consensus clustering (adaSC3)"*, replacing the principal component analysis of the original SC3 with diffusion maps that take the transition of single cells into account. Therefore, adaSC3 does not only respect the biological process of single cells methodologically but also improves accuracy. We evaluate the accuracy of SC3, adaSC3, and some competitive methods both on single-cell RNA-sequencing data and on simulation data, incorporating different subpopulation partitions. For all conducted studies we can state a convincing performance of *adaSC3*.

Contribution 2 proposes an *association accuracy heuristic* that allows an internal quality evaluation of unsupervised clustering. As internal validation cannot rely on external information, heuristics are the best one can hope for. Our contribution is motivated by an analogy to decision theory where high homogeneity of opinions among experts is a strong indicator to have chosen the right decision. We consider the groupings of single-cell data that are obtained by different methods as analogous to the referred expert opinions. For the assessment of associated clustering results, we adopt χ^2-based association measures and evaluate our heuristic on the same benchmark data as used in Contribution 1, as well as on simulation data with different degrees of dependence, including ground truth ex-post. We can state that highly associated clusterings result in an overall high accuracy and are therefore considered as trustworthy. In case of low association, there is a higher risk of choosing a clustering with bad performance.

Contribution 3 uses techniques from the generalized interval-probability theory to quantify the uncertainty in the measurement of single-cell RNA-sequencing data. We analyze multidimensional simulation data that also incorporate dependence structures using copulas and simulate three different scenarios, representing a homogeneous, a heterogeneous and an intermediate scenario that follow the same dependence structure. Based on these scenarios, we construct upper and lower distorted measurements using lower and upper distribution functions that lead to two further simulation settings. With the inclusion of a gold standard, we can provide instructions for possible calibration of measurement instruments in case of repeated measurements.

Contribution 4 combines clustering techniques with regularized regression. We propose the *"Discriminative Power Lasso (DP-Lasso)"*, which allows soft filtering. For our approach we consider different clustering evaluation metrics for each covariate separately. The better a variable can univariately decompose the underlying categorical target variable into distinct clusters, the higher the variable is weighted within the adaptive Lasso regularization. This proposal is motivated by the underlying genetic background, assuming that decisive genes are differently expressed in each target group and have a high discriminative power. After scaling, we incorporate the obtained weights into the adaptive Lasso. As we do not exclude any covariates before performing regularized regression, our approach can indeed be seen as a soft filtering method. In case that a covariate seems less important, the covariate has a higher penalty term, whereas the penalty is low in case of important covariates. The application of DP-Lasso to single-cell data sets and generated simulation data shows that DP-Lasso provides both a sparser solution and a more accurate prediction compared to Lasso.

Contents

Contributions of the thesis

This cumulative PhD thesis contains four publications, referred to as *Contribution 1* to *Contribution 4*:

1. *Fuetterer, C.*; Augustin, T.; Fuchs, C. (2020): Adapted Single-Cell Consensus Clustering (adaSC3). *Advances in Data Analysis and Classification*, 14:885–896. doi: https://doi.org/10.1007/s11634-020-00428-1

2. *Fuetterer, C.*, and Augustin, T. (2021). Internal Validation of Unsupervised Clustering following an Association Accuracy Heuristic. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM): Workshop on Machine Learning and Artificial Intelligence in Bioinformatics and Medical Informatics (MABM 2021)*, 2201–2210. Available under: `https://ieeexplore.ieee.org/document/9669782`

3. *Fuetterer, C.*, Schollmeyer, G., and Augustin, T. (2019). Constructing Simulation Data with Dependency Structure for Unreliable Single-Cell RNA-Sequencing Data Using Copulas. In De Bock, J., de Campos, C., de Cooman, G., Quaeghebeur, E., Gregory Wheeler, G. editors, *Proceedings of the Eleventh International Symposium on Imprecise Probabilities: Theories and Applications. PMLR*, 103:216–224. Available under: `http://proceedings.mlr.press/v103/fuetterer19a/fuetterer19a.pdf`

4. *Fuetterer, C.*, Nalenz, M., and Augustin, T. (2021). Discriminative Power Lasso – Incorporating Discriminative Power of Genes into Regularization-Based Variable Selection. Technical Report. Available under: `https://epub.ub.uni-muenchen.de/91666/1/DPL_TR_2022_03.pdf`

Declaration of the author's specific contributions

In the following listing, the own contribution of the author is specified:

- *Contribution 1:* The paper has been further developed based on the results that the first author presented at the European Conference on Data Analysis (ECDA 2019). The first author proposed replacing the principal component analysis of the original approach of Kiselev et al. (Nat Methods 14(5):483-486, 2017) by diffusion maps. Thomas Augustin had the idea to name the new method adapted single-cell consensus clustering (*adaSC3*). The author drafted and wrote the whole paper. Section 2 has been influenced by comments of Christiane Fuchs. The author included the requirements of the reviewers in each revision phase in close consultation with Christiane Fuchs and especially with Thomas Augustin. Thomas Augustin supported the first author with intensive discussions.

- *Contribution 2:* The project was brought up by the first author with the motivation of providing information of trusting unsupervised clustering results without knowing the truth. In close collaboration with Thomas Augustin, the association accuracy heuristic was formulated and sharpened. Furthermore, the proposed relation to the consensus of experts was extended by Thomas Augustin. The χ^2-based association measures needed for the investigation of associated methods were also proposed by the first author with the hint of Thomas Augustin that the Φ-coefficient is equivalent to a correlation in case of a binary coding. The simulation data as well as the real data were both initiated and analyzed by the first author. The author drafted the whole paper and included several suggestions of Thomas Augustin into the first draft as well as during revision.

- *Contribution 3:* The whole paper was drafted by the author. The author also conducted the analysis of existing approaches and constructed the simulation study, representing scenarios with different degrees of homogeneity, respecting the dependencies of genes by using copulas. The use of copulas has been proposed by Thomas Augustin. He also introduced lower and upper distribution functions, which were included by the author. The implementation of the dependence structure for the distorted simulation data, creating a fictive distribution family such that the copula function could be applied appropriately was joint work with Georg Schollmeyer. All authors contributed to proof-reading of the paper as well as to its revision.

- *Contribution 4:* The idea of the project was to bring clustering and regularization together. Including the information of clustering indices and the ANOVA as discount factors into the adaptive LASSO was proposed by the author. Malte Nalenz proposed the inclusion of the so-called discriminative power into regularized regression, which could be considered as a soft filter of genes. The technical report is based on the presentation given at the International Biometric Conference (IBC 2020) by the author, and a draft for a possible report provided by her. Malte Nalenz provided

R code for the framework of the cross validation, which was adapted by the author, including the discriminate power as penalty terms. The exact coding of the discriminative weights was finalized by both authors. The author launched and included the analysis of the single-cell RNA-sequencing data. The same R code was extended and then applied to the simulation data, which have been generated and analyzed by Malte Nalenz. The remaining parts of the technical report were rewritten by both first authors of the contribution. Thomas Augustin supported the idea from the beginning, enabled the author's contact to Gerhard Tutz in order to present the regularization approach directly to him. Furthermore, Thomas Augustin supported the way of notations throughout the technical report and provided advice for writing and structuring the paper appropriately.

1 Motivation: Challenges and opportunities of single-cell RNA-sequencing data

1.1 Biological background and workflow of single-cell analysis

With the invention of the microscope in the 17^{th} century, access to the shape and functions of cells has been enabled, which led to a categorization of groups of cells (Trapnell, 2015). In 1977, the technique of Sanger sequencing allowed the determination of a nucleotide sequence of the deoxyribonucleic acid (DNA), which had been discovered in 1953. Both have been rewarded with a Nobel prize. Since the advent of next generation sequencing in the year 2005, an intensive study of the genome can be realized based on huge amounts of data (Shendure et al., 2017). The most recent techniques of single-cell RNA-sequencing, such as Illumina sequencing, allow to study the diversity of cell populations. This includes the detection of rare or new cell types (Briggs et al., 2021; Duò et al., 2020), as the data provides insights into biological functions and complex biological systems on the level of a single cell (Pouyan et al., 2016). Therefore, single-cell data bring along a tremendous advantage, compared to the traditional bulk RNA-sequencing where only the average of gene expression is measured among all cells. The drawback of bulk data is that depending on the cellular composition of the tissue, the measurements are confounded because the genes are regulated in dependence of their cell state which gets lost by averaging the gene expression over all cells (Trapnell, 2015). As single cells do not develop synchronously, the average measurements of different sampling time points do not represent the underlying transition process. These transitions might be due to the development of certain gene functions or cells passing through several states from a zygote to an adult species. The development of an embryo is a continuous process, stimulated by certain genes (Trapnell, 2015). Both the decisive genes and the passed states are not known a priori because this is a very complex and dynamic biological process. The analysis of cells passing from one state to another is a very specific challenge related to single-cell RNA-sequencing data. However, the improved resolution of measuring the gene expression profile (GEP) of single cells (Zappia et al., 2017) also contributed to the mentioned opportunities of single cells, which enable valuable insights into the heterogeneity of cells. This new technique allows a better understanding of the underlying biology and human diseases (Angerer et al., 2017).

The Human Cell Atlas project aims at creating a data-driven reference for cell types that

are involved in the biological functioning or development of a specific organism (Angerer et al., 2017). With the understanding of a healthy organism, it is easier to explain the occurrence of diseases by mapping single-cell RNA-seq samples to reference atlases (Kiselev et al., 2019). As stated by Trapnell (2015) there is however no clear definition of cell types and therefore it is of highest priority to develop accurate unsupervised clustering methods for an appropriate creation of the Human Cell Atlas. Consequently, clustering methods are considered the most powerful tools for that purpose. This might explain the high interest into the research field of single-cell data, which is indicated by more than 120 published software packages in peer-reviewed journals or preprints in 2017 (Zappia et al., 2017) and more than 1000 tools counted in September 2021 (Zappia and Theis, 2021). The corresponding software packages target different steps of the single-cell RNA-sequencing analysis workflow, schematically presented in Figure 1.1. The workflow, described in the following includes the sequencing, pre-processing, dimension reduction and clustering of single cells, followed by its biological interpretation.

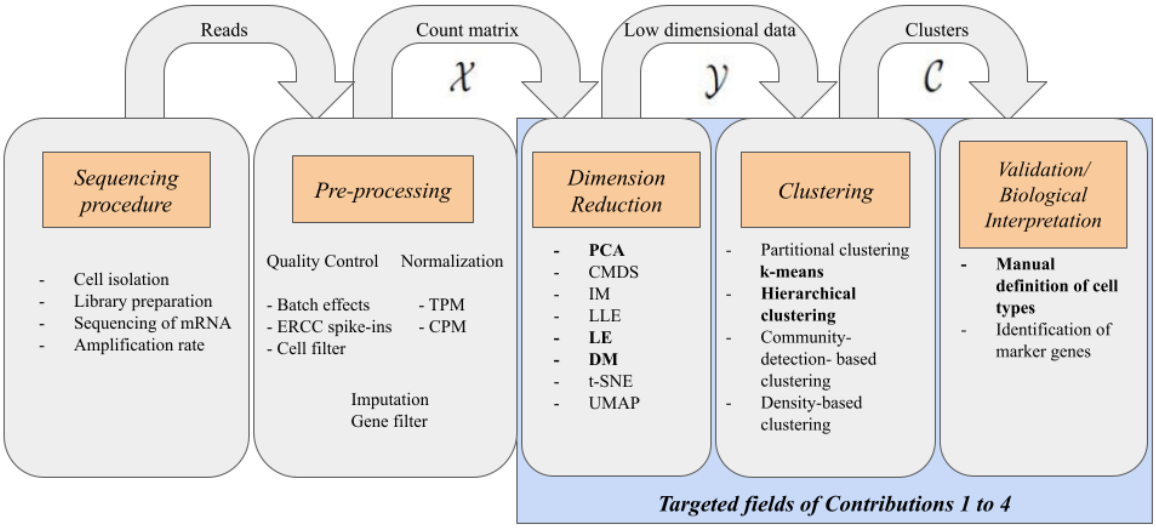

Figure 1.1: Workflow of scRNA-sequencing data analysis and targeted fields for contributed work with particular importance for this thesis, displayed in bold.

Single-cell RNA-sequencing procedure

The single-cell RNA-sequencing (scRNA-seq) procedure first isolates each single cell and measures its abundance of mRNA using a cell-specific sequencing library, which contains a collection of the mRNA fragments (Angerer et al., 2017). After library preparation, RNA fragments are sequenced. Each of these subsequences is called a *read*, and the number of nucleotides per fragment is defined as *read length*. The *sequencing depth* contains the

mean of how often each nucleotide was measured in the genome (Sims et al., 2014). In the step of alignment, each of these reads has to be aligned to a subsequence of the reference genome. The alignment to the gene depends on the sample location along the transcript of the sample (Svensson et al., 2018; Angerer et al., 2017). The output of the sequencing procedure are then *counts*, which represent the frequency how often one read can be assigned to the same gene for each single cell, resulting in a *count matrix* (Luecken and Theis, 2019). The data sets treated below contain annotated genes, which is the reason why we will refer to the term count matrix only.

Data preparation and data structure

Providing some background information, the quality control includes outlier detection, which can be an indicator of dying cells, broken membranes or doublets (Luecken and Theis, 2019) leading to the exclusion of cells (cell filter). Quality control also aims to circumvent confounding effects such as batch effects and library effects (Zappia et al., 2017), which are not considered in the following. We are also not going to elaborate on spike-in genes, which can serve as control genes, allowing a distinction between biological and technical dropouts because we assume that the scRNA-sequencing data used below are well prepared according to Kiselev et al. (2017).

The second part of the pre-processing procedure is the normalization, as illustrated in Figure 1.1. The normalization aims to obtain relatively comparable gene expression values between the different cells (Vasighizaker et al., 2022), preventing differing gene expression for identical cells due to e.g. sequencing uncertainty, which can be lead back to sampling effects. One of the most often used normalization techniques is the *counts per million (CPM)* normalization or in case of full-length scRNA-sequencing protocols the *transcripts per million (TPM)* normalization. Full length protocols allow that cells are comparable to each other, incorporating the gene length (see e.g. Patel et al. (2014), Kowalczyk et al. (2015), Soneson and Robinson (2018)). Especially for the pre-processing procedure there exists no golden standard and many tools are available.

Concerning the measurement of gene expression it is possible that genes have 0 counts. Here, one has to distinguish between biological reasons (e.g. species, tissue type, cell type, treatment, and cell cycle) or technical reasons (e.g. platform, protocol, or processing). The dropout of the latter can be explained by the amplification rate, where in case of choosing a too low sequencing depth, some gene expression is not measured (Zappia et al., 2017). Wagner et al. (2013) claim that the *zero-inflated negative binomial (ZINB)* distribution is the most appropriate distribution for the count data of single-cell RNA-sequencing. With this distribution the high frequency of zero counts as well as the gene-wise over-dispersion due to high noise can be respected (Svensson et al., 2018; Angerer et al., 2017). Kleiber and Zeileis (2016) describe the ZINB distribution as a mixture of a dropout rate and a negative binomial distribution, which will especially serve for the construction of simulation settings of *Contribution 1* and *Contribution 3*. According to literature, such as (see Brennecke et al. (2013); Grün et al. (2014)), the biological and technical variation could also be described by a Poisson distribution. In this case, only read counts strictly greater than 0 are taken

into account neglecting the zero-inflation rate.

After quality control and normalization one attemps to compensate missing values by imputation techniques (Svensson et al., 2018). Then, it is common standard to reduce the number of genes by a gene filter to exclude non-informative genes, as there are high dropout rates in the very high dimensional single-cell data. The challenge of *gene filtering* is the determination of thresholds (Zhang et al., 2021), which should guarantee sufficient data quality without the danger of data peeking. The problem hereby is that the quality control could be adapted after visualization or clustering, which would have an impact on the results (Luecken and Theis, 2019).

Dimension reduction and clustering

As a next step following the schematic overview of Figure 1.1, a combination of dimension reduction and clustering is applied, which will be the main focus within this dissertation. These approaches target at an insight into the data through the cell clusters, while maintaining the underlying data structure in a low dimensional space. Clustering single cells brings along the great opportunity to allow access to different cell states by the analysis of differing gene expression. The *visual inspection* of defining cell types is obtained by considering two dimensional plots obtained after dimension reduction. The *manual definition* of cell types requires a clustering, which leads to a data-driven categorization, relying on the abundance pattern of the obtained clusters (Duò et al., 2020). Within this dissertation we consider the cell types of the analyzed single-cell RNA-sequencing data as ground truth. Referring to the blue highlighted field of Figure 1.1 this dissertation targets dimension reduction techniques combined with clustering approaches and their validation, provided in *Contribution 1*, *Contribution 2*, and *Contribution 3*. We therefore give some theoretical background about dimension reduction and cluster analysis in Chapter 2. In addition, regularization techniques are combined with cluster theory in *Contribution 4*.

Validation / Biological interpretation

As the term cell type is not well defined one has to be careful at which level one searches for cell types. While for some researchers the level of "T cells" is sufficient, others might look for so called $CD4^+$ or $CD8^+$ T cells, which are subtypes of T cells (Luecken and Theis, 2019). Thanks to reference data bases such as the Mouse Cell Atlas (Han et al., 2018) and the Human Cell Atlas (Rozenblatt-Rosen et al., 2017), the annotation of clusters to cell types is supported. For the according procedure, a differential expression (DE) testing (e.g. Wilcoxon rank-sum test or the t-test and its adjustments allow multiple testing (Vasighizaker et al., 2022)) between each cluster and the remaining clusters is conducted. Significant genes are then ranked according to their adjusted p-values and further criteria with the assumption that especially up-regulated genes are of interest (Luecken and Theis, 2019). The corresponding marker genes serve for the annotation of the respective cluster to the closest cell type population, including an appropriate reference, such as the Gene Set Enrichment Analysis (GSEA) (Vasighizaker et al., 2022; Subramanian et al., 2005). Thus an analysis

on the single cell level also allows a deeper understanding of the behavior and development of diseases, providing the pathway to personalized medicine. However, this investigation cannot be done for bulk data and its medical applications as they allow no insight into the GEP at a cellular level. Compared to single-cell data, it is therefore not possible to explain the reason why gene expression might have an impact after drug prescription. For the subsequent investigation, we mainly extract the underlying single-cell states through cluster analysis, respecting different uncertainty scenarios. The identification of new cell types or the confirmation of predefined cell types aims at a better understanding of a cell's development and the emergence of diseases (Zappia et al., 2017). Especially the research field of immunology and health care improved substantially, as it was possible to measure cellular markers, such as proteins on the level of single cells (Pouyan et al., 2016). Also the opportunity of having access to intra-tumor heterogeneity (Yu and Du, 2022) contributed to the development of personalized medicine.

1.2 Current dimension reduction and clustering techniques used in the context of single-cell data

As described above, a manual determination of cell types is done in a data-driven way by cluster analysis, which is therefore the most relevant and decisive part in the scRNA-seq workflow (see Tian et al. (2021); Shapiro et al. (2013); Kolodziejczyk et al. (2015a); Kiselev et al. (2019)). Due to the high dimensional single-cell data, clustering methods might not lead to well-interpretable results in case of a direct application to the count data. It is therefore recommended to first project the high dimensional data into a lower dimension (Gan et al., 2022). The most commonly used dimension reduction techniques in this context are mainly based on the principal component analysis, according to the review papers of Kiselev et al. (2019) and Zhang et al. (2020). For instance, *SC3* (Kiselev et al., 2017), pcaReduce (Zurauskiene et al., 2016) and TSCAN (Ji and Ji, 2016) are all based on the principal component analysis (PCA). The Clustering through Imputation and Dimensionality Reduction (CIDR) (Lin et al., 2017) is an algorithm which includes principal coordinates for dimension reduction. For respecting the cells' differentiation trajectories, Haghverdi et al. (2015) propose the use of *diffusion maps* by Coifman and Lafon (2006), which include some stochasticity in the dimension reduction process. The general toolboxes Scanpy (Wolf et al., 2018) and Seurat (Butler et al., 2018) are specifically adapted to the analysis of single-cell gene expression data. In case of Seurat a pre-selection of highly variable genes (HVG) is done before dimension reduction and clustering. According to the review paper of Duò et al. (2020), especially Seurat and SC3 are dominating methods.

In the following an overview of clustering methods often used for benchmarking single-cell data is provided. RaceID-based algorithms (Grün et al., 2015, 2016; Herman et al., 2018) are of special interest for detecting rare cell populations. Concerning the clustering of pcaReduce, first several k-means clusterings (MacQueen et al., 1967) are applied, which are then combined with a subsequent *hierarchical clustering*. The same procedure holds

for the consensus clustering of *SC3*. Concerning the clustering part of CIDR, hierarchical clustering is performed. As an outlook, the category of community-detection-based clustering includes spectral clustering and the often used Louvain algorithm (Blondel et al., 2008), where the latter is part of Seurat and scanpy. Spectral clustering algorithms such as SIMLR (Wang et al., 2018) use several kernels to learn the similarity of single cells for the subsequent spectral clustering (Gan et al., 2022). Concerning the density-based clustering approaches, especially the algorithm DBSCAN (Ester et al., 1996) has to be mentioned, as well as its extensions GiniClust (Jiang et al., 2016) and Monocle2 (Qiu et al., 2017). These are particularly useful for detecting rare cell types.

For being able to sequence a large number of single-cells, some deep learning approaches can be applied for dimension reduction (see Gan et al. (2022); Tian et al. (2021) or the review of Zhang et al. (2020)). Especially neural networks or (variational) auto-encoders perform well (Tian et al., 2021), such as Deep Count Autoencoder (DCA) (Eraslan et al., 2019) or the variational autoencoder for scRNA-seq data (VASC) Wang and Gu (2018). However, these deep learning approaches will not be studied further within this dissertation.

1.3 Outline of this thesis

In this dissertation, we consider uncertainties that might occur during the single-cell workflow analysis described in Section 1.1. In Chapter 2 we first introduce the term high dimensional clustering, providing the concepts of dimension reduction techniques in Section 2.1 and clustering in Section 2.2. In Section 2.3 we give a short overview of the often cited single-cell consensus clustering of Kiselev et al. (2017). In Chapter 3 we present a study of each contribution on its own, including both a summary and possible perspectives. The contributions target the uncertainty of the biological background and the uncertainty of different clustering results. Furthermore, we included measurement uncertainty and clustering information as input for regularization approaches.

We explicitly propose

- ... better adapting the dimension reduction of the original SC3 to the biological background of single-cell data (Section 3.1).
- ... an association accuracy heuristic which aims to support the data-driven process of defining cell types, based on different clustering methods (Section 3.2).
- ... a simulation framework, considering dependence structure of genes, and a possibility to analyze the consequences of distorted measurements on the clustering performance (Section 3.3).
- ... an inclusion of univariate grouping information obtained by cluster indices into regularization approaches (Section 3.4).

Chapter 4 provides some general concluding remarks.

2 High dimensional clustering

For an appropriate analysis of high dimensional single-cell data, the first challenge is to extract valuable information in a compact way. In the unsupervised case, a dimension reduction technique can be performed, which is also the motivation for *Contribution 1* to *Contribution 3*. In supervised cases, methods such as regularization techniques are often applied (*Contribution 4*). In this dissertation, we use the term *high dimensional clustering*, referring to the use of clustering information of high dimensional data. In detail this means that generally a dimension reduction technique is applied prior to the subsequent clustering. In contrast to that, *Contribution 4* uses the univariate clustering information for the regularization of the high-dimensional single-cell data.

2.1 Dimension Reduction

Dimension reduction (DR) aims to embed the high dimensional data into a lower dimension. For observation $i \in \{1, \cdots, N\}$ we therefore denote the covariate vector of length p with $x_{\cdot i}$, and the lower dimensional vector with $y_{\cdot i}$, which contains $p' \leq p$ transformed covariates. The according embedding can be achieved directly by mapping $(x_{\cdot 1}, \cdots, x_{\cdot N}) \xrightarrow{DR} (y_{\cdot 1}, \cdots, y_{\cdot N})$, preserving the structure of the original data $\mathcal{X}_{p\times N}$, consisting of p covariates and N observations. The resulting data matrix of each DR $\mathcal{Y}^{(DR)}_{p'\times N}$ is thus obtained by:

$$\mathcal{X}_{p\times N} \xrightarrow{DR} \mathcal{Y}^{(DR)}_{p'\times N} \qquad \text{with } p' \leq p. \tag{2.1}$$

For dimension reduction techniques, it is in general of relevance whether the corresponding methods use a *local* or a *global* embedding. While local embeddings focus on "nearby" observations, global embeddings focus on preserving "faraway" observations in the low-dimensional embedding (Silva and Tenenbaum, 2002). Furthermore, dimension reduction techniques can be distinguished by linear and non-linear embeddings. As scRNA-seq data follow a non-linear pattern, the linear embedding of the principal component analysis (PCA) or the classical multidimensional scaling (CMDS) have the tendency to miss important information (see Tenenbaum (1997)). Therefore, especially for the high-dimensional gene expression data, many non-linear dimension reduction techniques have been proposed. In the subsequent part of this chapter we give an overview of DR techniques.

Principal Component Analysis (PCA)

The principal component analysis (PCA) is a linear dimension reduction technique, as higher dimensions are embedded into lower dimensions by linear combinations. PCA is a very popular tool for dimension reduction because it is well interpretable and can identify decisive factors, which are important genes in our case.

The PCA can be led back to Pearson (1901) and Hotelling (1933) and has the aim to reduce p correlated covariates to a considerably lower number of so called *principal components*, explaining as much of the variance as possible. This goal is achieved by transforming the observed covariates into linear combinations, which are the principal components. These principal components are uncorrelated and orthogonal to each other, ordered by decreasing variance. Especially in genetics, where genes are highly correlated, the PCA allows to reduce the number of dimensions substantially. While the first principal component still explains most of the variance, the proportion of explained variance to the total variance decreases more and more with further components, with the consequence that the last principal components explain only a negligible part of the total variance. Accordingly, e.g. thresholds or the scree test can serve for the determination of the number of selected principal components.

We consider the positive semidefinite covariance matrix Σ of the data matrix $\mathcal{X}$, but we can also think of Σ as correlation matrix, with $rank(\Sigma) = r \leq p$. As Σ is symmetric, the spectral decomposition allows to solve the following equation:

$$\Sigma = \Psi_{PCA} \Lambda_{PCA} \Psi_{PCA}^T. \tag{2.2}$$

The solution of Equation (2.2) are the eigenvectors $\psi_1^{(PCA)}, \cdots, \psi_r^{(PCA)}$, which form the orthogonal matrix Ψ_{PCA} with the corresponding eigenvalues $\lambda_1^{(PCA)}, \cdots, \lambda_r^{(PCA)}$. These are sorted in descending order and are part of the corresponding diagonal matrix Λ_{PCA}. The according lower dimensional embedding can be achieved by selecting the first $p' \leq r \leq p$ eigenvectors of Ψ_{PCA} leading to $\mathcal{Y}^{(PCA)} := (\psi_1^{(PCA)}, \cdots, \psi_{p'}^{(PCA)})^T \mathcal{X}$.

Classical Multi-Dimensional Scaling (CMDS)

The classical multi-dimensional scaling (CMDS) is a global dimension reduction technique, which was proposed by Torgerson (1952). The aim of the CMDS is to find a low dimensional embedding $\mathcal{Y}^{(CMDS)}$, which is described as configuration, such that Euclidean distances $d_2(x_{\cdot i}, x_{\cdot j})$ of observations i and j are maintained the best possible. It is hereby assumed that no pairwise distance is missing (Wang, 2012):

$$d_{\mathcal{Y}}(y_{\cdot i}, y_{\cdot j}) \approx d_2(x_{\cdot i}, x_{\cdot j}) = ||x_{\cdot i} - x_{\cdot j}||_2, \ \forall i, j \in \{1, \cdots, N\}. \tag{2.3}$$

The embedding of the CMDS is optimal when the distances of $\mathcal{Y}$ are as close as possible to the distances of $\mathcal{X}$ over all pair-wise observations. The according criterion is achieved by minimizing the so called stress function (see Hastie et al. (2009)). In general, any Euclidean

metric is accepted for the embedding of CMDS[1]. For each pair of observations i and j the squared distance is obtained with: $S_{ij} = D_{ij}^2$, which builds the matrix S, and allows the construction of the centered Euclidean square-distance matrix S^c.

The main advantage of multidimensional scaling is that only a dissimilarity matrix is needed instead of the original data, which is especially interesting for gene data, as $p >> N$. A generally obtained result of Wang (2012) shows that the centered Gram matrix G^c is equal to $G^c = -\frac{1}{2}S^c$. In general, this result allows to construct a squared dissimilarity matrix S instead of the centered original data for the inner product, leading to the gram matrix. We further denote the eigenvector matrix of CMDS with Ψ_{CMDS} and the diagonal matrix of eigenvalues with Λ_{CMDS}, which are obtained by the spectral decomposition of the centered Gram matrix G^c:

$$G^c = \Psi_{CMDS}\Lambda_{CMDS}\Psi_{CMDS}^T. \tag{2.4}$$

The low dimensional embedding can then be realized with the selection of p' eigenvectors for Ψ_{CMDS}, which leads to the reduced data set: $\mathcal{Y}^{(CMDS)} = (\psi_1^{(CMDS)}, \cdots, \psi_{p'}^{(CMDS)})^T \mathcal{X}$.

Isometric Feature Mapping (IM)

The isometric feature mapping (IM) by Tenenbaum et al. (2000) is a global dimension reduction technique. In contrast to the linear dimension reduction techniques, the underlying geometry of this nonlinear DR is a manifold, which is a topological space that is locally Euclidean (Lee, 2010). It is assumed that the data $\mathcal{X} \subset \mathbb{R}^p$ lie on a Riemann manifold $\mathcal{M} \subset \mathbb{R}^p$. The according isometric mapping $f^{(IM)} : \mathcal{M} \to \mathbb{R}^{p'}, f(x_{.i}) = y_{.i}$ aims to arrange the low dimensional data $\mathcal{Y} \subset \mathbb{R}^{p'}$, with $p' \leq p$ (Wang, 2012), such that the Euclidean distance d_2 matches the *geodesic distance* $d^{\mathcal{M}}$ as much as possible for all pairs of observations:

$$d_2(y_{.i}, y_{.j}) \approx d^{\mathcal{M}}(x_{.i}, x_{.j}), \forall i, j \in \{1, \cdots, N\}. \tag{2.5}$$

As approximation of the geodesic distance $d^{\mathcal{M}}$, the graph distance $d^{\mathbb{G}}$ of an undirected graph $\mathbb{G}$ is considered. With the constructed graph, we can then calculate the distance between each pair of data points that are connected, considering $k \in \{1, \cdots, K\}$ neighbors[2]. The set of all possible paths between $x_{.i}$ and $x_{.j}$ will be denoted with Γ. Thus, the corresponding vector for path $\gamma = (x_{.0}, \cdots, x_{.s+1})$ contains all connection points with $x_{.i} = x_{.0}$ as startpoint and $x_{.j} = x_{.s+1}$ as endpoint, which allows the calculation of the *path distance* d^{γ} using the Euclidean distance d_2:

$$d^{\gamma}(x_{.i}, x_{.j}) = d_2(x_{.0}, x_{.1}) + \cdots + d_2(x_{.s}, x_{.s+1}). \tag{2.6}$$

[1] We base our description on the pair-wise Euclidean distance $d_2(x_{.i}, x_{.j})$ of observation i and j, which was also used by Torgerson (1952). In case that CMDS is applied to Euclidean distances, the embedding is the same as if PCA was applied to the centered original data (see Hastie et al. (2009), and proof in Wang (2012)).

[2] see footnote 5 of Contribution 1 for the chosen algorithm, determining K.

Based on the set of path distances $\gamma \in \Gamma$, the *graph distance* $d^{\mathbb{G}}$ chooses the closest distance between all paths of $x_{\cdot i}$ and $x_{\cdot j}$ (Tenenbaum et al., 2000):

$$d^{\mathbb{G}}(x_{\cdot i}, x_{\cdot j}) = \min_{\gamma \in \Gamma} d^{\gamma}(x_{\cdot i}, x_{\cdot j}). \tag{2.7}$$

The advantage of the graph distance is that the detour of passing by s connection points might deliver a shorter distance compared to the direct way, underlying the Riemann manifold $\mathcal{M}$ for $\mathcal{X}$. As a next step, a CMDS is applied to the graph distance matrix $D_{\mathbb{G}}$, such that the intrinsic geometry is maintained in the low dimensional data set $\mathcal{Y}^{(IM)}$. The corresponding squared distance of isomap $S_{\mathbb{G}}^{(IM)}$ includes the square of the introduced graph distance matrix $D_{\mathbb{G}}$ for each pair of observations. Centering the obtained $S_{\mathbb{G}}^{(IM)}$ and following the same derivation as CMDS, the according centered Gram matrix G^c of the isometric mapping is obtained. In accordance to the above described methods, a spectral decomposition of G^c can be applied (Wang, 2012), providing the eigenvector matrix Ψ_{IM} and eigenvalue matrix Λ_{IM}:

$$G^c = \Psi_{IM} \Lambda_{IM} \Psi_{IM}^T. \tag{2.8}$$

Selecting the p' first eigenvectors $(\psi_1^{(IM)}, \cdots, \psi_{p'}^{(IM)})$ allows the embedding into the low dimensional data set $\mathcal{Y}^{(IM)}$, following the same DR procedures as shown above.

Locally Linear Embedding (LLE)

The locally linear embedding (LLE) introduced by Roweis and Saul (2000) can be seen as a local non-linear transformation technique. For the global embedding of the upper described isometric feature mapping (IM), the whole data structure is taken into account for the calculation of the geodesic distance. In contrast to IM, LLE focuses on a local region of the nearest neighbors. For the locally linear embedding, we assume a p-dimensional manifold $\mathcal{M} \subset \mathbb{R}^p$. In accordance to IM, we consider the same undirected graph $\mathbb{G} = [\mathcal{X}, A]$, with adjacency matrix A. If the adjacency matrix value A_{ij} of observation i and j is not equal 0, observation j is defined as a neighbor of observation i. All observations j, fulfilling the described requirement are part of the observation set $H(i)$, and form the according neighbors of observation i (Wang, 2012):

$$H(i) := \{j, A_{i,j} \neq 0\}, \tag{2.9}$$

The covariate values of the described observation set $H(i)$ build the according neighborhood set $O(i)$ of observation i, with:

$$O(i) := \{x_{\cdot j} \in \mathcal{X}, j \in H(i)\}. \tag{2.10}$$

The underlying geometric assumption is that each $x_{\cdot i}$ can be approximated by an orthogonal projection f from the manifold $\mathcal{M}$ to the tangent space $T_{x_{\cdot i}}$, leading to $\tilde{x}_{\cdot i} = f(x_{\cdot i})$. To avoid the computational effort of the tanget projection $\mathcal{M} \xrightarrow{f} T_{x_{\cdot i}}$ for the according

mapping, we try to minimize the distance of $x_{\cdot i}$ to its neighborhood set $O(i)$, underlying barycentric coordinates[3], as described in Wang (2012), including its observation set $H(i)$ only. As we assume a connected graph, at least one j is connected with i ($j \in H(i)$), the according weights $w_{i,j}$ are constructed to weight the impact of each $x_{\cdot j} \in O(i)$ on $x_{\cdot i}$. This leads to solving the optimization problem, described by Hastie et al. (2009):

$$\min_{W_{i,j}} ||x_{\cdot i} - \sum_{j \in H(i)} w_{i,j} x_{\cdot j}||^2, \text{ s.t. } \sum_{j \in H(i)} w_{i,j} = 1, \tag{2.11}$$

such that the L2 norm of $x_{\cdot i}$ is minimized with respect to the reconstructed $\tilde{x}_{\cdot i} = \sum_{j \in H(i)} w_{i,j} x_{\cdot j}$, including only $j \in H(i)$, with $H(i) \neq \emptyset$, which builds on the idea of barycenters. Thus, close neighbors should be weighted highly, whereas faraway neighbors are weighted less. To make Equation (2.11) identifiable, the number of considered neighbors K has to be less than p. Furthermore, the minimization problem has to fulfill the sparseness criterion setting $w_{i,j} = 0$ in case of $j \notin H(i)$, allowing only the reconstruction of $x_{\cdot i}$ by its neighbors $x_{\cdot j} \in O(i)$ (Wang, 2012). Equation (2.11) can be solved by constructing an inner product over a reference point $x_{\cdot k}$, with $k \in H(i)$. The according Gram matrix allows then access to an appropriate weighting for each pair of observations, leading to weight matrix W, respecting the above mentioned requirements. The weight matrix is then adapted to the LLE kernel $E = (I - W)^T (I - W)$, which is a positive semidefinite matrix, and thus can be used for the following spectral decomposition, leading to eigenvector matrix Ψ_{LLE} and diagonal eigenvalue matrix Λ_{LLE}:

$$E = \Psi_{LLE} \Lambda_{LLE} \Psi_{LLE}^T. \tag{2.12}$$

The low dimensional embedding can then be realized by using the p' eigenvectors, leading to the reduced data set: $\mathcal{Y}^{(LLE)} = (\psi_1^{(LLE)}, \cdots, \psi_{p'}^{(LLE)})^T \mathcal{X}$.

Laplacian Eigenmaps (LE)

The local embedding of Laplacian eigenmaps (LE) (Belkin and Niyogi, 2003) also aims to solve the sparse eigenvalue problem in a nonlinear manner. In accordance with the local dimension reduction of LLE, LE also consider an undirected graph $\mathbb{G} = [\mathcal{X}, A]$. The same neighborhood set $O(i)$ of observation i will be used for the calculation of the weights of the

[3] Möbius (1827) introduced barycentric coordinates, placing masses on a triangle, such that a point of interest is the gravity center. Barycentric coordinates allow to describe the position of each observation, using a set of vertices $\mathcal{V} = \{x_1, \cdots, x_N\}$ that form a simplex in $\mathbb{R}^{N-1}$. Point $\tilde{x}$ is by definition the gravity center of the weights $(w_1, \cdots, w_N)$ of vertices $(x_1, \cdots, x_N)$:

$$\tilde{x} = w_1 x_1 + \cdots + w_N x_N, \text{ such that } \sum_{i=1}^{N} w_i = 1.$$

In case of a simplex in $\mathbb{R}^{N-1}$, barycentric coordinates w are non-negative, given the point $\tilde{x}$ is part of the simplex (Wang, 2012).

graph Laplacian L. For LE a p-dimensional Riemann manifold $\mathcal{M}$ is assumed ($\mathcal{M} \subset \mathbb{R}^p$), such that $\mathcal{X} \subset \mathbb{R}^p$. With the function f the mapping $f : \mathcal{M} \to \mathbb{R}$ should be described. As stated by Wang (2012) the graph Laplacian L is approximated and discretized. These transformations result in the self-adjoint graph Laplacian L_{LE}. The construction of the according weight matrix W_{LE} can be based on a heat kernel, which contains the distance of $x_{\cdot i}$ to $x_{\cdot j}$ (see Belkin and Niyogi (2003)), delivering a weighting for each pair of observations of the underlying adjacency matrix. Alternatively, in case that $x_{\cdot j}$ is (not) part of $O(i)$, as the pair is (not) connected the corresponding weight is one (zero). With D_{LE} we denote the diagonal matrix of the weight matrix W_{LE}. After having obtained access to $L_{LE} = D_{LE} - W_{LE}$, and D_{LE} (Wang, 2012), the matrices L_{LE} and D_{LE} are used to solve the following eigen problem of Laplacian eigenmaps:

$$L_{LE} f = \lambda^{LE} D_{LE} f, \tag{2.13}$$

with eigenvalues: $0 = \lambda_0^{(LE)} \leq \lambda_1^{(LE)} \leq \cdots \leq \lambda_p^{(LE)}$, and eigenfunctions $f = (f^1, \cdots, f^p)$, which correspond to the eigenvectors $(\psi_1^{(LE)}, \cdots, \psi_p^{(LE)})$. The selection of p' eigenvectors or eigenfunctions then allow the low dimensional embedding $\mathcal{Y}^{(LE)}$ with $p' \leq N - 1$.

Diffusion maps (DM)

Diffusion maps (DM) (Coifman and Lafon, 2006) consider the connectivity for each pair of observations, which are defined as a probability, reaching each other by a random walk. Unlike IM, diffusion maps are considered as robust, as they look at all pair-wise paths, including t steps, in contrast to the shortest geodesic distance of IM, no matter how many steps have to be taken (De la Porte et al., 2008). Coifman and Lafon (2006) propose approximating the connectivity of $x_{\cdot i}$ and $x_{\cdot j}$ by a diffusion kernel. In accordance to the notation above, we denote the kernel of $x_{\cdot i}$ and $x_{\cdot j}$ with $w_{i,j}$, approximating its connectivity$_{i,j}$:

$$\text{connectivity}_{i,j} \approx w_{i,j}. \tag{2.14}$$

As a choice for $w_{i,j}$, one could think of the Gaussian kernel, which includes a parameter for the determination of the neighborhood size. For the exact relation of the pair-wise connectivity, a normalization constant for the pair-wise weighting Z_i is needed, leading to the discrete construction of connectivity:

$$\text{connectivity}_{i,j} = \frac{1}{Z_i} w_{i,j}, \text{ with } Z_i = \sum_{j=1}^{N} w_{i,j}. \tag{2.15}$$

The pair-wise probability interpretation $p_{i,j}$, also used for the application to single-cell data following Haghverdi et al. (2015), can be realized by respecting the whole neighborhood with:

$$p_{i,j} = \frac{1}{\tilde{Z}_i} \frac{w_{i,j}}{Z_i Z_j}, \text{ with } Z_j = \sum_{i=1}^{N} w_{i,j}, \text{ and } \tilde{Z}_i = \sum_{j \neq i} \frac{w_{i,j}}{Z_i Z_j}, . \tag{2.16}$$

Each row of the obtained transition probability matrix P sums up to 1 (Coifman and Lafon, 2006). Multiplying the corresponding matrix P with t repetitions allows to take into account in how many steps one observation to another should be reached for each pair of observations. The higher the parameter t, the more the path respects the underlying geometric structure, resulting in a path's overall high probability (De la Porte et al., 2008). Thus, P can be considered as a transition kernel of a Markov chain. This brings along the property that for a connected graph a unique stationary distribution can be expected and the chain is reversible. For finite $\mathcal{X}$ the chain is even ergodic. These attributes, in combination with some assumptions on the kernel allow the spectral decomposition of the Markov chain, leading to (Coifman and Lafon, 2006):

$$P = \Psi_{DM} \Lambda_{DM} \Psi_{DM}^T. \tag{2.17}$$

For the low dimensional embedding p' eigenvectors are taken leading to: $\mathcal{Y}^{(DM)} = (\psi_1^{(DM)}, \cdots, \psi_{p'}^{(DM)})^T \mathcal{X}$. As stated in Haghverdi et al. (2016), the eigenvectors of P and P^t are the same, however P^t includes the eigenvalues to the power of t.

Comparison in the context of single-cell data

The presented dimension reduction techniques are needed to embed the high dimensional single-cell RNA-sequencing data into lower dimensions. These techniques use either matrix factorization or neighbor graphs, and are part of the contributions, presented in the next chapter. The linear techniques PCA and CMDS use matrix factorization; IM, LLE, LE, and DM are built on neighbor graphs. The linear dimension reduction techniques PCA and CMDS might not be an accurate method for single-cell data, which have a high zero-inflation rate and are over-dispersed. Thus, the linearity assumption could be violated. PCA and CMDS are also both global embeddings and therefore cannot take into account the intrinsic data structure (Buettner and Theis, 2012; Sumithra and Surendran, 2015).

IM, LLE, LE and DM provide non-linear dimension reduction techniques in order to improve the limitations of PCA and CMDS as described by Wang (2012). Even though IM is a global embedding like PCA and CMDS it takes the nearest neighbors into account by using graph distances. IM perform well in case of dense enough data structures. LLE and LE embed the high dimensional data preserving the local neighborhood structure. LLE is based on a neighborhood set highly depending on the number K of included neighbors. If K is chosen too high the locally embedding cannot be guaranteed, whereas a too low number of K leads to no stable results. The drawback of LE is its sensitivity to outliers, resulting in an unstable embedding. The global DM rely on a Markov transition matrix and have the advantage that they are very effective in detecting the different states of single cells. Of course apart from the described techniques, there exist many more dimension reduction methods that are appropriate for single-cell RNA-sequencing data.

2.2 Cluster analysis

For high dimensional data, clustering is the subsequent step of dimension reduction, as indicated in the overview of Figure 1.1. The purpose of *clustering* is to detect the underlying subgroups of the compact data structure, obtained by the low dimensional data set $\mathcal{Y} = \{y_{\cdot 1}, \cdots, y_{\cdot N}\}$. The aim of clustering in general is to combine observations that are similar to each other into groups, called *clusters*, in which the distances to dissimilar observations in other clusters are higher compared to the distances of observations within a cluster as described by Hastie et al. (2009). Clustering is also called *data segmentation* as the underlying process aims to find the natural group structures in an unsupervised way (Kiselev et al., 2019). Hereby, the challenge is to cluster the set of N observations $\mathcal{I} = \{1, \ldots, N\}$ into a set of clusters $\mathcal{C} = \{1, \ldots, K\}$ based on a clustering algorithm M, with $K << N$. The clustering process thus corresponds to the mapping of:

$$\mathcal{I} \overset{M}{\rightarrow} \mathcal{C}. \tag{2.18}$$

In the following, first different clustering approaches are described which are needed in *Contribution 1* to *Contribution 3*, followed by validation measures assessing the cluster quality, which serve as preparation for *Contribution 2* and *Contribution 4*.

Clustering algorithms

Clustering algorithms can be divided into four categories: the general *partitional clustering*, where *k-means* is the most often used algorithm, *hierarchical clustering*, *community-detection-based clustering*, and *density-based clustering*.

K-means:
As a representation of the *partitional clustering* we present the *k-means* algorithm (MacQueen et al., 1967). In the first step a specific number of clusters, which is denoted with K has to be defined. Then, the algorithm randomly sets K centers and creates K clusters by allocating each observation to the closest center. In our case, the closeness is obtained by the squared Euclidean distance between the transformed covariates of observation i of length p' ($y_{\cdot i}$) and the according cluster center $m_{\cdot k}$. In the second step, the center of each cluster is recalculated and new clusters are built by moving the observations to the nearest center, called *centroids*. This step is repeated until the number of maximum iterations is reached, or the assignments remain the same after checking for each observation, whether the centroid of the allocated cluster C^* is still the closest, fulfilling the following criterion for each observation i (Hastie et al., 2009):

$$C^*(i) = \underset{1 \leq k \leq K}{\operatorname{argmin}} ||y_{\cdot i} - m_{\cdot k}||^2. \tag{2.19}$$

The k-means algorithm can be initiated with different starting values in order to check the stability of the obtained partitions. Different initiations are of high relevance giving

the chance to find the global minimum of Equation (2.19) over all observations ($i \in \mathcal{I}$).

Hierarchical clustering:
Hierarchical clustering contains two strategies: *agglomerative* and *divisive clustering*. As a starting point of the agglomerative clustering, each observation forms its own cluster. The remaining observations are then iteratively combined with the most similar observations until all observations are unified in one single cluster. The union of clusters can be obtained by *single linkage, complete linkage, or average linkage* (Hastie et al., 2009). In case of single linkage, the dissimilarity d of the closest observations of each cluster is considered. In case of complete linkage the maximal dissimilarity is considered[4]. Average linkage uses the average dissimilarity between the groups $\left(\frac{1}{N_{\mathcal{C}_i} N_{\mathcal{C}_j}} \sum_{i \in \mathcal{C}_i} \sum_{j \in \mathcal{C}_j} d_{ij} \right)$. $N_{\mathcal{C}_i}$ and $N_{\mathcal{C}_j}$ correspond to the number of observations that are part of of the underlying cluster $\mathcal{C}$ of observation i and j. Divisive clustering proceeds in reverse order, starting with all observations in one cluster. The subsequent splits aim at dissimilarity, such that the obtained clusters are most dissimilar to each other until each observation is part of its own cluster.

Community-detection-based clustering:
With increasing sample size of scRNA-seq data sets, both the k-means and the hierarchical clustering become computationally very expensive. The *community-detection-based clustering* aims at the identification of neighborhoods, which are connected groups of observations (Kiselev et al., 2019). The idea is related to a graph-based clustering. In our case the nodes of a K nearest neighbors graph represent single-cells and the edge weights contain the pair-wise distances of single cells (Zhang et al., 2020).

Density-based clustering:
Density-based clustering considers highly dense regions of observations as clusters, which are aimed to be well separated from low density regions. The approach of the *density-based spatial clustering of applications with noise (DBSCAN)*, introduced by Ester et al. (1996), needs a minimum number of observations forming the circle of the considered similarity. The advantage of DBSCAN is that the obtained clusters can be of different size and shape, allowing to handle outliers. As stated by Hennig (2015), the choice of a certain clustering method depends on the research objective. In contrast to identifying clusters of subpopulations that are assumed to be represented in a balanced way, algorithms such as GiniClust (Jiang et al., 2016), and RaceID (Grün et al., 2015) have been proposed in the context of single-cell clustering, in order to identify rarely represented subgroups. This is of special interest in cancer research, as rare cell types might be (cancer) stem cells or circulating tumor cells (Grün et al., 2015).

[4]Dissimilarity measures for quantitative variables can be for example the absolute difference, Euclidean distance, or the Pearson correlation.

Cluster validation

Apart from choosing a clustering algorithm, cluster validation plays an important role for evaluating the quality of the obtained clustering result (Liu et al., 2010) because especially in scRNA-sequencing analysis the subsequent downstream analysis relies on them. Analyzing the low-dimensional data after dimension reduction by visual inspection aims to get insights whether distinct patterns can already be detected. *Visualization validation* techniques, such as biplots of the PCA or CMDS deliver a first impression about the underlying grouping structure. In bioinformatics applications, visual inspection has often been applied. However, these inspections are highly subjective, and might lead to biased decisions (Handl et al., 2005). Therefore, validation measures are needed for preventing the subjective influence especially in gene expression data. For that purpose there mainly exist two categories of validation procedures (Halkidi et al., 2001; Handl et al., 2005): *external*, and *internal validation*.

External validation:
External validation includes the gold standard or ground truth, which is a rarity in practice, as normally the class labels are not known (Liu et al., 2010; Handl et al., 2005). Thus, the external information is not used for clustering but is included for the subsequent validation. The adjusted Rand Index (ARI) (Hubert and Arabie, 1985; Rand, 1971), purity (Rendón et al., 2011), normalized mutual information (Strehl and Ghosh, 2002) and the F-measure (Larsen and Aone, 1999) are examples of external validation measures. Since information about the aimed partitioning is available, the number of clusters doesn't have to be determined as the "true" number is known (Liu et al., 2010). It is of higher relevance to choose the optimal clustering algorithm for the investigated situation.

Internal validation:
Internal validation assesses separateness and compactness based on quantitative measures such as distance or variance measures of the defined clusters (Liu et al., 2010). For example the Davies-Bouldin index (Davies and Bouldin, 1979), and the silhouette index (Rousseeuw, 1987) target both separation and compactness. While the silhouette index determines separation and compactness observation-wise, the Davies-Bouldin index considers the different clusters. Furthermore, there exists also the special case of cluster validation based on stability (see e.g. Ben-David et al. (2006); Handl et al. (2005); Ullmann et al. (2021)). For the stability investigation, the same clustering algorithm has to be applied several times to different data situations. A good stability can be obtained if different data sets with the same data distribution result in similar partitions. Accordingly, data sets are re-sampled, artificially perturbed, or already clustered data are clustered again in order to investigate stability.

Apart from finding the best partition, internal validation also serves for the determination of the number of underlying clusters K. In general, the partition with the highest compactness and separation leads to the choice of K. Also, the sum of squares plot or di-

rectly the internal validation measure plot among all possible Ks might support the choice for the number of clusters. Alternatively, the Gap Statistic, proposed by Tibshirani et al. (2001) might be helpful. The latter is an often applied approach in bioinformatics (Handl et al., 2005).

A difficulty of internal validation is that some algorithms are specifically targeted to some of the validation criteria, such as the k-means algorithm which is focused to reach maximal compactness. As compactness is also part of internal validation, the underlying cluster algorithm should be taken into account during the validation process. Hennig et al. (2015) claim that it is of high relevance to have a qualified understanding of the data situation, the vised clustering target, and the clustering methods which are available in order to obtain a suitable application. Especially *Contribution 2*, *Contribution 3*, and *Contribution 4* will develop on this issue. *Contribution 2*, which proposes an association-accuracy heuristic builds on the dilemma of "true" versus "real" clusters, as discussed by Hennig (2015).

2.3 Single-cell consensus clustering (SC3)

The *single-cell consensus clustering (SC3)* of Kiselev et al. (2017), which will be the foundation of *Contribution 1* and *Contribution 2*, has been cited a lot and is considered as one of the most appropriate clustering methods for single cells (Duò et al., 2020). The clustering of SC3 mainly builds on the principal component analysis (PCA) and Laplacian eigenmaps (LE), applied to the preprocessed dissimilarity matrices, as described in the paper of Kiselev et al. (2017) and *Contribution 1*. To each combination of dissimilarity matrix and dimension reduction technique, an eigenvalue decomposition is applied, resulting in the eigenvectors Ψ_{DR}. Then, an automatic selection of eigenvectors is done, generating different low-dimensional data sets[5] $\mathcal{Y}^{(DR)}$ to which a k-means clustering is applied, using the algorithm of Lloyd (1982). Based on all the performed k-means clusterings, which can be considered as ensembles, a consensus matrix is built, including the average value of how often each pair of observations was clustered together. These frequencies are denoted as consensus values and form the resulting consensus matrix for all pair-wise combinations. In the default setting, the final clustering result is obtained by hierarchical clustering, applying complete linkage to the consensus matrix, choosing the partition with the number of clusters used in k-means.

[5]The choice of the number of selected ordered eigenvectors depends on the number of total observations N. In detail, this means that a clustering is performed on the rounded 4% of N, increasing the number of included eigenvectors by 1, until the rounded number of 7% of N is reached, for each of the six combinations of different transformations. However, this guideline applies only if the number of k-means clusterings which have to be applied by one combination remains below 15. Otherwise, for the clustering 15 first eigenvectors are randomly selected out of the obtained range.

3 Summaries and perspectives of the contributing material

In the following, we provide a detailed overview of each contribution, summarizing its main content, describing the analysis, and reporting the most important findings. After each summary, we comment on the submission and discuss possible perspectives for further research. All four contributions propose methods for the analysis or prevention of uncertainties in the context of clustering high dimensional single-cell data during the workflow of single-cell RNA-sequencing data analysis (see Figure 3.1). The figure shows

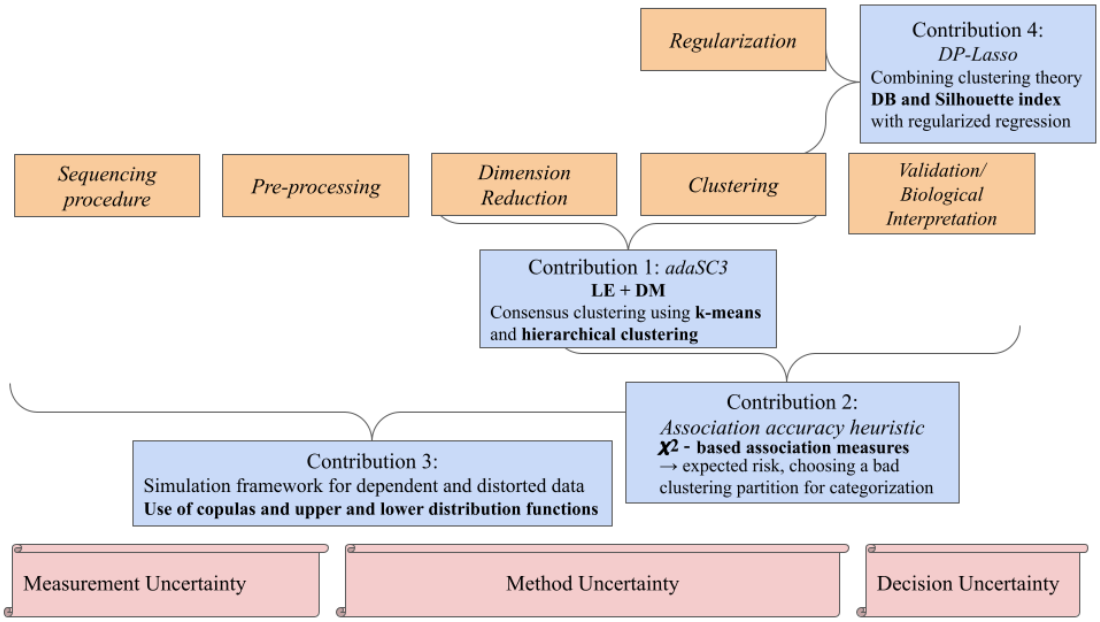

Figure 3.1: Connections of Contributions 1 to 4 to the underlying workflow of the single-cell RNA-sequencing data analysis.

the interrelationships of the individual contributions, targeting high dimensional clustering in *Contribution 1* to *Contribution 3*, and using cluster indices for regularization in *Contribution 4*. Especially *Contribution 1* and *Contribution 2* target method uncertainty, including decision uncertainty in *Contribution 2*. *Contribution 3* incorporates measurement uncertainty, and *Contribution 4* aims at leading regularization techniques into the right direction, reducing method and decision uncertainty.

3.1 Contribution 1: Method uncertainty meets biological background

Fuetterer, C., Augustin, T., and Fuchs, C. (2020). Adapted single-cell consensus clustering (adaSC3). *Advances in Data Analysis and Classification*, 14(4):885–896.

3.1.1 Summary

In *Contribution 1*, we start with a methodological improvement of the introduced single-cell consensus clustering (SC3) of Kiselev et al. (2017) by replacing the principal component analysis with diffusion maps. As already argued in Chapter 1 and Chapter 2, dimension reduction combined with clustering is of high relevance in the research field of single-cell RNA-sequencing data. Especially with regard to the manual determination of cell types, appropriate clustering algorithms are needed, but also the detection of newly defined cell types might be of interest to find new marker genes. This could lead to new biomarkers, which aim to serve as predictive or prognostic markers of diseases. In the context of single-cell consensus clustering, studies such as Duò et al. (2020), Freytag et al. (2018), Menon (2018) state that the approach of Kiselev et al. (2017), which includes a combination of principal component analysis and Laplacian eigenmaps as dimension reduction techniques, is part of the most recommended methods.

Replacement of principal component analysis by diffusion maps in SC3

We propose adapting the original SC3 (see Section 2.3) better to the biological background of single cells. The principal component analysis (PCA) is replaced by the diffusion maps (DM). Following the hint of Haghverdi et al. (2016), the differentiation of single cells should not be considered as a linear continuous process (Bendall et al. (2014); Buettner and Theis (2012)). Therefore, the PCA as linear dimension reduction technique is considered as not appropriate and with diffusion maps, we respect the transition of single cells, passing from one state to another in a non-linear way. With the inclusion of diffusion maps we thus adapt SC3 more to an explicit application of single-cell data resulting in the proposed *adapted single-cell consensus clustering (adaSC3)*. AdaSC3 contains Laplacian eigenmaps (LE) and diffusion maps (DE) as two non-linear dimension reduction techniques compared to the original SC3, which includes the linear PCA and the non-linear LE. Replacing the linear PCA by the non-linear DM for the combination of clustering results with those obtained by the LE still allows performing the subsequent clustering, similar to SC3. Thus, in general, the obtained clustering result enables to catch the decisive states, in which specific biological functions start to develop or benign cells mutate into malignant cells. Apart from the change of the dimension reduction method, adaSC3 follows the same procedure as SC3. Accordingly, adaSC3 includes the same pair-wise distance matrices for each embedding. We thus take the same dissimilarity measures for single cells into account, constructing a corresponding graph on which we base the low dimensional embedding of diffusion maps.

Results of real data and the simulation study

The investigated real data sets, which are also part of the original benchmarking of Kiselev et al. (2017), as well as the constructed simulation data show a better performance of the proposed *adaSC3* compared to SC3. Apart from replacing the PCA with DM, we also considered the dimension reduction techniques classical multidimensional scaling (CMDS), isomaps (IM), and locally linear embedding (LLE) of Section 2.1 as competitors, which are applied to the same dissimilarity matrices as part of the normal SC3.

For the construction of simulation data, we considered the zero-inflated negative binomial (ZINB) distribution as stated in Section 1. For each covariate $x_{j\cdot}$ with $j \in \{1, \ldots, p\}$ we generate the simulated gene expression with a mixture of the negative binomial distribution (NB) and zero-inflation parameter ϕ_j for each covariate j:

$$f_{ZINB}(X_j = x_{j\cdot}) = \begin{cases} \pi_j + (1-\pi_j) f_{NB}(0) & \text{if } x_{j\cdot} = 0 \\ (1-\pi_j) f_{NB}(x_{j\cdot}) & \text{if } x_{j\cdot} \in \mathbb{N}. \end{cases} \tag{3.1}$$

With the generalized form of the negative binomial distribution a gamma distribution Γ of the Poisson rate is included into the mixture of Poisson distributions in Equation (3.2). Following Zeileis et al. (2008), the size parameter of each covariate ϕ_j can be considered as continuous in the part of the negative binomial (NB) distribution. Furthermore, the expectation parameter of each covariate is denoted with μ_j:

$$f_{NB} = f(x_{j\cdot}|\mu_j, \phi_j) = \frac{\Gamma(x_{j\cdot} + \phi_j)}{\Gamma(\phi_j) x_{j\cdot}!} \cdot \frac{\mu_j^{x_{j\cdot}} \cdot \phi_j^{\phi_j}}{(\mu_j + \phi_j)^{x_{j\cdot} + \phi_j}}. \tag{3.2}$$

The according ZINB has been used for simulating two groups with all possible partitions of N observations. Both the expectation and the size parameter are considered to have an impact on the cluster performance and have been included into an intensive simulation framework provided in Section 5 of the original contribution. We can state that *adaSC3* reached the overall highest performance measured by the adjusted Rand index (ARI) compared to all included competitors. Concerning the benchmark data, we have shown in four out of five cases that *adaSC3* results in better or equal performance than SC3 and reaches the highest ARI in three out of five cases among all competitors. We can further state that *adaSC3* performs considerably better in the case of extremely unbalanced settings, especially observed for the simulated data. This result corresponds to the findings of Coifman and Lafon (2006) that the upper and lower tails of the distribution can be maintained very well by diffusion maps. All in all, we come to the conclusion that *adaSC3*, including two non-linear DR techniques, reaches a methodological improvement and leads to more accurate and stable results, while the biological background of single-cells is respected.

3.1.2 Comments and perspectives

AdaSC3, the extension of SC3 was developed following the same framework as proposed by SC3 and takes the biological structure of single-cells into account. As described in Section 2.3, the original SC3 performs an automatic selection of eigenvectors for dimension reduction based on the total number of observations N. Adapting SC3 and its extension *adaSC3* more to the statistical standards, one could replace the automatic selection of p' components with some statistically motivated tools, such as those mentioned in Section 2.1.

Concerning the clustering part, as further investigation it could be interesting to replace the k-means algorithm by density based clustering, such as DBSCAN whose strength lies in the detection of rare cell clusters. Setting a focus more to rarely represented cells in tissue samples is of special interest for example for the investigation of how healthy cells turn into cancer cells. The correct detection of mutated single cells is hereby of high relevance, especially during the first stages of cancer development.

Related to that, clustering single-cell data serves as a data-driven approach for the determination of cell types. It is thus of interest whether the visual inspection corresponds to the data-driven manual inspection and vice versa. The identified groups and the according marker genes allow to support or counteract cell development obtained by visual inspection. If the data-driven partition corresponds to the visual findings, one could replace the visual grouping by the manual grouping, which is especially efficient in high dimensions of single-cell data. Otherwise, the detected data structure might lead to the discovery of new marker genes. Whether the clusters found by *adaSC3* contribute to the findings of new or more trustworthy marker genes compared to SC3 is not part of *Contribution 1* as the focus of the described contribution lies on the methodological improvement. As biomarkers are developed on marker genes, our method might also provide new insights into that research area. Unfortunately, this has not been investigated within this dissertation. The according field of biomarker discovery has already brought and will further deliver a lot of benefits with regard to drug development - however ethical drawbacks have to be kept in mind.

As stated by Duò et al. (2020) not every combination of clustering results in a higher performance but for *adaSC3* the chosen combination leads to an improvement. Furthermore, Kiselev et al. (2017) argue that the original SC3 consensus clustering leads to stable results. The according stability can be reached by first combining several k-means clusterings on which a hierarchical clustering is performed. The according combinations can thus prevent unstable results of k-means. Another aspect that has been investigated in SC3, as well as in the first proposal of consensus clustering, introduced by Monti et al. (2003), is the determination of the number of clusters K. As the consensus value indicates the relative frequency for each pair being clustered together, the consensus matrix gives an insight into the stability of clusterings for different numbers of chosen partitions K. The optimal choice of K according to stability validation would be the one with the most stable clustering. The according internal validation measure of the consensus value, which is reclustered by

hierarchical clustering aims at a highly stable result as it includes several k-means clustering results obtained from different (data) transformations. The most stable clustering is obtained in case that either each pair is always or never clustered together. The same thought can be transferred to applying different clustering algorithms and considering the consensus of different methods on the same data set. This could bring along much effort in the subsequent biomarker discovery step, bringing us to the content of *Contribution 2*.

3.2 Contribution 2: Expert decisions meet clustering decisions

Fuetterer, C., and Augustin, T. (2021). Internal Validation of Unsupervised Clustering following an Association Accuracy Heuristic. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM): Workshop on Machine Learning and Artificial Intelligence in Bioinformatics and Medical Informatics (MABM 2021)*, 2201–2210.

3.2.1 Summary

With *Contribution 1* we provided another competitor to the original SC3. However, as stated in Section 1.1, there exist many unsupervised clustering methods, on which subsequent decisions depend on, such as the determination of cell types. In general, clustering is performed if the true underlying grouping structure is unknown or hardly accessible. In case that no external information is available, internal validation of clustering is often the only way to validate the clustering. In *Contribution 2*, we relate the consensus of expert opinions to the agreement of clustering results. We therefore propose an association-accuracy heuristic, which allows to judge the trustworthiness of different clustering methods. With the construction of association measures, we can estimate the risk of choosing a bad clustering among the considered well-established methods for a given application.

Association accuracy heuristic

Inspired by decision making theory in which the opinions of experts are considered to be independent of each other, one has higher confidence in decisions in which all experts share the same opinion. In this case it is common standard to trust unanimous decisions and no further studies are requested. However, the more heterogeneous the experts' opinions are, the less one trusts them, and caution is required. In these heterogeneous settings, further studies would be required for identifying the best choice.

Relating to the idea of the consensus of experts, we state that if different well-established methods target the same application situation and deliver comparable clustering results, we are close to the true underlying groupings. Referring to decision theory, where experts are needed for proper decision making, we request well-established methods. These have to

be well defined and should not contain stupid methods which would immediately destroy our heuristic because the consensus of trivial estimates would not contribute to a reliable decision. Under these assumptions, we propose the following association accuracy heuristic, introduced in Section I of the original contribution (p. 2202):

> *"The degree of association between clustering results derived by different methods is an indicator for the extent of trustworthiness of the results. Under high association, a high accuracy of each method can be expected, while lowly associated clustering results indicate a high risk of choosing a method with bad performance."*

For being able to judge the described association of clustering results, we propose χ^2-related association measures. With the consideration of several clustering methods, the underlying concept can be referred informally to the approach of bootstrapping where the idea is to pull oneself up by its own bootstraps, generating more samples for an adequate efficient testing procedure. With our approach, we also generate more information being able to assess the degree of concordance between the partitions obtained by different methods. We refer to the groupings as trustworthy in case that the different methods deliver highly associated clustering results described by the following association measures.

Construction of association measures

For the measurement of associated clustering results, we propose χ^2-based association measures as internal validation to which we refer as *method-association-measures*. In contrast to the classical compactness and separation criteria within one clustering method, we consider the (dis)similarity of the clustering results, obtained by different methods. For analyzing the association of individual groupings among different (clustering) methods we propose a reinterpretation of the *adjusted contingency coefficient C*, and *Cramér's V*. For the pair-wise evaluation, we consider the Φ-*coefficient*, where it is only of interest whether a pair of observations is clustered together or not among the clustering results of two different methods M_k and M_l being elements of the set of all methods $\mathcal{M} = \{M_1, ..., M_q\}$. The vectors $\mathcal{P}_{M_k}$ and $\mathcal{P}_{M_l}$ of method M_k and M_l contain the (dis)similarities of each pair leading to the following association measure constructed on $\binom{N}{2}$ pairs of observations:

$$\Phi(\mathcal{P}_{M_k}, \mathcal{P}_{M_l}) = \frac{\chi^2(\mathcal{P}_{M_k}, \mathcal{P}_{M_l})}{\binom{N}{2}}, k \neq l, k, l \in \{1, ..., q\}, \tag{3.3}$$

With the proposed association measures we directly take into account the number of observations N and the number of clusters K. As we apply the Φ-coefficient to 0/1-coded variables, the Φ-coefficient is equivalent to the Pearson correlation as proven by Cramer (1946) and can thus be interpreted in the same way. The proposed method-association-measures allow the user to be aware of the risk choosing a partition that might lead to low accuracy. In addition, comparing clustering results with differing K might lead to a selection of K, which represents the highest association of clustering methods.

Results of simulation study and real data

For the investigation of our heuristic for both simulated data and real scRNA-seq data, we first assess the association of context-related methods. Apart from SC3 and adaSC3 we further include the *Potential of Heat-diffusion for Affinity-based Trajectory Embedding (PHATE)* (Moon et al., 2019), as well as *t-distributed stochastic neighbor embedding (t-SNE)* (Van der Maaten and Hinton, 2008) and *Uniform Manifold Approximation and Projection (UMAP)* (McInnes et al., 2018) in combination with k-means. As simulation data we generate correlated groups as described in Section IV of *Contribution 2*, which can be reached assuming a multivariate normal distribution. With increasing correlation, we aim to simulate increasing dependence of the simulation groups. From these scenarios, we expect that lowly correlated simulation groups achieve highly associated clustering results and highly correlated simulation groups achieve a lower association among the clustering results. Including the underlying cluster labels after the investigation of association allows the judgement that stronger dependence between the different simulation groups leads to less accuracy of the considered clustering methods. The results obtained by the bachelor thesis of Stermann (2021), which was supervised by Thomas Augustin and me, confirm the expectations that also in the case of simulating ZINB data, lower correlated simulation data result in higher performance. In *Contribution 2* we further showed that it was possible to rank both the real data sets as well as the simulated data sets according to their median of associated methods. The ranking correlated strongly with the performance ranking of the different methods. We thus argue that our association accuracy heuristic can provide valuable insights, delivering method-association-measures that are indeed an indicator for expected accuracy, maintaining an internal validation measure.

3.2.2 Comments and perspectives

Referring to the above stated situation that no external validation is feasible for the obtained clustering situation, one has to rely on internal validation measures, such as compactness and separation. As stated in Section 2.2, certain algorithms are constructed to fulfill some of the requested criteria. Instead of blindly trusting one clustering algorithm, we propose investigating several methods, which are well-established for the application at hand. Comparing the results of the different algorithms with independent experts we claim that our association measures provide a deeper insight into the expected accuracy of clustering results. However, as truth is not known, our heuristic can only provide a measurement of an estimated risk. Thus, our heuristic is the best you can hope for as no prediction is possible with missing truth, as no model can be trained in this case.

The constructed method-association-measures allow an internal validation that is also comparable across different data sets with differing K and N. The proposed association measures can be connected to the assumption underlying the approach of general consensus clustering, introduced by Monti et al. (2003). Instead of classical stability validation for example by resampling procedures, we analyze different methods on the same data sets to

get insights into the (dis)similarity behavior of methods. This view could relate the proposed measures also to stability validation, which can be considered as a part of internal validation (see Section 2.2). Furthermore, with our association measures we are able to identify situations with low and high risk, completely trusting data-driven partitions. The distinction of strongly and lowly associated methods is especially of interest with regard to increasing sample size, leaving only lowly associated situations to the work of experts with the indication that trusting one single method in a situation with lowly associated methods is a risky choice.

We consider our heuristic as generally and easily applicable, as well as quite useful and powerful. This is especially true for situations where the ground truth is extremely costly to generate for external validation, like for single-cell RNA-sequencing data. Nevertheless, domain experts are still needed, choosing appropriate methods for assessing the association of clustering results and deciding in which situation it is better to consult an expert in order to prevent blind trust into one single method or our heuristic. During the process of developing the framework, we observed that clustering methods that were most associated over all methods also had the tendency of highest performance. However, this result is not part of published work but could open the pathway to selecting the best method in a completely unsupervised manner.

In case of an overall low association of clustering results, we protect ourselves by considering these situations as not trustworthy. As we incorporate all pair-wise combinations of methods into our application, we build our assessment on stable associated clustering results. However, a specification of a lower bound for the number of included well-established methods q would be desirable, which we expect to be different depending on the application areas.

3.3 Contribution 3: Uncertainty meets measurement distortion

Fuetterer, C., Schollmeyer, G., and Augustin, T. (2019). Constructing Simulation Data with Dependency Structure for Unreliable Single-Cell RNA-Sequencing Data Using Copulas. In De Bock, J., de Campos, C., de Cooman, G., Quaeghebeur, E., Gregory Wheeler, G. editors, *Proceedings of the Eleventh International Symposium on Imprecise Probabilities: Theories and Applications. PMLR*, 103:216–224.

3.3.1 Summary

In the previous contributions a new clustering method has been proposed as well as an association accuracy heuristic, which allows to relate the clustering results of the different methods to each other. With the simulation framework presented in *Contribution 3*, it is possible to analyze the influence of distorted measurements, including dependence structures of covariates. We therefore simulated downward and upward distorted measurements based on upper and lower distribution functions. Furthermore, we respected the multivariate dependence of genes by including copulas into the simulation process. Apart from the uncertainties within a method, which played a role in the last two contributions, we investigate technical uncertainty in this contribution. This uncertainty can occur during the sequencing process due to e.g. the amplification rate, missing gene expression, or measurement instrument variation. But also, stochastical noise and imputed values might contribute to distorted data. In accordance with the contributions above, this work has also arisen in relation to single cells. With this contribution we provide a framework which allows an intensive study of clustering performance comparing simulation groups of 1) undistorted measurements with each other, simulating a homogeneous, an intermediate, and a heterogeneous scenario, 2) upper and lower distorted measurements to no distortion, and 3) include the influence of dependence structures by different copula families.

As orientation for the construction of the upper and lower parameter ranges of undistorted measurements, a zero-inflated negative binomial (ZINB) distribution is assumed with parameters estimated from the real single-cell RNA-sequencing data of (Kolodziejczyk et al., 2015b), simulating two target groups ($g = 2$).

Undistorted measurement scenarios

For the simulation setting within this contribution we first construct three undistorted simulation scenarios with $l = \{1, 2, 3\}$, including different ranges of varying parameter intervals, which overlap more and more with increasing heterogeneity. The first scenario, which is also the most homogeneous setting contains the smallest range for expectation parameter values μ_1, size parameter values ϕ_1, as well as for the fraction of zero-inflation

values π_1. The second scenario (μ_2, ϕ_2, π_2) includes broader bounds for each parameter range, and the third scenario (μ_3, ϕ_3, π_3) has the broadest parameter range. The first scenario represents a homogeneous setting, followed by an intermediate setting in the second scenario, and the third scenario consists of the most heterogeneous setting, which leads to the overall parameter set of simulation group 1:

$$\theta^{(1)} = \{\mu_1^{(1)}, \phi_1^{(1)}, \pi_1^{(1)}, \mu_2^{(1)}, \phi_2^{(1)}, \pi_2^{(1)}, \mu_3^{(1)}, \phi_3^{(1)}, \pi_3^{(1)}\},$$

and overall parameter set of simulation group 2:

$$\theta^{(2)} = \{\mu_1^{(2)}, \phi_1^{(2)}, \pi_1^{(2)}, \mu_2^{(2)}, \phi_2^{(2)}, \pi_2^{(2)}, \mu_3^{(2)}, \phi_3^{(2)}, \pi_3^{(2)}\}.$$

For group 1, and scenario l, a $ZINB(\mu_l^{(1)}, \phi_l^{(1)}, \pi_l^{(1)})$ is generated. For group 2 a different range of parameters is included into the according distribution $ZINB(\mu_l^{(2)}, \phi_l^{(2)}, \pi_l^{(2)})$ for simulating the covariates of each scenario. The obtained simulation groups are then combined for each scenario, leading to the simulation setting of independent, undistorted scenarios.

Lower and upper distribution functions

Based on the undistorted scenarios, we construct upper and lower distorted count data using the concept of lower and upper distribution functions as described by Montes et al. (2015) in the context of imprecise probability theory. For that purpose, we consider a set of distribution functions $\mathcal{F}_j^{(g)}$, specific to target group g and covariate j. The aimed lower and upper distribution functions $\underline{F_j^{(g)}}(x_{j\cdot})$ and $\overline{F_j^{(g)}}(x_{j\cdot})$ are thus constructed by the extraction of the infimum and supremum of $F_j^{(g)} \in \mathcal{F}_j^{(g)}$ for the considered covariate value $x_{j\cdot}$, which are again distribution functions (Montes et al., 2015):

$$\underline{F_j^{(g)}}(x_{j\cdot}) = \inf\{F_j^{(g)}(x_{j\cdot}) : F_j^{(g)} \in \mathcal{F}_j^{(g)}\}, \tag{3.4}$$

$$\overline{F_j^{(g)}}(x_{j\cdot}) = \sup\{F_j^{(g)}(x_{j\cdot}) : F_j^{(g)} \in \mathcal{F}_j^{(g)}\}. \tag{3.5}$$

In the applied simulation framework the lower and upper distribution functions are obtained by choosing the infimum and supremum of the empirical cumulative distribution functions leading to $\underline{\hat{F}_j^{(g)}}$ and $\overline{\hat{F}_j^{(g)}}$. The simulated read counts $x_{j\cdot}$ of each target group g of the simulated undistorted scenarios $(l = 1, 2, 3)$ are obtained with:

$$\underline{\hat{F}_j^{(g)}}(x_{j\cdot}) = \inf_{l=1,2,3} \hat{F}_j^{(g)}(x_{j\cdot} \mid \theta_l^{(g)}), \tag{3.6}$$

$$\overline{\hat{F}_j^{(g)}}(x_{j\cdot}) = \sup_{l=1,2,3} \hat{F}_j^{(g)}(x_{j\cdot} \mid \theta_l^{(g)}). \tag{3.7}$$

The empirically based lower and upper distribution functions allow the simulation of upper

and lower distorted measurements. The artificial construction of distorted data leads to the proposal how to simulate unprecise data. The obtained lower distribution function represents situations, in which for each covariate higher count data are measured, whereas the upper distribution function reflects situations measuring lower count data. As the upper and lower cumulative distribution functions do not follow a ZINB distribution anymore, we underlie the upper and lower quantile functions as marginal distributions. For the construction of dependent simulation data, using copulas, we maintain the underlying marginal distributions, which results in a ZINB for the non-distorted simulation scenarios.

Simulating dependence structure using copulas

For the integration of the dependence structure of the single-cell data of (Kolodziejczyk et al., 2015b), we use copulas fulfilling the attributes described by Nelsen (2007). Copulas allow to generate a joint distribution, maintaining the specified marginal distributions for each covariate. As we build each simulation group on its own marginal distribution, we are able to underlie the same group-specific dependence structure. A copula of a pairwise dependence structure aims to map two distribution functions to one joint distribution function. Also, in case of a high number of covariates p, it is possible to find a joint distribution function F_X, maintaining the univariate marginals, leading to a copula C of a specific family v (see theorem of Sklar (1959)):

$$F_X^{(g)}(x_{1\cdot}, \cdots, x_{p\cdot}) = C_v(F_1^{(g)}(x_{1\cdot}), \cdots, F_p^{(g)}(x_{p\cdot})). \tag{3.8}$$

As a next step copulas were applied to the above defined lower and upper distribution functions $\overline{F_j^{(g)}}$ and $\underline{F_j^{(g)}}$. We therefore adapt Sklar's theorem according to Montes et al. (2015) and Škulj (2018) leading to joint distribution functions of the distorted measurements. For the applied simulation study, copulas are built on the estimated cumulative distribution functions of each scenario, allowing the artificial construction of upper and lower distorted measurements, including the same multivariate dependence structure as in the undistorted case. In our study, we consider the Gaussian copula, the Clayton copula, and the Frank copula as possible copula families v.

Results of simulation data including the dependence structure of real data

The simulation framework thus consists of the undistorted simulation settings, considering both an independence structure and the different dependence structures obtained by using the three different copula families. These settings are also investigated with regard to the distorted simulation data. All simulation data contain two balanced target groups ($g = 2$) with differing p. To each setting, a k-means clustering is applied with $K = 2$. For the undistorted settings, we expect that with increasing heterogeneity, the underlying data structure can be better detected. These expectations hold true for the settings with independent simulation groups, as well as for the Gaussian copula. In contrast to the lower distorted setting, the upper distorted setting still leads to a good clustering performance.

This is especially true for the independent setting as well as for the Frank copula. The downward distortions result in considerably poorer clustering performance. Apart from that scenario, an increase in sample size has the tendency to reach a higher clustering performance for all considered dependence structures.
All in all, the proposed simulation framework constructed for count data of single cells allows the investigation of dependence structures in scenarios with increasing heterogeneity and to analyze the consequences of distorted data sets, which can occur during the technical preparation of the data.

3.3.2 Comments and perspectives

With the proposed simulation framework, a complex simulation study is possible with regard to different degrees of heterogeneity providing the simulation of distorted data. With the lower distribution function, we simulate the tendency to measure higher read counts, as the according cumulative distribution function leads to a lower probability for lower read counts compared to the upper distribution function. The opposite holds for the upper distribution function. Comparing the lower and upper distorted measurements, the latter are limited by the supremum of the simulated undistorted settings, and thus allow more heterogeneity between the groups. In comparison to that, the value range of the lower distorted setting is limited by 0, and thus the two underlying groups cannot be distinguished appropriately. Of course, the degree of heterogeneity as basis of the undistorted settings can be left to the user. Alternatively, repeated measurements could be included as a possible set of distribution functions. The higher the heterogeneity of the underlying settings, the higher the probability that the resulting distorted settings include extreme values. Nevertheless, with these adjustment screws, the consequences of heterogeneity can be investigated directly. Respecting the multivariate dependence of gene data, we can extend the simulation settings by including copulas of different families maintaining the marginal distribution of each gene. Thus, the influence of dependence can be simulated, including an analysis of its consequences for the clustering performance. We consider our complex simulation framework as possibility for an extensive study of methods allowing both the incorporation of specific uncertainties and different dependence structures.

Including repeated measurements as foundation of the undistorted scenario provides insights into measurement variation, which would be especially interesting in case that a gold standard could be included as well. Within our framework, the repeated measurements determine a possible range of the measured gene expression. Furthermore, the construction of the downward and upward distorted measurements based on repeated measurements, compared to a possible gold standard might provide hints for calibrating the measurement device. In contrast to adding a constant term as measurement error, or underlying a certain distribution for simulating measurement distortions, the proposed framework allows the simulation of complex data structures. This is also of special interest for investigating more complex algorithms compared to k-means, especially in combination with the non-linear dimension reduction techniques provided in Section 2.1. The measurements

of a biased instrument would influence a subsequent dimension reduction in case of no constant distortions. With our contribution we provide a complex simulation framework, investigating the measurement uncertainty with regard to the clustering performance of different high dimensional clustering approaches. Referring to the underlying distribution assumption within this dissertation, the simulated distorted measurements can take into account the differing parameters of the ZINB distribution. Simulating genes with increasing zero-inflation rates could show an influence of the high amplification rate during the measurement of single cells, which lead to a high number of missing measurements. This would be the extreme case of the upper distribution function. In contrast to that, expected dispersion and the effect of extreme outliers could be analyzed with the lower distribution function providing an upper range of possible measurement uncertainty.

For further research, the result of including specific dependence structures might be of interest as it adapts the simulation settings as close as possible to the underlying data structure. A specific research project could further analyze imprecise copulas, allowing other distributions, such as the Poisson distribution, which is also discussed as being appropriate for single-cell data. Our approach is not limited to the application of single-cell data and can be generalized and adapted to other underlying distributions. It is also possible to investigate more than two subpopulations. However, the inclusion of possible scenarios is limited because the higher the number of scenarios, the more extreme the according upper and lower distribution functions.

3.4 Contribution 4: Clustering information meets regularized regression

Fuetterer, C., Nalenz, M., and Augustin, T. (2021). Discriminative Power Lasso – Incorporating Discriminative Power of Genes into Regularization-Based Variable Selection. Technical Report. Available under: `https://epub.ub.uni-muenchen.de/91666/1/DPL_TR_2022_03.pdf`

3.4.1 Summary

Apart from the unsupervised learning tasks, single-cell data can also be used for classification. With regard to their high dimension, it is of relevance to reliably extract only the most decisive genes. Especially for the development of biomarkers the selection of only a few genes significantly reduces the effort as fewer candidates have to be investigated during intensive additional studies. With *Contribution 4* we propose the *Discriminative Power Lasso (DP-Lasso)*, which allows to connect the grouping information of univariate covariates based on clustering theory with the approach of regularized regression.

Discriminative Power Lasso (DP-Lasso)

With the Discriminative Power Lasso (DP-Lasso) we aim at strongly penalizing covariates not directly contributing to the target variable. For the construction of the DP-Lasso we incorporate univariate information of covariates into the regularization process of the adaptive Lasso (Zou, 2006). The novelty of this method is to first tune a regularization model based on the training data by including the grouping quality of each covariate with regard to the target variable instead of univariate estimates as it is the case for the adaptive Lasso. We refer to the grouping quality, inspired by clustering indices, as *discriminative power (DP)*, providing hints to the multivariate model, which univariate information should be penalized less with regard to high compactness and separation. We then integrate the discriminative power as penalty factors into the adaptive LASSO, resulting in the *Discriminative Power Lasso (DP-Lasso)*. In accordance to the adaptive Lasso (Zou, 2006), the overall loss function, provided by the following equation has to be minimized:

$$L(z, X, \beta, \lambda, w) = \mathcal{E}(z, \hat{z}, \beta) + \sum_{j=1}^{p} \lambda_j |\beta_j|, \tag{3.9}$$

with $\mathcal{E}$ being the loss function of the true and predicted values of the response vector z and $\hat{z}$, obtained by the design matrix X and parameter vector β. The second term of Equation (3.9) is the penalization term, which is the sum of the products between the local shrinkage parameter λ_j for each covariate and the absolute regression coefficient over all p covariates. We replace the original penalty term of the adaptive Lasso by the covari-

ate specific DP, leading to the following penalization: $\lambda_j^{(DP)} := \lambda w_j^{(DP)} = \lambda \frac{1}{DP_j}$. Thus, with a higher discriminative power value of covariate j, denoted with DP_j, we aim at less penalization, and therefore the according discount factor $w_j^{(DP)}$ is smaller. In accordance with regularized regression models, the data are split into training and test data. For the construction of the discriminative power we include the true underlying class labels of z to assess compactness and separation of each covariate based on the training data. So, instead of the original clusters we investigate the K given classes of the target variable. With the cluster indices it is aimed to measure for each covariate the compactness within each category of the target group and the separation between two categories of the target group. We thus adapt the Davies-Bouldin (DB) index (Davies and Bouldin, 1979) and the silhouette index (Rousseeuw, 1987) to the underlying context. The according formulas for the covariate-specific Davies-Bouldin index DB_j and the silhouette index S_j can be found in Section III of *Contribution 4*. The adapted interpretation of the DB index $\in [0, 1]$ is that the lower its value, the more compact and the more separated are the underlying groups among the considered covariate. The resulting discount factor is thus as follows: $w_j^{(DB)} = DB_j$. The silhouette index is interpreted in the opposite way. The higher the absolute value of the silhouette index $S_j \in [-1, 1]$, the more compact and the better the underlying classes are separated by the investigated covariate, leading to the according discount factor $w_j^{(Sil)} = 1/|S_j|$.
As a second idea, we modify the underlying concept of the analysis of variance (ANOVA) (Fisher, 1992), such that it is possible to explore whether there is a difference of gene expression with regard to the categorical target variable. For the covariate specific weighting of the discriminative power, we consider each covariate as grouping factor. Thus, in our case, the according F-statistic is calculated among the true underlying groups $g = (1, \ldots, K)$, reflecting the ratio of between-group variability and within-group variability, including the group specific index h for each covariate value (x_{hj}), and the number of observations, which are part of the same group (n_g):

$$F_j = \frac{(N-K)\sum_{g=1}^{K} n_g(\bar{x}_{\cdot j}^{(g)} - \bar{x}_{\cdot j})^2}{(K-1)\sum_{g=1}^{K}\sum_{h=1}^{n_g}(x_{hj}^{(g)} - \bar{x}_{\cdot j}^{(g)})^2}. \tag{3.10}$$

Based on the obtained F-value F_j, we can state that the higher the F-value of covariate j, the more the mean values of the respective groups differ, as the degree of freedoms are the same for univariate investigations. The corresponding discount factor is thus defined with $w_j^{(ANOVA)} = \frac{1}{F_j}$. For a very low F_j value, the according discount factor results in extremely high penalization. In order to avoid numerical instabilities we therefore include the logarithmic transformation of $w_j^{(ANOVA)}$ as penalty factors.

The constructed discriminative power indices DP-L$_{DB}$, DP-L$_{Sil}$, and DP-L$_{ANOVA}$ based on the DB index, the silhouette index and on the ANOVA will be compared to the performance of Lasso, Elastic Net (Zou and Hastie, 2005), and adaptive Lasso.

Results of simulated and real single-cell data

Within our simulation study we simulated for a fixed number of N, and differing number of covariates p, a normal distribution with increasing heterogeneity $\sigma^2 \in \{1, 2, 3\}$. In each study, 10 relevant covariates are assumed to be differentially expressed genes, with differing expectation values of -1, and 1 for the corresponding subpopulations ($K = 2$). The remaining $p-10$ covariates are simulated with a mean equal to 0. For the validation of our proposed DP-Lasso models in comparison to its competitors we consider precision, relating the number of differently simulated covariates to the selected number of covariates of each model. In case of high precision, we can trust the model, selecting decisive covariates. Furthermore, we are interested in the recall, which measures the fraction of differently simulated covariates detected. Regarding the recall, Lasso, Elastic Net, and adaptive Lasso seem slightly better, which could mean that the discriminative power Lasso selects a too low number of covariates. With regard to precision, all DP-Lasso approaches result in the highest precision in case of $\sigma^2 = 1$. In case of $\sigma^2 = 2$ and $\sigma^2 = 3$, the DP-L$_{ANOVA}$ shows the highest precisions for each simulated number of covariates p.

Concerning the real single-cell RNA-sequencing data, the number of selected covariates is stable and considerably lower compared to all competitors. For the number of selected covariates, the ANOVA-based approach DP-L$_{ANOVA}$ selects the lowest number of covariates, whereas the adaptive Lasso reaches a sparser model compared to the remaining DP-Lasso models. In the binary classification task, the misclassification rate is very well comparable to Lasso, and Elastic Net but performs better than the adaptive Lasso. In case of the multiclass classification task the DP-Lasso models are considerably better than the adaptive Lasso but result in a tendency towards a higher misclassification rate compared to the Lasso and the Elastic Net.

3.4.2 Comments and perspectives

The proposed DP-Lasso allows a stricter selection of covariates maintaining a high performance of the regularization model. Adapting clustering indices instead of clusters to the underlying target category enables to investigate the impact of each covariate to compactness and separation of the target variable. We make use of explorative tools from cluster theory, assessing compactness and separation of the real target group instead of the original purpose, which assesses partitions obtained by clustering. These adaptions serve as discriminative power measures. We further incorporate a reinterpretation of the ANOVA as discriminative power measure. Instead of a univariate regression, we investigate how well the target variable is explained by each covariate. This is in contrast to the original ANOVA, which analyzes the influence of the grouping variable on the continuous target variable. Both the cluster indices and the ANOVA are not directly affected by the number of classes and are independent of a reference category. Only the according frequencies of each class have an impact on the determination of compactness and separation and thus on the discount factor, which includes the discriminative power. This can be seen as an advantage over the challenge described by Tutz and Ulbricht (2009) and Tutz et al. (2015),

where in case of a multiclass classification task each category is penalized, which brings along the decision problem whether the covariate itself remains part of the model.

Nevertheless, the applied discriminative power measures fulfill their purpose leading the multivariate model into the right direction and providing informative univariate weights as confirmed by the conducted simulation study. With the proposed DP-Lasso we use a soft filter instead of a pre-selection of genes. It is clear that soft filtering needs more computation time compared to a hard gene filter. However, in contrast to a hard filter no critical threshold has to be determined. Furthermore, the proposed method leads to stable and reproducible results and is efficiently implemented, as it makes use of the Lasso framework. In comparison to the competitors Lasso and Elastic Net, the DP-Lasso approach delivers a considerably lower number of selected covariates maintaining a comparable performance to its competitors. Furthermore, the DP-Lasso shows a high stability selecting a stable number of covariates. Especially with regard to precision, the DP-Lasso models perform quite well, whereas with regard to recall the competitors are slightly better, with the Elastic Net performing best. However, the relation of precision and recall is best met by the DP-Lasso approaches with a preference to the DP-L_{ANOVA}. Thus, the included compactness and separation leads for all DP-Lasso models to a considerably lower number of covariates, which is also decisive for the underlying classification task. As a further extension of our method, we could think of integrating the different discriminative power measures of covariates into the group Lasso, or alternatively into a multitask learning approach.

The combination of variable selection and regularization plays a very important role in genetics, that is why special attention has to be paid to the correlation of covariates such as investigated by Tutz and Ulbricht (2009) and Tutz et al. (2015). These approaches are needed to target the well-known problem of Lasso, which has stability problems in case of strongly correlated data. The combination with ridge regression, which uses a L2 norm instead of the L1 norm of Lasso, also aims to prevent the stability pitfalls by including the strong convex penalty term. The Elastic Net still chooses both covariates in case that they are strongly correlated. With group Lasso the problem of correlated covariates is aimed to be circumvented by penalizing groups of covariates instead of single ones. Alternatively, approaches which cluster correlated covariates are also often used. Examples include the *Cluster Elastic Net (CEN)* (Witten et al., 2014), and the *octagonal shrinkage and clustering algorithm for regression (OSCAR)* (Bondell and Reich, 2008). With DP-Lasso we can give different weights to highly correlated covariates if they behave differently with regard to the target variable. We therefore consider our approach as very promising, as also in case of a correlated design matrix a distinction of variables with higher importance to the target variable can be obtained. One step further is the *Integrative LASSO with Penalty Factors (IPF-Lasso)* proposed by Boulesteix et al. (2017), which penalizes different sources of data individually for regularized regression. In the investigated context of single-cell data it would definitely be of interest to include different data sources to which our approach should be extendable as well.

4 General concluding remarks

In this dissertation we provide methodological contributions, analyzing the consequences of measurement uncertainty, method uncertainty and its consequences to clustering performance. We investigate different dependence structures, including possible transitions of single-cells from one state to another. With the replacement of the principal component analysis of the original *single-cell consensus clustering (SC3)* (Kiselev et al., 2019) by diffusion maps, the underlying biological function of single cells can be taken into account. This leads to the proposed method *adapted SC3 (adaSC3)* in *Contribution 1*. For validating clustering partitions from different methods, we propose adapted association measures with the according *association accuracy heuristic*, provided in *Contribution 2*, preventing the risk of choosing a bad partition due to high method uncertainty. *Contribution 3* allows a construction of simulation studies with increasing heterogeneity or repeated measurements, which enables the generation of artificially lower and upper distorted data. In addition, complex dependence structures can be included into the proposed simulation framework for an intensive study of the performance of different clustering algorithms. The proposals of all three contributions allow to analyze the consequences of distorted single-cell RNA-sequencing measurements with well suited clustering algorithms and internal validation. In contrast to these completely unsupervised procedures, *Contribution 4* incorporates validation measures of clustering theory into the regularization framework, leading to a selection of a lower number of decisive covariates. Based on the training data, the test set of the investigated single cells can be classified into cell types in a supervised way.

In Chapter 3 concluding remarks and outlooks are provided for each contribution. A more general perspective is given concerning the future of this research field. As many single-cell RNA-sequencing studies have been conducted, and open science plays a very important role, a large number of data bases are shared, especially due to the overall aim of completing the Human Cell Atlas. The high dimensional clustering and its underlying workflow essentially support the manual determination of cell types. As stated in this thesis, the clustering of single cells considerably contributed to the labeling of cells as the foundation of prediction and might be the future methodology to determine biomarkers. Also, the labeled cell data bases obtained by manual or visual inspection provide classifiers for a joint classification analysis, with the expectation to gain consistent labels.

Clustering approaches as well as regularization techniques aim at identifying decisive genes of the different groupings. While single-cell clustering enables the detection of new cell types, classical regularization models are restricted to firmly fixed cell types. During the year 2018 there was an increasing trend of clustering single-cell RNA-sequencing data. This trend decreased in the year 2021, as stated by Zappia and Theis (2021). Nowadays there is more a tendency to integrate different samples into classification tasks. Due to the increasing data volume, classification tasks will replace cluster analysis. A future challenge will be an appropriate definition of the term cell type as it is not yet properly defined. Also due to the very complex designs, such as including several conditions, replicates, data sources, and different cancer types, future analysis will profit from the integration of suitable data sets for classification approaches. Multiple data sources such as genomics, proteomics, metabolomics or microbiomics data should be considered for classification tasks.

However, before further expensive studies are conducted, it might be beneficial to invest in some more benchmarking studies because of the high number of methods in this research field. Especially the open data and open source code might support this avenue.

The proposed contributions have shown that taking into account the data structure as well as having the application task in mind, delivers a great opportunity for the development and validation of methods with regard to accurate applications.

Further references

Angerer, P., Simon, L., Tritschler, S., Wolf, F. A., Fischer, D., and Theis, F. J. (2017). Single cells make big data: New challenges and opportunities in transcriptomics. *Current Opinion in Systems Biology*, 4:85–91.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396.

Ben-David, S., Luxburg, U. v., and Pál, D. (2006). A sober look at clustering stability. In *International Conference on Computational Learning Theory*, pages 5–19. Springer.

Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, 157(3):714–725.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123.

Boulesteix, A.-L., De Bin, R., Jiang, X., and Fuchs, M. (2017). IPF-LASSO: Integrative-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and Mathematical Methods in Medicine*, 2017:1–14.

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1095.

Briggs, E. M., Warren, F. S., Matthews, K. R., McCulloch, R., and Otto, T. D. (2021). Application of single-cell transcriptomics to kinetoplastid research. *Parasitology*, 148(10):1223–1236.

Buettner, F. and Theis, F. J. (2012). A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics*, 28(18):i626–i632.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420.

Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and Computational harmonic Analysis*, 21(1):5–30.

Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):224–227.

De la Porte, J., Herbst, B., Hereman, W., and Van Der Walt, S. (2008). An introduction to diffusion maps. In *Proceedings of the 19th Symposium of the Pattern Recognition Association of South Africa (PRASA 2008), Cape Town, South Africa*, pages 15–25.

Duò, A., Robinson, M. D., and Soneson, C. (2020). A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 7:1–23.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1):1–14.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of 2nd International Conference on Knowledge Discover and Data Mining*, volume 96, pages 226–231.

Fisher, R. A. (1992). *Statistical methods for research workers*. Springer.

Freytag, S., Tian, L., Lönnstedt, I., Ng, M., and Bahlo, M. (2018). Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Research*, 7:1–29.

Gan, Y., Huang, X., Zou, G., Zhou, S., and Guan, J. (2022). Deep structural clustering for single-cell RNA-seq data jointly through autoencoder and graph neural network. *Briefings in Bioinformatics*, 23(2):1–13.

Grün, D., Kester, L., and Van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and Van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255.

Grün, D., Muraro, M. J., Boisset, J.-C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., van den Born, M., Van Es, J., Jansen, E., Clevers, H., et al. (2016). De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell*, 19(2):266–277.

Haghverdi, L., Buettner, F., and Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998.

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods*, 13(10):845–848.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the mouse cell atlas by microwell-seq. *Cell*, 172(5):1091–1107.

Handl, J., Knowles, J., and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer.

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, 64:53–62.

Hennig, C., Meila, M., Murtagh, F., and Rocci, R. (2015). *Clustering strategy and method selection*. Chapman & Hall/CRC.

Herman, J. S., Grün, D., et al. (2018). FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nature Methods*, 15(5):379–386.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Ji, Z. and Ji, H. (2016). TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research*, 44(13):e117–e117.

Jiang, L., Chen, H., Pinello, L., and Yuan, G.-C. (2016). Giniclust: detecting rare cell types from single-cell gene expression data with gini index. *Genome Biology*, 17(1):1–13.

Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, 20(5):273–282.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods*, 14(5):483–486.

Kleiber, C. and Zeileis, A. (2016). Visualizing count data regressions using rootograms. *The American Statistician*, 70(3):296–303.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015a). The technology and biology of single-cell RNA sequencing. *Molecular Cell*, 58(4):610–620.

Kolodziejczyk, A. A., Kim, J. K., Tsang, J. C., Ilicic, T., Henriksson, J., Natarajan, K. N., Tuck, A. C., Gao, X., Bühler, M., Liu, P., et al. (2015b). Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, 17(4):471–485.

Kowalczyk, M. S., Tirosh, I., Heckl, D., Rao, T. N., Dixit, A., Haas, B. J., Schneider, R. K., Wagers, A. J., Ebert, B. L., and Regev, A. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Research*, 25(12):1860–1872.

Larsen, B. and Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 16–22.

Lee, J. (2010). *Introduction to topological manifolds*, volume 202. Springer Science & Business Media.

Lin, P., Troup, M., and Ho, J. W. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biology*, 18(1):1–11.

Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. (2010). Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining*, pages 911–916. IEEE.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. Oakland, CA, USA.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Menon, V. (2018). Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data. *Briefings in Functional Genomics*, 17(4):240–245.

Möbius, A. F. (1827). *Der barycentrische Calcul, ein Hülfsmittel zur analytischen Behandlung der Geometrie (etc.)*. Barth.

Montes, I., Miranda, E., Pelessoni, R., and Vicig, P. (2015). Sklar's theorem in an imprecise setting. *Fuzzy Sets and Systems*, 278:48–66.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1-2):91–118.

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., Elzen, A. v. d., Hirn, M. J., Coifman, R. R., et al. (2019). Visualizing structure and transitions in high-dimensional biological data. *nature biotechnology*, 37(12):1482–1492.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Pouyan, M. B., Jindal, V., Birjandtalab, J., and Nourani, M. (2016). Single and multi-subject clustering of flow cytometry data for cell-type identification and anomaly detection. *BMC Medical Genomics*, 9(2):99–110.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, 14(10):979–982.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Rendón, E., Abundez, I., Arizmendi, A., and Quiroz, E. M. (2011). Internal versus external cluster validation indexes. *International Journal of Computers and Communications*, 5(1):27–34.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Rozenblatt-Rosen, O., Stubbington, M. J., Regev, A., and Teichmann, S. A. (2017). The human cell atlas: from vision to reality. *Nature*, 550(7677):451–453.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., and Waterston, R. H. (2017). Dna sequencing at 40: past, present and future. *Nature*, 550(7676):345–353.

Silva, V. and Tenenbaum, J. (2002). Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems*, 15.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231.

Soneson, C. and Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261.

Stermann, F. (2021). On the Influence of Dependence Structures on Clustering Performance (in German). Accessible at `https://epub.ub.uni-muenchen.de/77439/1/BA_Stermann.pdf`, Last access: 2022-03-31.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:397–423.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Sumithra, V. and Surendran, S. (2015). A review of various linear and non linear dimensionality reduction techniques. *International Journal of Computer Science and Information Technologies*, 6:2354–2360.

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604.

Tenenbaum, J. (1997). Mapping a manifold of perceptual observations. *Advances in Neural Information Processing Systems*, 10.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Tian, T., Zhang, J., Lin, X., Wei, Z., and Hakonarson, H. (2021). Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data. *Nature Communications*, 12(1):1–12.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Research*, 25(10):1491–1498.

Tutz, G., Pößnecker, W., and Uhlmann, L. (2015). Variable selection in general multinomial logit models. *Computational Statistics and Data Analysis*, 82:207–222.

Tutz, G. and Ulbricht, J. (2009). Penalized regression with correlation-based penalty. *Statistics and Computing*, 19(3):239–253.

Ullmann, T., Hennig, C., and Boulesteix, A.-L. (2021). Validation of cluster analysis results on validation data: A systematic framework. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1444.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11).

Vasighizaker, A., Danda, S., and Rueda, L. (2022). Discovering cell types using manifold learning and enhanced visualization of single-cell RNA-Seq data. *Scientific reports*, 12(1):1–16.

Škulj, D. (2018). Imprecise copulas constructed from shock models (Slides of a talk at the 11th Workshop on Principles and Methods of Statistical Inference with Interval Probability (WPMSIIP)). Last access: 2022-03-31.

Wagner, G. P., Kin, K., and Lynch, V. J. (2013). A model based criterion for gene expression calls using RNA-seq data. *Theory in Biosciences*, 132(3):159–164.

Wang, B., Ramazzotti, D., De Sano, L., Zhu, J., Pierson, E., and Batzoglou, S. (2018). Simlr: A tool for large-scale genomic analyses by multi-kernel learning. *Proteomics*, 18(2):1700232.

Wang, D. and Gu, J. (2018). VASC: dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder. *Genomics, proteomics & bioinformatics*, 16(5):320–331.

Wang, J. (2012). *Geometric structure of high-dimensional data and dimensionality reduction*, volume 13. Springer.

Witten, D. M., Shojaie, A., and Zhang, F. (2014). The cluster elastic net for high-dimensional regression with unknown variable grouping. *Technometrics*, 56(1):112–122.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5.

Yu, Z. and Du, F. (2022). AMC: accurate mutation clustering from single-cell DNA sequencing data. *Bioinformatics*, 38(6):1732–1734.

Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome biology*, 18(1):1–15.

Zappia, L. and Theis, F. J. (2021). Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. *Genome biology*, 22(1):1–18.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in R. *Journal of statistical software*, 27(8):1–25.

Zhang, S., Li, X., Lin, Q., and Wong, K.-C. (2020). Review of single-cell RNA-seq data clustering for cell type identification and characterization. *arXiv preprint arXiv:2001.01006*.

Zhang, Z., Cui, F., Lin, C., Zhao, L., Wang, C., and Zou, Q. (2021). Critical downstream analysis steps for single-cell rna sequencing data. *Briefings in Bioinformatics*, 22(5).

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

Zurauskiene, J., Yau, C., et al. (2016). pcareduce: hierarchical clustering of single cell transcriptional profiles. *BMC bioinformatics*, 17(140):1–11.

Attached contributions

Contribution 1: p. 48–59;
accessible at doi: https://doi.org/10.1007/s11634-020-00428-1

Supplementary Material to Contribution 1: p. 60–67;
accessible at `https://static-content.springer.com/esm/art%3A10.1007%2Fs11634-020-00428-1/MediaObjects/11634_2020_428_MOESM1_ESM.pdf`

Contribution 2: p. 68-77

Contribution 3: p. 78–86;
accessible at `http://proceedings.mlr.press/v103/fuetterer19a/fuetterer19a.pdf`

Contribution 4: p. 87–96;
accessible at `https://epub.ub.uni-muenchen.de/91666/1/DPL_TR_2022_03.pdf`

Advances in Data Analysis and Classification (2020) 14:885–896
https://doi.org/10.1007/s11634-020-00428-1

REGULAR ARTICLE

Adapted single-cell consensus clustering (adaSC3)

Cornelia Fuetterer[1] · **Thomas Augustin**[1] · **Christiane Fuchs**[2,3,4]

Received: 3 July 2019 / Revised: 11 August 2020 / Accepted: 8 November 2020 /
Published online: 15 December 2020

Abstract

The analysis of single-cell RNA sequencing data is of great importance in health research. It challenges data scientists, but has enormous potential in the context of personalized medicine. The clustering of single cells aims to detect different subgroups of cell populations within a patient in a data-driven manner. Some comparison studies denote single-cell consensus clustering (SC3), proposed by Kiselev et al. (Nat Methods 14(5):483–486, 2017), as the best method for classifying single-cell RNA sequencing data. SC3 includes Laplacian eigenmaps and a principal component analysis (PCA). Our proposal of unsupervised *adapted single-cell consensus clustering (adaSC3)* suggests to replace the linear PCA by diffusion maps, a non-linear method that takes the transition of single cells into account. We investigate the performance of $adaSC3$ in terms of accuracy on the data sets of the original source of SC3 as well as in a simulation study. A comparison of $adaSC3$ with SC3 as well as with related algorithms based on further alternative dimension reduction techniques shows a quite convincing behavior of $adaSC3$.

Keywords Diffusion maps · Non-linear embedding · Single-cell consensus clustering · Simulation data · Single-cell RNA sequencing data

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11634-020-00428-1.

✉ Cornelia Fuetterer
Cornelia.Fuetterer@stat.uni-muenchen.de

Thomas Augustin
Thomas.Augustin@stat.uni-muenchen.de

Christiane Fuchs
christiane.fuchs@helmholtz-muenchen.de

1 Department of Statistics, Ludwig-Maximilians-University Munich, 80539 Munich, Germany

2 Faculty of Business Administration and Economics, Bielefeld University, 33615 Bielefeld, Germany

3 Institute of Computational Biology, Helmholtz Zentrum Munich, 85764 Neuherberg, Germany

4 Department of Mathematics, Technical University of Munich, 85747 Garching, Germany

Mathematics Subject Classification 62H30 · 68U20

1 Introduction

Personalized medicine based on genomic data promises the precise and individualized treatment of diseases using information from a patient's genome (Cho et al. 2012). There is tremendous research interest in this field, especially with regard to cancer. Hereby it is of interest to determine the different stages of cancer as well as the understanding of the complex development of organs for instance, by analyzing the single cells that are obtained from the single-cell RNA sequencing. Data-driven approaches have led to projects such as The Human Cell Atlas (2020), which aims to establish an interpretable structure for the different cell types of single cells and serves as an orientation for the study of diseases. The mission of the Human Cell Atlas is "(t)o create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease." Based on the genetic profiles of these single-cell RNA sequencing data, an unsupervised classification allows a data-driven distinction of intra- and intertumoral heterogeneities as well as the determination of different pathways during the development (Duò et al. 2018). The approach of single-cell consensus clustering (SC3) by Kiselev et al. (2017) has gained much attention, not only due to its superior performance in the comparison study of Duò et al. (2018). SC3 is also explicitly tailored to single cell data. Nevertheless, the original incorporation of the linear dimension reduction of principal component analysis may offer a potential for improvement. Following Bendall et al. (2014) and Buettner and Theis (2012), for single cells the transition from one state to another is a non-linear continuous process. Therefore, we propose to replace the PCA of the SC3 method by diffusion maps (Haghverdi et al. 2015), resulting in a new unsupervised algorithm, which we call *adapted single-cell consensus clustering (adaSC3)*. The use of diffusion maps is not only motivated by the biological behavior of single cells but is also supported empirically: First, diffusion maps allow for a natural modeling of the transition of single cells by Markov processes. Secondly, according to Haghverdi et al. (2015), when applying diffusion maps to single cell data, they also seem to perform best compared to other non-linear transformation methods such as independent component analysis, Kernel PCA, Isomap, or Hessian Local Linear Embedding.

Our paper ist structured as follows: We first give an introduction into single-cell RNA sequencing data in Sect. 2. In Sect. 3, we present the methodological background, starting in Sect. 3.1 with the proposed *adapted single-cell consensus clustering (adaSC3)*, as well as some related competing methods. In Sect. 3.2, we focus on the special suitability of diffusion maps, included in our framework of $adaSC3$. Analyzing some characteristic single-cell RNA sequencing data sets, introduced already in Sect. 2, $adaSC3$ is compared to its competing methods in Sect. 4. The performance of $adaSC3$ and its competing methods is further evaluated with partition-wise simulation data in Sect. 5. Section 6 concludes with a brief discussion and outlook.

Table 1 Characteristics of the scRNA-seq data sets, with N single cells, G genes and k categories of cell types

Data set	N	G	k
Biase et al. (2014)	49	13,322	3
Deng et al. (2014)	269	10,333	10
Goolam et al. (2016)	124	11,154	5
Treutlein et al. (2014)	80	5757	5
Yan et al. (2013)	90	10,077	7

2 Single-cell RNA sequencing data (ScRNA-Seq data)

The health and development of living organisms can heavily be impacted by the kind and activity of their genes, referred to as gene expression. With the technique of single-cell RNA sequencing (scRNA-seq) introduced by Tang et al. (2009), it is possible to measure the gene expression for single cells. In scRNA-seq, genomic profiles are measured in terms of read counts, that is the number of small sequences ("reads") that result from a cell's RNA that can be identified as belonging to a particular gene. A data set comprising N cells and G genes will hence be a $N \times G$ matrix containing non-negative integers (including zero). The scRNA-seq data of Biase et al. (2014), Deng et al. (2014), Goolam et al. (2016) and Yan et al. (2013), provided by the Hemberg Group of the Sanger Institute (2020), contain read counts of single cells of a mouse or a human with different cell states that are passed during differential transcription for targeting the analysis of cell division in a pedigree. Gene expression is stochastic, and often the reads follow different distributions for different cell types. Another topic of interest is the examination of having reached varying pathway stages. For example, the experiments of Treutlein et al. (2014) were carried out to investigate cell transition during lung development. In detail, these experiments aim to analyze the development of the distal lung epithelium of the mouse based on the different transcriptional states.

The data sets shown in Table 1 had been used for the evaluation of SC3 by Kiselev et al. (2017). Fulfilling the reproducibility and the sample size criterion for the unsupervised classification leads to the data situation[1] described in Table 1. Except for the data set of Biase et al. (2014), the distributions of cell types are quite unbalanced encompassing between 80 and 269 single cells.

[1] Since in the SC3 framework, as described in the transformation step of $adaSC3$ later, components in higher dimensions are chosen randomly, it is important to focus on data sets with a small amount of single cells, in order to keep the analysis replicable. Moreover, for providing the same data situation as Kiselev et al. (2017), we had to adapt the data set of Biase et al. (2014) such that the number of single cells corresponds to the data description of the original SC3 paper.

3 Methods

In this section, we first present the proposed *adapted single-cell consensus clustering (adaSC3)*, which allows, by replacing PCA with diffusion maps, to take the varying pathways and their different transcriptional states into account. Furthermore, we consider several SC3-like approaches that later serve as competitors including different transformation techniques.

In the second part of this section, we look more closely at diffusion maps, which allow an appropriate embedding of the complex data structure of single cells in the transformation step of $adaSC3$.

3.1 Unsupervised adapted single-cell consensus clustering (*adaSC3*) and its competitors

Single-cell consensus clustering aims at classifying the gene expression of single cells in an unsupervised way such that groups are determined in a data-driven manner for detecting new subgroups or confirming manually determined cell types. The classification process subdivides the cell population with regard to the homogeneity of the genetic profile into subgroups of single cells, which represent, for example, different stages of a disease or of a development process within a patient or within a mouse. The original SC3 is implemented in the software *R* (R Core Team 2020) and can be described as a pipeline consisting of several transformation steps including an automatic dimension reduction, resulting in a clustering respecting all combinations. For the construction of the adapted single-cell consensus clustering (*adaSC3*) and its competitors, we rely on the same principle framework as SC3 but consider different transformations.[2] The concrete procedure of $adaSC3$ consists of the following steps:

1. *Preprocessing* As a result of the scRNA-seq process, one obtains the gene expression matrix E containing the read counts of N single cells and G genes. As a preprocessing step, the original matrix E is reduced by a gene filter, as proposed in the original work, that aims to exclude rare and omnipresent genes.[3] This leads to the expression matrix E' of dimension $N \times G'$.
2. *Calculation of distance matrix D* Based on the expression matrix E', the Euclidean distance matrix is constructed for each pairwise single cell combination. Furthermore, two measures of dissimilarity are applied on the log-transformed data of E' using the Pearson and the Spearman correlation, respectively. For the sake of simplicity, the obtained distance and dissimilarity matrices will each be referred to as distance matrix D.
3. *Transformation technique T* For each of the obtained distance matrices D, we apply Laplacian eigenmaps[4], introduced by Belkin and Niyogi (2003) as proposed in the original SC3. In addition, we apply diffusion maps, described in more

[2] AdaSC3 and its competitors are also implemented in R and are available from the first author upon request.

[3] Genes with an expression value of > 2 in less than 6% of the cells as well as genes with a positive expression value in more than 94% of the single cell population are excluded.

[4] That are implemented as a spectral embedding in the Python software library `Scikit-learn` (Pedregosa et al. 2011) as the best size of the neighborhood for the aimed embedding.

detail below, instead of the originally proposed principal component analysis. To each of the six combinations, consisting of three different distance matrices D and two transformation techniques, an eigenvalue decomposition is applied. This leads in total to $3 \times 2 = 6$ different eigenvalue decompositions, resulting each in respectively $N - 1$ eigenvectors $\psi_1, \ldots, \psi_{N-1}$ with their ordered eigenvalues $1 > \lambda_1 \geq \cdots \geq \lambda_{N-1}$.

4. *Consensus clustering* In accordance to SC3, we adopt the automatic selection of the number of eigenvectors to be considered. For each eigenvalue decomposition, a k-means clustering, with k being deterministic, representing the number of categories of the underlying cell types as proposed in the original paper of Kiselev et al. (2017), is conducted. The automatic selection of eigenvectors of the described scenario starts incorporating the first eigenvector until the rounded integer of the 4% quantile of the set $\{1, \ldots, N\}$. The subsequent clusterings include each one further eigenvector until the 7% quantile is reached for including the maximal range of eigenvectors for the last clustering of each combination. The result of each k-means clustering m is summarized in a consensus matrix $\mathcal{C}$, indicating the relative frequency of how often a pair of single cells is grouped together over all n_m clusterings. Based on the obtained consensus matrix $\mathcal{C}$, a final complete-linkage clustering is performed. It aims to achieve higher performance and a more robust result for the classification of single cells, leading to the final grouping of k subgroups. The quality of clustering is evaluated ex-post by the Adjusted Rand Index (ARI) as proposed in the original work of Kiselev et al. (2017).

Apart from the original SC3 that includes a PCA and Laplacian eigenmaps as transformation techniques, we construct further competing algorithms following the same principle as of $adaSC3$. Instead of diffusion maps and Laplacian eigenmaps, we propose additional algorithms leading to two different types of constructions, differing in the number of incorporated transformations. The first construction only uses one single transformation technique T in Step 3 of $adaSC3$. We therefore analyze the influence of the non-linear manifolds of isomaps (IM), locally linear embedding (LLE) as well as the multidimensional scaling (MDS), in addition to the transformation of the principal component analysis (PCA), Laplacian eigenmaps (LE), and diffusion maps (DM), each on their own.[5] The second type of construction consists of the combination of each mentioned transformation T with Laplacian eigenmaps. This leads to the algorithms named by the incorporated transformations, resulting in LE + IM, LE + LLE and LE + MDS, in addition to the original SC3 (PCA + LE) and $adaSC3$ (DM+LE).

3.2 Diffusion maps

In the following, we describe the motivation of embedding the complex structure of single cells during transition into an appropriate global non-linear manifold, using

[5] For the construction of IM and LLE, Kayo (2006) proposes to use for IM and LLE the same estimate for the optimal neighborhood size, implemented by the R function `calc_k` of the R package `lle` (Diedrich et al. 2012). Isomaps are then constructed using the R package `vegan` (Oksanen et al. 2019); the locally linear embedding further relies on the R package `lle` (Diedrich et al. 2012) and MDS is based on the R package `stats` (R Core Team 2020).

diffusion maps. As stated in the introduction of Angerer et al. (2016), diffusion maps allow the reconstruction of the different states that are connected via different transitions. One possible transition is the mutation of one single cell into another. For the following construction of diffusion maps, we only consider the transition of one single cell into another within one step. Another decisive fact is the robustness of diffusion maps to noise. Furthermore, with the normalization, described in the following, diffusion maps are able to detect lowly represented cell types. Coifman and Lafon (2006) provide the general framework of diffusion maps that can be adapted to single cells following (Angerer et al. 2016) that is presented in the next steps. For the construction of diffusion maps, consider two states $x, y \in \Omega$, with Ω as the appropriate state space. x and y represent single cells; their gene expressions, measured by count data, lead to the pairwise distance $D(x, y)$.

1. For each choice of the distance measure D, each point (single cell) is considered as a node of a symmetric graph with weight function K_D

$$K_D(x, y) = \exp\left(-\frac{D(x, y)}{2\alpha^2}\right),$$

indicating the affinity of a pair of single cells with scale parameter α, reflecting the best size of the included neighborhood.[6]
2. In the following, we construct the core of a transition kernel of a Markov chain

$$P_D(x, y) = \frac{K_D(x, y)}{Z(x)}, \quad \text{with } Z(x) = \sum_{y \in \Omega} K_D(x, y)\,.$$

3. With a density interpretation of the upper term, the following density normalized transition probability matrix

$$\tilde{P}_D(x, y) = \frac{1}{\tilde{Z}(x)} \frac{K_D(x, y)}{Z(x)Z(y)}, \quad \text{with } \tilde{Z}(x) = \sum_{y \in \Omega \setminus x} \frac{K_D(x, y)}{Z(x)Z(y)}$$

can be obtained. As the research question consists of mapping the differentiation behavior of single cells, we are only interested in the transition between single cells. Thus, the diagonal of $\tilde{P}_D(x, y)$ is set to zero, and the normalization is adapted appropriately, summing up only the gene expression of differing pairs of single cells with $y \neq x$.
4. Based on the normalized matrix $\tilde{P}_D$, indicating the transition of one state to another by an ergodic Markovian diffusion process, the aimed transformation is obtained.

4 Results

In this section, we evaluate the clustering performance of $adaSC3$ and its competitors. The accuracy of combining each of the mentioned transformations in combination with Laplacian eigenmaps is illustrated in Table 2.

[6] The estimation of α relies on the methods implemented in the R package `destiny` (Angerer et al. 2016).

Table 2 ARI of all algorithms including two transformation techniques with
*: overall best clustering performance of the combination Laplacian eigenmaps (LE) with isomaps (IM), locally linear embedding (LLE), and multidimensional scaling (MDS), as well as SC3 and $adaSC3$; bold: best performance comparing SC3 and $adaSC3$

Data Set	LE+ IM	LE + LLE	LE + MDS	SC3	$adaSC3$
Biase et al. (2014)	0.95	1.00*	1.00*	**0.95**	**0.95**
Deng et al. (2014)	0.68	0.70	0.56	0.67	**0.76***
Goolam et al. (2016)	0.54	0.69*	0.54	**0.69***	0.68
Treutlein et al. (2014)	0.53	0.42	0.56	0.66	**0.77***
Yan et al. (2013)	0.75	0.75	0.65	0.65	**0.75***

$AdaSC3$ leads in three out of five cases (Deng et al., Treutlein et al. and Yan et al.) to better clustering results, compared to the original SC3, and it is identical in the case of the data set of Biase et al.. Concerning the competing algorithms, the combinations of LE with IM and MDS tend to deliver worse results compared to $adaSC3$. However, LE + LLE achieves two times the best performing classification but fails extremely in the case of the Treutlein et al. data set. The slightly worse performance of $adaSC$ compared to SC3 concerning the complete Goolam et al. data set of Table 2 should not be over-interpreted as the resampling results based on leaving out each single cell once, we reach considerably higher performance compared to SC3. Furthermore, we can state that $adaSC3$ delivers the highest overall performance concerning both the resampling study as well as using only one transformation technique as illustrated in the Supplementary Material. We therefore consider our proposal as generally the best approach among its competitors, of SC3 and its related approaches, based on the benchmarking data sets. This result is especially surprising as the scRNA-seq data sets were originally used for determining the proposed default settings of SC3, such as e.g. the automatic choice for the lower dimension.

5 Simulation data

The classification accuracy of the simulation data is investigated in the same way as the scRNA-seq data. We are interested in the consensus clustering accuracy of two simulation groups, which are constructed with different ranges of distribution parameters describing the read counts. With shifted parameter ranges, one can consider the simulation groups as representing a healthy and a diseased population. According to the literature, the use of a zero-inflated negative binomial (ZINB) distribution is recommended as the most adequate approximate distribution for modeling the read counts of single cells. It allows larger variability of read counts compared to the former used Poisson distribution (see e.g. Wagner et al. 2013). Based on the constructed simulation data following a (generalized version of a) ZINB distribution for the expression of each gene, we will investigate the influence of various parameters describing each gene for all possible group partitions for a fixed total number N of single cells.

5.1 Construction of simulation data

The ZINB distribution (Kleiber and Zeileis 2016) is a mixture between a negative binomial probability mass function and a point mass at zero. For generating ZINB distributed simulation data we use the R package `emdbook` (Bolker B, Bolker Maintainer Ben and Imports, MASS 2020), which is based on a generalization of the negative binomial (NB) distribution with parameters μ and ϕ for the non-zero inflated part.[7] The parameter μ is a continuous positive real value, describing the mean. The dispersion parameter ϕ represents the shape parameter of the gamma distribution underlying the generalization of the NB. The fraction of zero-inflation is taken into account by the parameter π.

In the following scenarios, we investigate the influence of different parameters of the distribution family. In order to mimic a realistic situation, the scRNA-seq data of Kolodziejczyk et al. (2015) is taken for estimating the parameters of a ZINB distribution[8] and allow the construction of ranges for each parameter looking at the shifted quantiles of the estimates of the parameters μ and ϕ. This leads to the parameter ranges $\boldsymbol{\mathcal{M}}^{(1)}$ and $\boldsymbol{\Phi}^{(1)}$ for cell population 1 and $\boldsymbol{\mathcal{M}}^{(2)}$ and $\boldsymbol{\Phi}^{(2)}$ for cell population 2. The parameter range $\boldsymbol{\Pi}$ for π is set to be the same for both populations.[9] Thus, the simulated read counts of each gene follow a ZINB(μ_1, ϕ_1, π_1) distribution for simulation group 1 and a ZINB(μ_2, ϕ_2, π_2) for simulation group 2, according to the following scenarios:

- Simulation scenario (a) for different ranges of μ:
  $\mu_1 \in \boldsymbol{\mathcal{M}}^{(1)}$ and $\mu_2 \in \boldsymbol{\mathcal{M}}^{(2)}$, $\phi_1 = \phi_2 \in \boldsymbol{\Phi}^{(2)}$, $\pi_1 = \pi_2 \in \boldsymbol{\Pi}$
- Simulation scenario (b) for different ranges of ϕ:
  $\mu_1 = \mu_2 \in \boldsymbol{\mathcal{M}}^{(2)}$, $\phi_1 \in \boldsymbol{\Phi}^{(1)}$ and $\phi_2 \in \boldsymbol{\Phi}^{(2)}$, $\pi_1 = \pi_2 \in \boldsymbol{\Pi}$
- Simulation scenario (c) for different ranges of μ and ϕ:
  $\mu_1 \in \boldsymbol{\mathcal{M}}^{(1)}$ and $\mu_2 \in \boldsymbol{\mathcal{M}}^{(2)}$, $\phi_1 \in \boldsymbol{\Phi}^{(1)}$ and $\phi_2 \in \boldsymbol{\Phi}^{(2)}$, $\pi_1 = \pi_2 \in \boldsymbol{\Pi}$
- Simulation scenario (d) for the same range of μ and ϕ:
  $\mu_1 = \mu_2 \in \boldsymbol{\mathcal{M}}^{(2)}$, $\phi_1 = \phi_2 \in \boldsymbol{\Phi}^{(2)}$, $\pi_1 = \pi_2 \in \boldsymbol{\Pi}$

For each of the simulation scenarios (a) to (d), we sample N_1 times out of ZINB$(\mu_1$ $\phi_1, \pi_1)$ and N_2 times out of ZINB(μ_2, ϕ_2, π_2) such that, for comparison purposes, the gene-specific parameters remain the same when generating all possible partitions of $N_1 : N_2$ (with $N_1 + N_2 = N$), starting with $1 : (N-1)$ until $(N-1) : 1$, with $N = 50$ for the respective scenario. In order to obtain simulation data with the dimension $N \times G$ for each partition, we repeat this procedure 200 times. Thus, read counts of $G = 200$ genes are generated with the new parameter values drawn uniformly from the respective intervals.

[7] Details explaining the generalization of the negative binomial distribution function based on a mixture of Poisson distributions with gamma distributed Poisson rates can be found e.g. in Fuetterer et al. (2019). They investigate the influence of different heterogeneity degrees of count data using simulation data as well as up- and downwardly distorted measurements via the ZINB distribution describing the case of measurements tending to lower read counts and upper read counts.

[8] The manual construction of two cell populations rely on the differentially cultured murine embryonic stem cell populations "2i" and "serum" for each of the 38.616 genes.

[9] The constructed parameter ranges are part of the Supplementary Material.

5.2 Clustering performance based on simulation data

The following plots show how the group partition (x-axis) of a combination of two transformation methods influences the clustering accuracy, measured by the Adjusted Rand Index (ARI) (y-axis). In the ideal case, the accuracy is 1 for each of the partitions, which would indicate that the classification perfectly corresponds to the underlying group allocation. This criterion is best met for $adaSC3$, not only in the case of using only one transformation technique (see Supplementary Material), but also in combination of those with Laplacian eigenmaps (LE) for simulation scenarios (a) to (c). Simulation (d) serves as a reference where no difference in the gene-specific parameters was simulated and no accurate grouping should be detected. Each partition of each scenario is repeated 10 times and the accuracy of the respective clustering results is visualized by boxplots. Results of simulation scenario (c) and (d) can be found in the Supplementary Material.

5.2.1 Simulation scenario (a): variation in expectation parameter μ

In the case of differing parameter μ represented in scenario (a), $adaSC3$ seems to perform best among the combined methods (see Fig. 1) as well as compared to each method on its own. It can also be seen that the inter quantile range of boxplots have the tendency to be shorter for $adaSC3$ and reach higher ARI values compared to its competitors. Therefore, we conclude in general that our approach generates

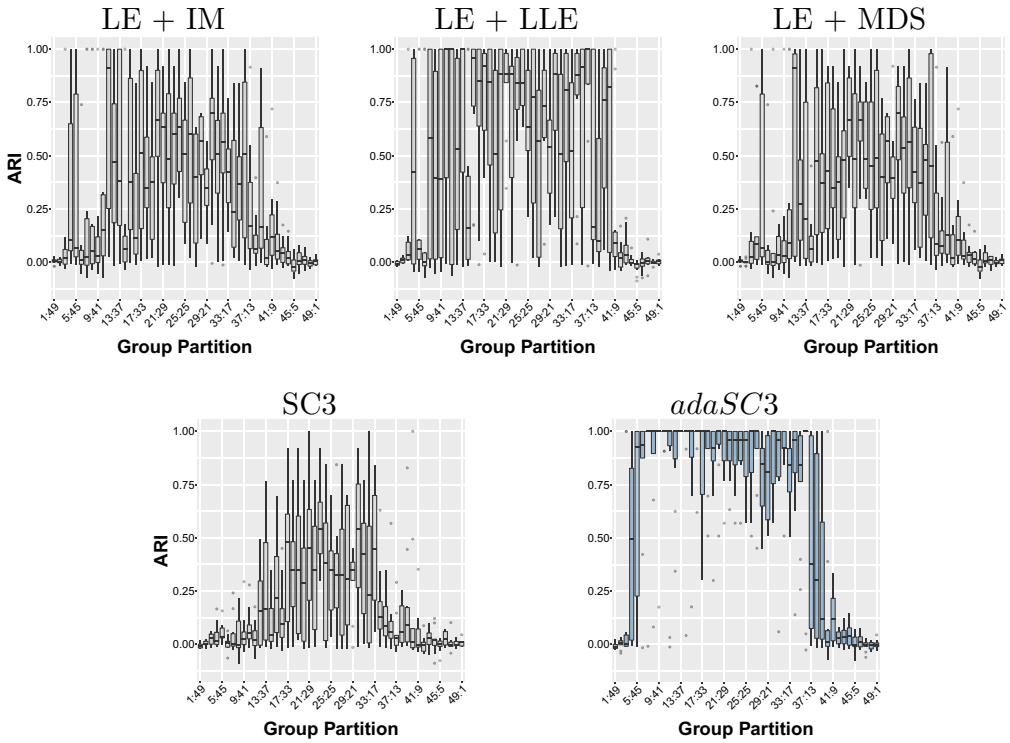

Fig. 1 ARI among all partitions of $N = 50$ with regard to the combination of different transformation techniques for simulation scenario (a)

a more efficient clustering allocation. Furthermore, it should also be noted that the performance of the method $adaSC3$ seems to be better when more single cells are part of simulation group 1 compared to simulation group 2. This might allow the interpretation that the detection of true positives achieves higher accuracy compared to the detection of false negatives, given the diseased population has on average higher gene expression.

5.2.2 Simulation scenario (b): variation in size parameter ϕ

For partition-wise created simulation data differing in the parameter ϕ one can state, referring to Fig. 2, that apart from $adaSC3$ the combination of LE + LLE performs quite well, too. However, this method needs more partitions before it starts detecting the difference in the simulation groups and fails drastically earlier compared to $adaSC3$. For approximately balanced data, LE + LLE often leads to worse results. The tendency that the clustering performance depends on the partitions can be confirmed over all methods for simulation scenarios (a) to (c), in which $adaSC3$ is affected the less.

With regard to scenario (c), the simulated differences of both parameters μ and ϕ lead to a quite accurate classification for most methods with an overall superiority of $adaSC3$, representing the scenario being the closest to the reality. For the simulation design with no difference in the simulation groups, the allocation of single cells is as expected and represents random allocation.

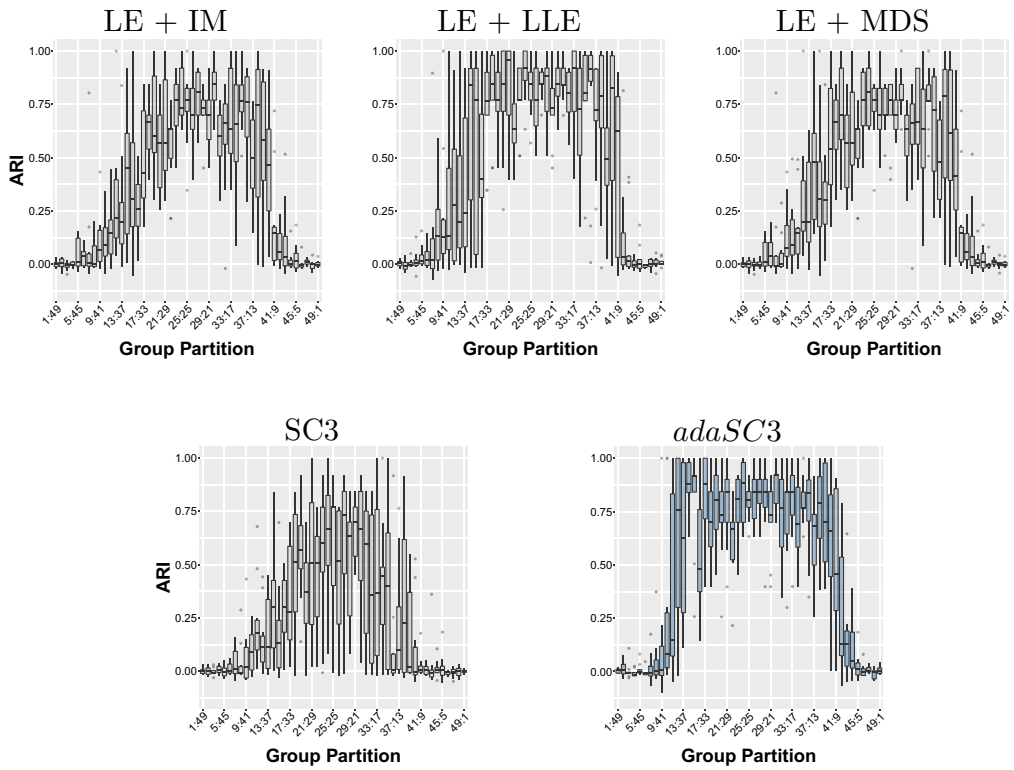

Fig. 2 ARI among all partitions of $N = 50$ with regard to the combination of different transformation techniques for simulation scenario (b)

6 Conclusions

The approach of *adapted single-cell consensus clustering (adaSC3)* is tailored to the clustering of single cells. It reflects the biological structure of single cells by including diffusion maps with the aim to respect the transition process of the underlying data. Indeed, the inclusion of diffusion maps instead of the originally proposed PCA led to a better clustering performance. We consider $adaSC3$ to be the best method compared to all investigated competitors, both in the analyzed scRNA-seq data as well as in the simulation study. This motivates further research that takes into account the biological basis of the data before constructing or combining some methods, as this could be rewarded both in terms of interpretation and accuracy, as shown in this paper.

Based on the discovery that balanced data seems to be detected correctly with higher quality, the distribution of classified classes could be taken into account for an unsupervised evaluation. Furthermore, studies of additional scRNA-seq data and further simulations are needed to reinforce the results of this paper. This is especially due to the fact that $adaSC3$ was evaluated on the same scRNA-seq data used for the development of the original SC3 method. This makes the overall superiority of $adaSC3$ over SC3 even more surprising, while on the other hand some bias of these data sets favoring SC3-like methods cannot be excluded.

Acknowledgements We are very grateful to the two anonymous referees and the editors for their stimulating and constructive comments. We also want to thank Florian Pfisterer for discussions about diffusion maps. Furthermore, the public data sharing of the scRNA-sequencing data by the Hemberg Group of the Sanger Institute is gratefully acknowledged. In addition, the first author is very thankful to the LMUMentoring program, connecting young researchers with experienced researchers and providing financial support.

Funding Open Access funding enabled and organized by Projekt DEAL.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Angerer P, Haghverdi L, Büttner M, Theis FJ, Marr C, Buettner F (2016) Destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics 32(8):1241–1243

Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15(6):1373–1396

Bendall SC, Davis KL, el Amir AD, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP, Pe'er D (2014) Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell 157(3):714–725

Biase FH, Cao X, Zhong S (2014) Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. Genome Res 24(11):1787–1796

Bolker B, Bolker Maintainer Ben and Imports, MASS (2020) Package 'emdbook' R package version 1.3.11

Buettner F, Theis FJ (2012) A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. Bioinformatics 28(18):i626–i632

Cho SH, Jongsu J, Seung IK (2012) Personalized medicine in breast cancer: a systematic review. J Breast Cancer 15(3):265–272

Coifman RR, Lafon S (2006) Diffusion maps. Appl Comput Harmonic Anal 21(1):5–30

Deng Q, Ramsköld D, Reinius B, Sandberg R (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343(6167):193–196

Diedrich H, Abel M., Diedrich Maintainer Holger (2012) Package "lle" R package version 1.1

Duò A, Robinson MD, Soneson C (2018) A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F10000Research 7.2

Fuetterer C, Schollmeyer G, Augustin T (2019) Constructing simulation data with dependency structure for unreliable single-cell RNA-sequencing data using copulas. ISIPTA '19. Proc Mach Learn Res 103:216–224

Goolam M, Scialdone A, Graham SJ, Macaulay IC, Jedrusik A, Hupalowska A et al (2016) Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. Cell 165(1):61–74

Haghverdi L, Buettner F, Theis FJ (2015) Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics 31(18):2989–2998

Hemberg Group at Sanger Institute (2020) scRNA-Seq Datasets. https://hemberg-lab.github.io/scRNA.seq.datasets/. Accessed 11 Aug 2020

Kayo O (2006) Locally linear embedding algorithm–Extensions and applications. Technical Report, Faculty of Technology, Department of Electrical and Information Engineering, University of Oulo

Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Rei W, Barahona M, Green AR et al (2017) SC3: consensus clustering of single-cell RNA-Seq data. Nat Methods 14(5):483–486

Kleiber C, Zeileis A (2016) Visualizing count data regressions using rootograms. Am Stat 70(3):296–303

Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, Bühler M, Pentao L, Marioni JC, Teichmann SA (2015) Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell 17(4):471–485

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H (2019) vegan: Community Ecology Package. Package 'vegan' R package version 2.5-6

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg R, Vanderplas J, Passos A, Cournapeau D, Perrot Brucher M, Duchesnay ME (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/. Accessed 11 Aug 2020

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA (2009) MRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6:377–382

The Human Cell Atlas. https://www.humancellatlas.org. Accessed 11 Aug 2020

Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509(7500):371–375

Wagner GP, Kin K, Lynch VJ (2013) A model based criterion for gene expression calls using RNA-seq data. Theory Biosci 132(3):159–164

Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J et al (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol 20(9):1131

1

Clustering Performance of Complete Single-cell RNA-seq Data

As announced in the paper, Table 1 contains the classification results for the individual transformations of the complete scRNA-seq data in comparison to SC3 and *adaSC3*:

Table 1: ARI of consensus clustering using a specific combination of methods PCA: Principal Component Analysis, LE: Laplacian Eigenmaps, DM: Diffusion Maps; IM: Isomaps and MDS: Multidimensional Scaling;
*: best clustering performance among all transformations;
bold: better performance comparing SC3 and *adaSC3*.

Data Set	PCA + LE (SC3)	LE + DM (adaSC3)	PCA	LE	DM	IM	LLE	MDS
Biase et al. [1]	**0.95**	**0.95**	0.95	0.95	0.95	0.95	1.00*	1.00*
Deng et al. [2]	0.67	**0.76**	0.67	0.95*	0.34	0.44	0.43	0.50
Goolam et al. [3]	0.69	0.68	0.69	0.68	0.94*	0.54	0.69	0.54
Treutlein et al. [5]	0.66	**0.77***	0.58	0.63	0.52	0.34	0.36	0.32
Yan et al. [6]	0.65	**0.75***	0.65	0.75*	0.64	0.60	0.91	0.65

Resampling on Single-cell RNA-seq Data

For classification we used the scRNA-seq data, described in Section 2 of the manuscript, and consider a resampling scheme to analyze how stable the results of the individual approaches are. For this purpose, each single cell was omitted once from each scRNA-seq data set, such that the classification was performed N times on $N-1$ data points. The sample that is drawn without replacement was compared to the corresponding underlying cell types of the sampled $N-1$ single cells. By these resampling experiments, we evaluated each iteration and underline the good performance of *adaSC3* with the scRNA-seq data of Biase et al. [1], Deng et al. [2], Goolam et al. [3] and Treutlein et al. [5] in Figure 1 and of Yan et al. [6] in Figure 2.

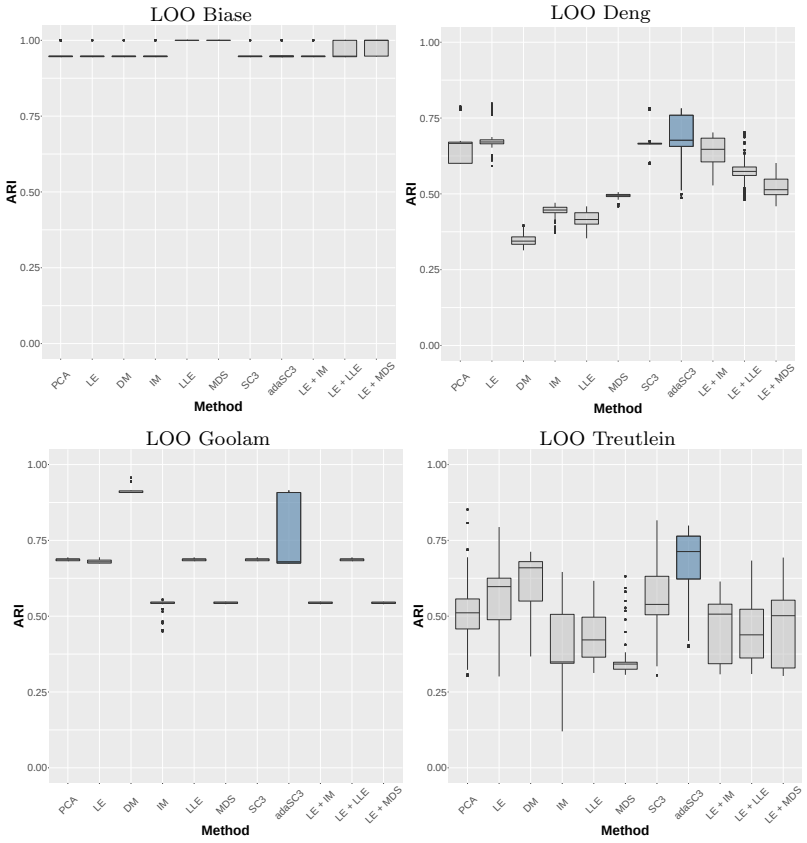

Fig. 1: Resampling results of the scRNA-seq data sets for *adaSC3* and its competitors

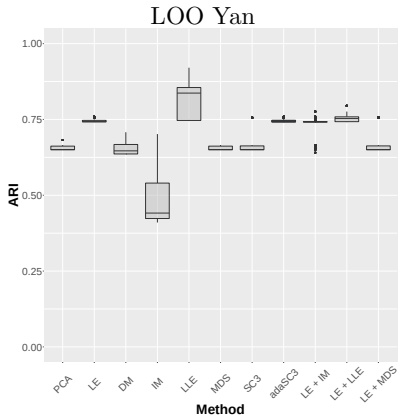

Fig. 2: Resampling results of the scRNA-seq data set Yan et al. [6] for $adaSC3$ and its competitors.

Construction of Simulation Data

To determine realistic simulation data, we mimic the data set of Kolodziejczyk et al. [4] that contains embryo stem cells of a mouse. This data set was selected for the construction of the simulation data because it contains a high number of single cells ($N = 704$) and a high number of genes ($G = 38.616$). Another advantage of this data is that it contains two quite balanced subgroups with 295 and 250 single cells representing the state "2i" and "serum", which are used for the construction of the simulation groups. The remaining 159 single cells of "a2i" are not considered for the construction of the simulation data.

We estimated the parameters of the zero-inflated Negative Binomial (ZINB) distribution of each gene for the respective cell population. From these estimates we determined ranges for the parameters of our simulation data with the 35% to 80% quantiles of the estimated parameters of the labelled group $2i$ for simulation group 1 and the 15% to 60% quantiles of the estimated parameters of the labeled group *serum* for simulation group 2. The lower quantiles of simulation group 2 should represent the healthy population with lower mean and lower dispersion of gene expression compared to simulation group 1. This leads to the following ranges of parameters μ and ϕ of the ZINB distribution for simulation group 1 and 2 shown in Table 2:

Table 2: Intervals of the estimated parameters of a ZINB distribution based on the cell types "2i" and "serum" used by steps of $^{(I)}$: 0.1; $^{(II)}$: 0.001; $^{(III)}$:0.0001.

Parameter Range	Constructed Ranges based on 2i	Constructed Ranges based on serum
$\boldsymbol{\mathcal{M}^{(1)}}$	$\boldsymbol{\mathcal{M}}^{(1)} := [45, 293]^{(I)}$	$\boldsymbol{\mathcal{M}}^{(2)} := [12, 112]^{(I)}$
$\boldsymbol{\Phi}$	$\boldsymbol{\Phi}^{(1)} := [0.24, 0.94]^{(II)}$	$\boldsymbol{\Phi}^{(2)} := [0.12, 0.47]^{(II)}$
$\boldsymbol{\Pi}$	$\boldsymbol{\Pi} := [0.001, 0.01]^{(III)}$	$\boldsymbol{\Pi} := [0.001, 0.01]^{(III)}$

Results of Simulation data

Simulation Scenario (a) - Variation in Expectation Parameter μ

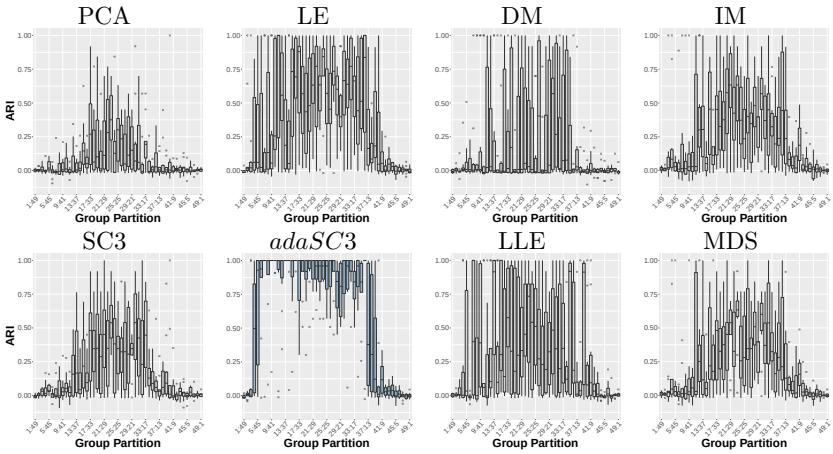

Fig. 3: Clustering accuracy (ARI) among all partitions of N= 50 with regard to the individual transformation techniques for simulation scenario (a) in comparison to SC and $adaSC3$.

Simulation Scenario (b) - Variation in Size Parameter ϕ

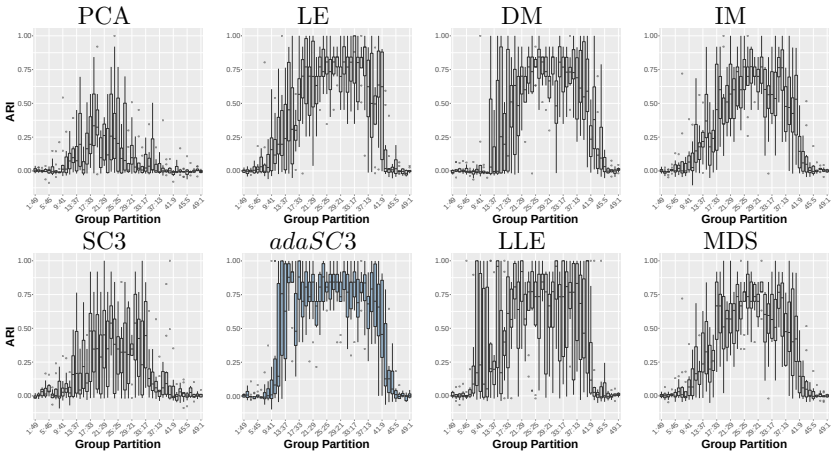

Fig. 4: Clustering accuracy (ARI) among all partitions of N= 50 with regard to the individual transformation techniques for simulation scenario (b) in comparison to SC and $adaSC3$.

Simulation Scenario (c) - Variation in both Parameter μ and Parameter ϕ

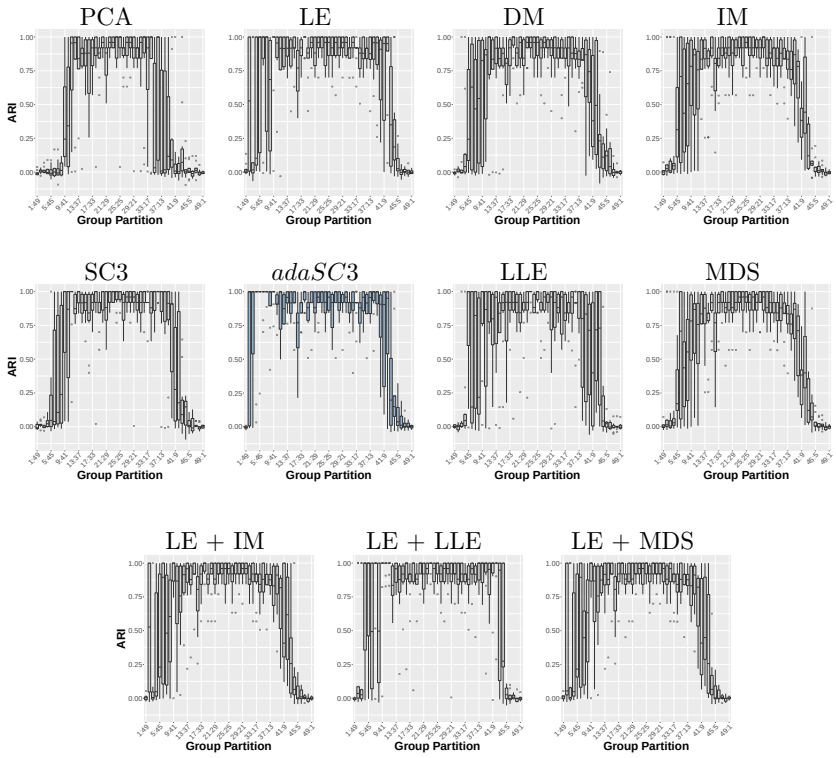

Fig. 5: Clustering accuracy (ARI) among all partitions of N= 50 with regard to the individual transformation techniques as well as in combination with LE for simulation scenario (c).

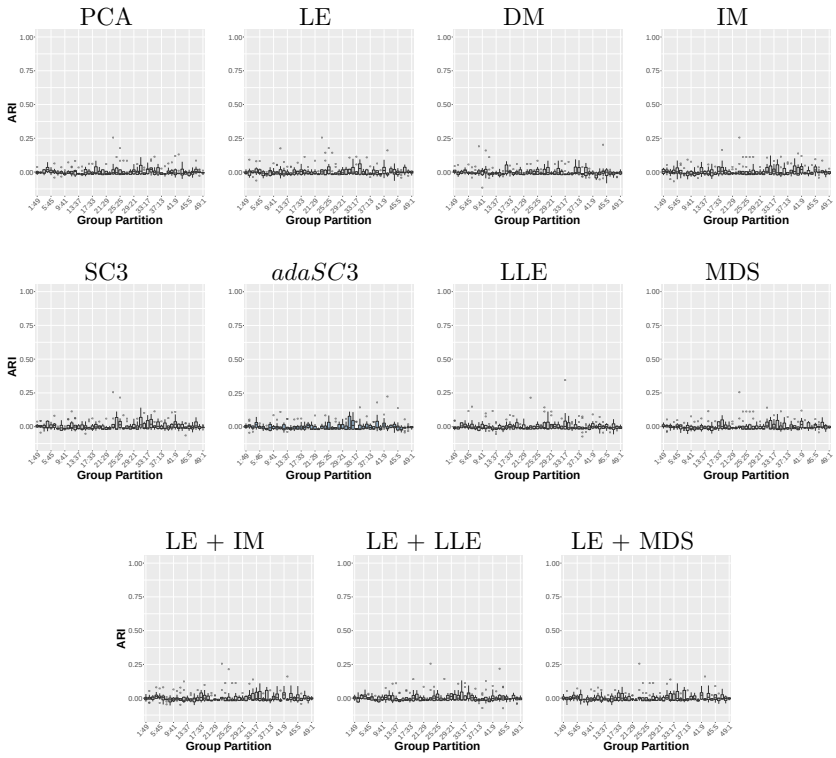

Fig. 6: Clustering accuracy (ARI) among all partitions of N= 50 with regard to the individual transformation techniques as well as in combination with LE for simulation scenario (d).

References

1. Biase FH, Cao X, Zhong S (2014) Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. Genome Research 24.11:1787-1796.
2. Deng Q, Ramsköld D, Reinius B, Sandberg R (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343.6167:193-196.
3. Goolam M, Scialdone A, Graham SJ, Macaulay IC, Jedrusik A, Hupalowska A et al. (2016) Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. Cell 165.1:61-74.
4. Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, Tuck AC, Gao X, Bühler M, Pentao L, Marioni JC, Teichmann SA (2015) Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell 17.4:471-485.
5. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509.7500:371-375.
6. Yan L, Yang M, Guo H, Yang L, Wu J, Li R , Liu P, Lian Y, Zheng X, Yan, J. et al. (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nature structural and molecular biology 20.9:1131.

Internal Validation of Unsupervised Clustering following an Association Accuracy Heuristic

Cornelia Fuetterer
Department of Statistics
Ludwig-Maximilians-University Munich
80539 Munich, Germany
Cornelia.Fuetterer@stat.uni-muenchen.de

Thomas Augustin
Department of Statistics
Ludwig-Maximilians-University Munich
80539 Munich, Germany
Thomas.Augustin@stat.uni-muenchen.de

***Abstract*—One challenge of unsupervised clustering is that its clustering results cannot be directly evaluated in terms of expected accuracy. In case of internal validation, the clustering is validated based on the compactness within a cluster as well as the separation of clusters. Especially in high dimensional settings, internal validation as well as user inspection, becomes more difficult and expensive the higher the dimension of the data. We therefore propose an *association accuracy heuristic*, relating the association of results obtained by different methods to their accuracy. This heuristic is based on an analogy to decision making where high homogeneity among the opinions of independent experts is a widely accepted indicator for having chosen the right decision. Analogous to expert opinions, we assess the groupings of different state-of-the art clustering methods. To measure the (dis)similarity of the clustering results, we propose *method-association-measures*, that are built on an adaption of χ^2-based association measures. Our heuristic is investigated in a simulation study as well as on single-cell RNA sequencing data. Incorporating the ground truth allows a validation of the proposed association accuracy heuristic. Our results provide the opportunity to distinguish between situations where clustering results are expected to be trustworthy and settings where external intervention is indispensable to protect ourself against high risk of bad clustering results.**

***Index Terms*—Unsupervised Clustering, Internal validation, Association measure, Non-linear embedding, Single-cell RNA sequencing data**

I. Introduction

With single-cell RNA sequencing (scRNA-seq), enabled by [1], progress has been made by allowing sequencing on the level of single cells. The measured gene expression of one gene is obtained by counting the number of cDNA-fragments in the sequencing library ("reads") that can be assigned to the underlying gene sequence. The result of the measurement process over all genes and all single cells is a count matrix, containing the reads of N single cells and p genes ($N << p$).

As stated by [2], the groups of single cells, called cell types, are usually unknown and are mostly obtained by an explorative analysis applying different context-adapted clustering algorithms. For a manual grouping, which is the case of the used scRNA-seq data, much effort is needed, which becomes even infeasible in case of a high number of markers (genes that are used to identify specific cell types). In addition, manual inspection is influenced by the researcher's experience and interpretation of visual embeddings leading to human mistakes [3]. These are the reasons why automatic clustering methods are more and more applied in order to achieve accurate groupings. In case of single cells, methods such as Seurat [4] and SC3 [5] are often used for determining cell types [6].

Far beyond genetics, the evaluation of unsupervised clustering is a well-known issue for different application fields such as for example marketing, biology, insurance, earthquake studies as well as text clustering targeting natural language processing tasks. While computation time becomes cheaper and cheaper, expert decisions remain expensive. Accordingly, the development and application of automated algorithms is increasing, which leads to fewer decisions that have to be made by humans. The cluster validation can be based on internal or external validation, on stability measures, hypothesis testing or visual validation as shown by [7] and [8]. Especially on the field of internal validation, no involvement of domain experts is planned. Therefore the automatic validation of unsupervised clustering becomes increasingly important. Since no external information about the ground truth is included in the internal validation at all, the quality assessment can only be performed using heuristics. Remaining strictly in an unsupervised setting, such heuristics can guide an assessment of the uncertainty of making wrong decisions and preferably create measures that can recommend clustering approaches that indicate low uncertainty.

In general decision making, one relies on the consensus of experts. In case of an agreement of expert opinions, one trusts them. Otherwise, caution is required, and, accordingly, deeper studies are then needed to determine which opinion is best. If we now transfer the above setting to unsupervised clustering, we can focus on the extent of the agreement of different well-established clustering methods[1]. This assumption allows interpreting the agreement of the resulting clusterings as the consensus of experts' opinions, applying our heuristic for internal validation. Based on this decision making analogy,

The first author is very grateful to the LMUMentoring program, supporting young researchers. 978-1-6654-0126-5/21/$31.00 ©2021 European Union

[1]We stress the well establishments of the methods to be considered, i.e. their expertise. Consensus as agreement itself is not sufficient; it is crucial that the consensus occurs within a well informed group of experts. So it goes without saying that stupid clustering methods that come to the same result can be constructed without showing reasonable accuracy.

we consider the following *association accuracy heuristic:*

> *The degree of association between clustering results derived by different methods is an indicator for the extent of trustworthiness of the results. Under high association, a high accuracy of each method can be expected, while lowly associated clustering results indicate a high risk of choosing a method with bad performance.*

For the measurement of association of the methods we construct *method-association-measures*, that are inspired by adapted χ^2 association measures, describing the (dis)similarity of methods either observation-wise or pair-wise observations. By highly(lowly) associated methods, we understand a high(low) average association value with low(or high) variation of association within one method-association-measure. The proposed association measures can also serve for internal validation, allowing to quantify and rank the association of the methods' results among context-related data sets. Our heuristic is evaluated in high-dimensional settings in order to assess whether the proposed method-association-measures are an indicator for the expected performance. We first examine the association accuracy heuristic under systematically varied dependence of the data structure in our simulation study. Then, we investigate the heuristic in the context of single-cell RNA sequencing data, in comparison to the internal validation measure of the silhouette index [9]. In both settings, we investigate the association of context-related clustering algorithms that include different non-linear embeddings: *SC3* [5], *adaSC3* [10], *phate* [11], and the combination of *umaps* [12] with *k-means*.

The paper is structured as follows. In Section II, we review already existing validation measures of clustering algorithms. In Section III, we provide method-association-measures, allowing the assessment of observation-wise as well as pair-wise (dis)cordant clustering results of methods, described in Section III-B and Section III-C. We provide their interpretation in Section III-D. Including ground truth for evaluating the accuracy of the methods after the investigation of their (dis)similarities allows an evaluation of our heuristic for both the simulation study (see Section IV), as well as for some real scRNA-sequencing data (see Section V). Section VI concludes and gives a short outlook.

II. Related background

Clustering is part of unsupervised learning [13], which is also known as a data mining task [14], allocating the most similar objects to the same cluster. The validation of clustering can be based on external, relative or internal validation measures. For the external validation, external information such as the gold standard or the underlying ground truth is required for assessing the clustering quality [15]. External validation then measures the "purity" of the obtained clustering and the underlying class labels [16]. Relative clustering validation compares the clustering results of the same algorithm, achieved by different parameter values, based on the same data set. Internal evaluation uses distance metrics and variances to assess the inter-cluster separation and the intra-cluster cohesion (compactness), without the inclusion of external information [15]. One internal validation measure is the classical silhouette index [9], which is constructed on the individual level of compactness and separation. This index takes the averaged compactness within the individual's cluster as well as the averaged separation to the closest cluster into account, and is then averaged over all observations. An absolute value of the silhouette index of 1 indicates that for each observation the corresponding cluster is compact itself and well separated to all other clusters. The more the absolute value approaches 0, the less homogeneous are the observations within the same cluster, and the less heterogeneous are the observations to the observations of the closest cluster.

In general, all internal evaluation measures have the aim to recommend the best clustering algorithm on a specific data set, see also [16] for general reviews of classical measures. As external information, such as the true class labels or ground truth, is often not available, internal validation is the only possibility for assessing a clustering algorithm. This in particular applies to high-dimensional settings, where it is very likely that the knowledge of underlying ground truth is neither complete nor correct [14]. Especially in genetics, we have high dimensional settings where the number of variables p is substantially higher than the number of observations N ($p >> N$) (see e.g. [17] and [18]), which causes additional challenges for internal evaluation. For those settings, subspace clustering or dimension reduction approaches with a subsequent clustering is often performed. Under some assumptions, with increasing dimensions, the difference between the highest distance and the smallest distance gets very small, in relation to the smallest distance, which is a result of the curse of dimensionality [19]. It is therefore not recommendable to base internal validation of high-dimensional data on distances, as they are not confidently interpretable.

III. How to measure and to interprete the association of methods

In the following, we present the construction of our method-association-measures that aim to measure the (dis)similarity of different methods. Throughout the paper an appropriate group alignment has been conducted in advance. This can be achieved by the Hungarian Algorithm of [20] (Bioconductor package `cola`, [21]). The relabeling mainly serves for a clear description and allows access to a more intuitive understanding of the aim of our heuristic. The χ^2-based construction of the association measures derived in this section are invariant to swapping labels. In Section III-A a motivation for constructing appropriate method-association-measures is provided. After the visual motivation, we propose the according association measures in Section III-B and SectionIII-C, followed by their interpretation, given in Section III-D.

A. Visual Motivation

Within this paper we consider different state-of-the art clustering methods in the context of single-cell data. In order to get a better idea of our association accuracy heuristic, we will visualize the results of the clustering methods *SC3*, *adaSC3*, *phate*, and *umaps* in combination with *k-means*. The relabeling was done by taking the clustering method *SC3* as reference. Aiming at a direct comparison of methods in the following radar chart, we set the number of clusters K equal to the number of underlying number of groups G of the below introduced data sets of Biase et al. [22] and Treutlein et al. [23]. In Fig. 1 we can see the relative frequencies of the underlying cell types (truth), as well as the relative frequencies of each cluster within one method for each of the displayed data sets. The relative frequencies are indicated by the polar coordinates for each clustering method. All relative frequencies within one plot sum up to 1, as each single cell can only be grouped to one single cluster. If all clustering methods assign the same frequencies to the different groupings, and agree with the relative proportions of the underlying ground truth, the radii of the radar chart are the same for each method, resulting in a perfect circle. Such almost perfect circle is obtained in case of the Biase et al. data set (left of Fig. 1). Here, all clustering methods result in the same grouping frequencies but all of these methods slightly differ to the the underlying truth. We stress out that the more similar the clustering results of the different algorithms, the rounder the K circles of the radar charts, visualizing the strength of association of the clustering results. The radar chart on the right of Fig. 1 represents the relative frequencies of group assignments of the Treutlein et al. data set [23] and indicates more dissimilar clusterings. Referring to the expert analogy above, identifying the clustering methods with experts, we investigate the associations among the clustering results of the different methods.

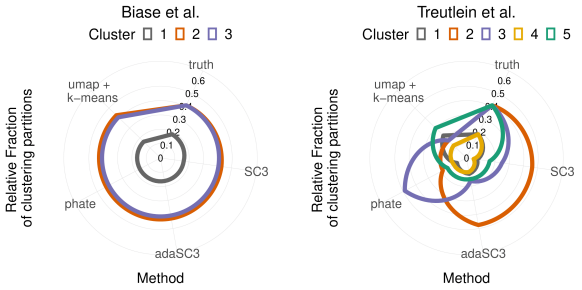

Fig. 1: Radar charts indicating the relative fraction of observations that were assigned to the specific clusters based on the considered clustering methods that are arranged in clockwise order for the clustering results of the data set of Biase et al. [22] (left) and Treutlein et al. [23] (right).

For that purpose, we construct association measures meeting the following requirements:

- In a setting with completely homogeneous assignments, each clustering method proposes the same grouping. The respective association should be described as perfect association.
- With descending concordance, reflecting a setting with more heterogeneous assignments, the association measure should indicate less association.

To achieve this aim in the construction of association measures, we consider the clustering results as attributes of the specific clustering method. This allows an adaption of classical χ^2-based association measures, in our case providing an assessment of association strength of methods. For the internal evaluation of our heuristic, we suggest the below presented method-association-measures, which include the adaption of classical association measures. On the one hand, the *contingency coefficient C* and *Cramér's V*, presented in Section III-B respect the obtained clusterings for each observation on its own. On the other hand, the introduced method-association-measure of the Φ-coefficient is constructed on pair-wise observations and introduced in Section III-C. In Section III-D we provide the interpretation of our method-association-measures.

B. Adjusted Contingency Coefficient C and Cramér's V

We consider the set $\mathcal{I} = \{1, \ldots, N\}$ of N observations, where each observation is part of one of the K adapted clusters for each of the considered clustering methods $\mathcal{M} = \{M_1, \ldots, M_q\}$. As each method $M \in \mathcal{M}$ is seen as a mapping from the set of units $\mathcal{I}$ to the set $\mathcal{C} = \{1, \ldots, K\}$ of K cluster labels, each clustering method M is reinterpreted as a variable with K attributes. In the contingency table (Table I) we have

TABLE I: Contingency table showing the observed frequencies of cluster $(c_{M_k}, c_{M_l}) \in \{1, \ldots, K\}^2$ for method M_k and M_l.

M_k \ M_l	1	$\cdots$	K	
1	n_{11}	$\cdots$	n_{1K}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
K	n_{K1}	$\cdots$	n_{KK}	$n_{K\cdot}$
	$n_{\cdot 1}$	$\cdots$	$n_{\cdot K}$	N

the same number of rows and columns as we only consider methods with the same number of clusters K. For every method M_k and M_l, we specify $n_{c_{M_k} c_{M_l}}$ as the absolute frequency of observations that are part of group label $c \in \{1, ..., K\}$ with the respective methods M_k and M_l ($\forall\ k \neq l$, $k, l \in \{1, ..., q\}$). To apply χ^2 (see e.g. [24, Chapter 4]), contrasting $n_{c_{M_k} c_{M_l}}$ with $\tilde{n}_{c_{M_k} c_{M_l}} = n_{c_{M_k}\cdot} \cdot n_{c_{\cdot M_l}} / N$, the number expected when assuming independence of the variables under given marginal distributions. With these absolute frequencies, the corresponding classical association measure χ^2 can be calculated for each possible combination of clusterings of method M_k and M_l. Then, $\chi^2(M_k, M_l)$ indicates the degree of

association of the clustering results of method M_k and method M_l:

$$\chi^2(M_k, M_l) = \sum_{c_{M_l}=1}^{K} \sum_{c_{M_k}=1}^{K} \frac{(n_{c_{M_k} c_{M_l}} - \tilde{n}_{c_{M_k} c_{M_l}})^2}{\tilde{n}_{c_{M_k} c_{M_l}}}. \quad (1)$$

As it is known that χ^2 is influenced by the number of considered observations N, as well as by the number of K clusters, it is common practice to adjust for these parameters (see e.g. [24, Chapter 4]). For that purpose, we base our subsequent analysis of the method-association-measures on the adjusted *contingency coefficient* C

$$C(M_k, M_l) = \sqrt{\frac{K}{K-1}} \cdot \sqrt{\frac{\chi^2(M_k, M_l)}{N + \chi^2(M_k, M_l)}}, \quad (2)$$

and Cramér's V

$$V(M_k, M_l) = \frac{\chi^2(M_k, M_l)}{N \cdot (K-1)}, \quad (3)$$

which allow a comparison of association values across the data sets.

C. Φ-Coefficient

For the following construction we take a different perspective, directly investigating whether pairs of observations are assigned (dis)similarly. It is therefore of relevance whether a pair of unequal observations $\{i, j\}$, with $i, j \in \mathcal{I}$ is grouped into the same cluster or not by method M:

$$P_M(\{i, j\}) = \begin{cases} 1, & \text{if } M(i) = M(j) \\ 0, & \text{else.} \end{cases}$$

Considering the (dis)similarity over all pair-wise combinations results in the $\binom{N}{2}$ dimensional vector $\mathcal{P}_M$. With the same aim as above, declaring the categories of concordant and discordant clusterings of method M ($\mathcal{P}_M$) as attributes, allows to adapt the classical Φ-coefficient (see e.g. [24, Chapter 4]). In contrast to the upper defined method-association-measures we now incorporate the (dis)similarity of methods based on pair-wise observations:

$$\Phi(\mathcal{P}_{M_k}, \mathcal{P}_{M_l}) = \frac{\chi^2(\mathcal{P}_{M_k}, \mathcal{P}_{M_l})}{\binom{N}{2}}, k \neq l, k, l \in \{1, ..., q\}. \quad (4)$$

This method-associated measure describes the association of pair-wise obtained clusterings of method M_k and M_l and can here also be interpreted as the Pearson correlation[2].

D. How to interprete the association measures

The proposed association measures quantify the strength of association of methods in a very convenient way, delivering an easily interpretable range of values. An association of 0 indicates no association between the different methods, whereas an association of 1 describes a perfect association of the methods. The above introduced equivalence between the Φ-coefficient and the Pearson correlation (see footnote 2) provides easy access for interpretation, as correlation measures are a popular tool in applied science. Furthermore, there exists a broad range of the general interpretation of the association measures for different application fields such as medicine, psychology, politics, and social science (see e.g. [26], [27], [28], [29], and [30]). These guidelines allow for a context-specific categorization of different degrees of associations. However, it is of note that these guidelines are based on humanly conducted experiments instead of observations obtained by different data-driven methods. As far as we know, we are the first investigating the degree of association obtained from different clustering methods instead of human experiments, so no appropriate guideline is available yet as more applications are needed.

IV. Simulation data

For the investigation of our association accuracy heuristic, we construct simulation data of two sub-populations (groups) with three different degrees of mutual dependence. We assume a multivariate normal distribution $\mathcal{MVN}_1 = \mathcal{MVN}(\mu_1, \Sigma_1)$ for simulation group 1 and $\mathcal{MVN}_2^{(\rho)} = \mathcal{MVN}(\mu_2, \Sigma_2^{(\rho)})$ for simulation group 2. For Σ_1 we assume the empirical covariance matrix of the underlying cell population "4cell" of the data set of [31] as true, and extract the covariance matrix for p randomly chosen variables. $\Sigma_2^{(\rho)}$ contains the same variances as simulation group 1. The covariances of simulation group 2 are constructed in dependence of the standard deviations given by Σ_1. With fixed correlation ρ, it is therefore possible to generate $\Sigma_2^{(\rho)}$ which enables a simulation study with different dependence structures[3]. This allows to simulate the second group as *weakly* (ρ = 0.1), *moderately* (ρ = 0.5) and *highly* (ρ = 0.9) correlated to the first group. We refer to these simulation settings as simulation data with a *low*, *moderate*, and *strong* dependence structure.

The purpose of our simulation data is that in case of a *low* dependence structure a *low* overlap of the two subgroups is simulated. With a higher simulated dependence it is more difficult to distinguish the underlying groups and the different clustering algorithms should differ more in their clustering results. This simulation setting aims to construct settings with lowly, moderately and strongly associated clustering results. Given our heuristic is correct, we expect that overall highly associated methods, simulated by the *low* dependence structure

[2]Following [25] the Φ-coefficient is equivalent to the Pearson correlation applied to $0/1$ variables.

[3]The formula $Cov(X_q, X_r) = \rho \cdot \sqrt{Var(X_q)} \cdot \sqrt{Var(X_r)} \ \ \forall q \neq r \in \{1, \ldots, p\}$ allows the construction of the covariance matrix for all $p = 1000$ covariates of simulation group 2 in dependence of $\rho \in \{0.1, 0.5, 0.9\}$, maintaining the same variances as simulation group 1, leading to $\Sigma_2^{(\rho)}$. We randomly sample $N_1 = 50$ times out of $\mathcal{MVN}_1 = \mathcal{MVN}(\mu_1, \Sigma_1)$ and $N_2 = 50$ times out of $\mathcal{MVN}_2^{(\rho)} = \mathcal{MVN}(\mu_2, \Sigma_2^{(\rho)})$. For robustness, we repeat this procedure 10 times, generating 10 simulation data sets for each of the three determined dependence structures, resulting in 30 simulation data sets with $N = 100$ observations and $p = 1000$ covariates. For each iteration, we sample p times randomly out of the set μ_1 and μ_2, which have been determined with $[100, 200]$ and $[0.41, 20]$ with steps of 0.0001. In a Bachelor thesis supervised by us, [32] investigates the influence of different dependence structures within the simulation groups, underlying a zero-inflated negative binomial distribution.

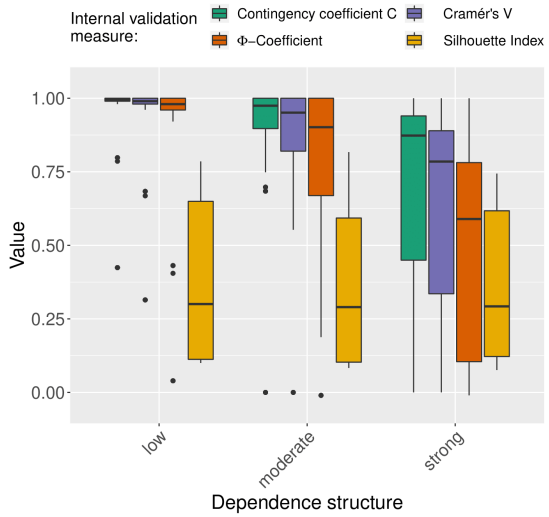

Fig. 2: Association values of the adjusted contingency coefficient C (green), Cramér's V (violet), the Φ-coefficient (orange), and the silhouette index (yellow) for the clustering methods of the simulation data with low, moderate and strong dependence structure.

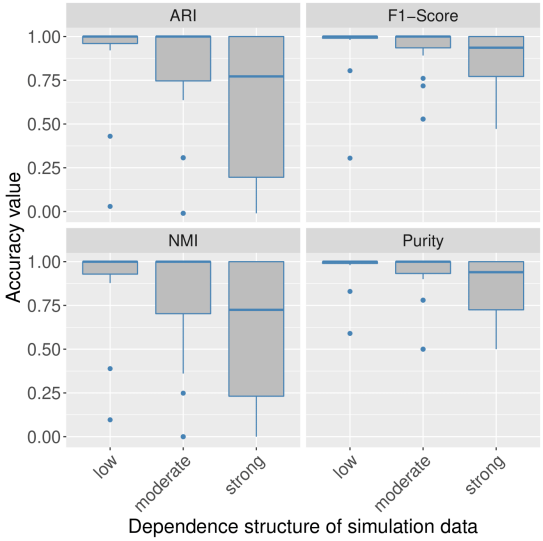

Fig. 3: Accuracy values of clustering methods for each of the simulation data with low, moderate and strong dependence structure, evaluated by the adjusted Rand index (ARI), the Normalized Mutual Information (NMI), the F1-score, and the purity.

result in an overall high accuracy, whereas lower associated methods, simulated with a *strong* dependence structure are expected with an overall lower accuracy, including methods with bad performance as the underlying subgroup structures are expected to be hard to detect for some algorithm(s).

As we will consider high dimensional data sets, a dimension reduction or manifold embeddings are needed before the subsequent clustering can be assessed. We apply the context-related clustering algorithms for single-cell RNA-sequencing data: *SC3* [5], *adaSC3* [10], *phate* [11], as well as *umaps* [12] in combination with *k-means* [33]. The generation of the method-association-measures, including the adjusted contingency coefficient C and Cramér's V, can still be calculated using the *DescTools* R-package [34]. The Φ-coefficient is calculated with the correlation function of the *stats* R-package [35]. As a competitor for internal validation, we include the silhouette index of the R-package *cluster* [36]. To study the performance of the clusterings, we will include the underlying ground truth. This allows the consideration of the adjusted Rand index (ARI) [37], which was applied for the original evaluation of *SC3* and *adaSC3* [5], [10], using the R package *mclust* of [38]. We further investigate the Normalized Mutual Information (NMI), accessible by the R package *aricode* [39], as well as the F1-score, which is provided by the R package *MLmetrics* [40]. As a fourth accuracy measure we include the purity, which is based on the R package *funtimes* [41].

Compared to the highly elaborate process of data collection of single-cell RNA-sequencing data, as well as the high effort for the alternative of manual inspection, we see the effort of analyzing the sequenced data simultaneously by several different clustering methods as more than justified. The simulation study has been conducted with a computer containing a 2.60 GHz Intel (R) Xeon(R) processor with 64 GB 3000 Mhz of RAM. The computation time for generating the simulation data is 43 minutes. The performed clustering for all simulation data takes in total 4.03 minutes. The calculation of all association measures of the whole simulation study is performed in 11.95 seconds.

A. Association

With regard to Fig. 2, we can see the values of the internal validation measures of the different clustering methods. The figure contains the results of the constructed method-association-measures: adjusted Contingency Coefficient C (green), Cramér's V (violet) and the Φ-coefficient (orange) for each simulated dependence structure. In addition, the silhouette index (yellow) for each method is included. We see that with higher simulated dependence structures, the values of our method-association-measures decrease. In addition, the variation of associated clustering results gets higher with increasing dependence structure. However, the silhouette index remains the same over all three settings. In addition, the boxplots, illustrating the *moderate* setting, indeed lie in between of the results of the more extreme settings. In the case of simulated *low* dependence, the deviation of association among the same clustering methods is very low, represented by small boxes. In case of simulated *strong* dependence, the

deviations of associated clustering results get larger, which is also reflected by the resulting larger boxes including their lower bounds of the corresponding boxplots.

B. Accuracy

For the simulated *low* dependence, the accuracy values shown in Fig. 3 are very high with very low deviations for all four accuracy measures. In case of the *moderate* setting, the median of the accuracy value remains very high in all measures. The ARI and the NMI show now higher deviation of accuarcy compared to the previous setting. For the simulation data that was generated with a *strong* dependence structure, the median slightly drops in case of the F1-score and the purity, and even a little bit more for the remaining two accuracy measures. Especially in the simulation setting underlying a strong dependence structure, a very poor performance is reached for some clustering methods.

C. On the Relationship between Association and Accuracy

Concerning the simulation data with *low* dependence structure, the median of all method-association-measures is especially high and become smaller the stronger the dependence structure. In contrast to these method-association-measures, the silhouette index remains approximately the same over the different settings. Over all four accuracy measures one can see that the median of the accuracy value is lower the stronger the dependence structure of the simulation data. Furthermore, the deviation of performance gets higher, the stronger the dependence structure of the simulation data. Comparing the association values of the method-association-measures to the accuracy values, we can see that Cramér's V is quite close to the accuracy values of the ARI and the NMI. As the two figures contain a different number of observations, a direct comparison of the boxplot value ranges is not recommendable. This is the reason, why we consider the rankings in the following. With increasing dependence structure of our simulation settings, the rankings of our method-associated-measures show the same order for the accuracy. The rankings of each method association measure for increasing dependence structure is perfectly correlated to each other as well as to the according ranks of accuracy, which would respectively lead to a Spearman's rank correlation coefficient of 1 (see e.g. [24, Chapter 4]). However, this is not the case for the internal validation measure of the silhouette index. With regard to our method-association-measures, we can state that the claim of our association accuracy heuristic is completely fulfilled in the simulation data considered.

V. Clustering of Single-cell RNA-Sequencing Data

The validation of the association accuracy heuristic is studied on the same single-cell RNA-sequencing data that are also part of the original works of [5] and [10], proposing *SC3* and *adaSC3*[4]. In addition, we add the data set of Darmanis et al. [42] and Fan et al. [43] as further benchmark data sets. Each data set examines another biological research questions, which can be categorized by the following. The benchmark data describe cell differentiation (e.g. [22], [42], [44], [45] and [46]), known cell types (e.g. [47], [48]), as well as the discovery of decisive genes characterizing mammalian cells (e.g. [49], [43]). The experiments contain N single cells, p genes as well as G true underlying groups, as specified in Table II.

TABLE II: Original single-cell RNA-sequencing data sets including N single cells, p genes and G cell types.

Data set	N	p	G
Biase et al. [22]	49	25,734	3
Darmanis et al. [42]	466	22,088	9
Deng et al. [49]	269	22,431	10
Fan et al. [43]	63	26,357	7
Goolam et al. [44]	124	41,428	5
Kolodziejczyk et al. [45]	704	38,616	3
Pollen et al. [47]	64	23,710	4
Treutlein et al. [46]	80	23,271	5
Yan et al. [48]	90	20,214	7

For the following analysis, we first evaluate the internal validation given the number of clusters K corresponds to the number of specified cell types G, setting $K := G$. In a second step, we investigate the association measures as well as the silhouette index with $K = 2, \cdots, 10$ clusters. Doing so, we aim to assess whether the introduced association measures might also serve for choosing the correct number of clusters. With regard to the application of our association accuracy heuristic on scRNA-seq data sets, we see the additional calculation time for the association values also as more than justified here. The clustering and the calculation of association was run on the same computer as described above. The clustering of all methods for one K takes about 1 minute for most of the data sets. In case of the Kolodziejczyk et al. data set which contains the highest number of single cells, one run of all clustering methods takes 45 minutes. The calculation of the association measures over all Ks, lies between 1 and 16 minutes.

In this section we start with the description of the association and accuracy of the clustering results of the benchmark data provided in Section V-A and in Section V-B. We then utilize both measures to evaluate the association accuracy heuristic (see Section V-D). In Section V-C we take a look at the capability of the proposed method-association-measures for choosing the correct number of underlying groups.

[4]The subsequent application of *SC3* and *adaSC3* differs from the original version in that no gene filter is applied to favor no clustering method. In addition, we limited our application to $N = 500$ single cells to ensure traceable, unsupervised clustering, as justified in [5], [10].

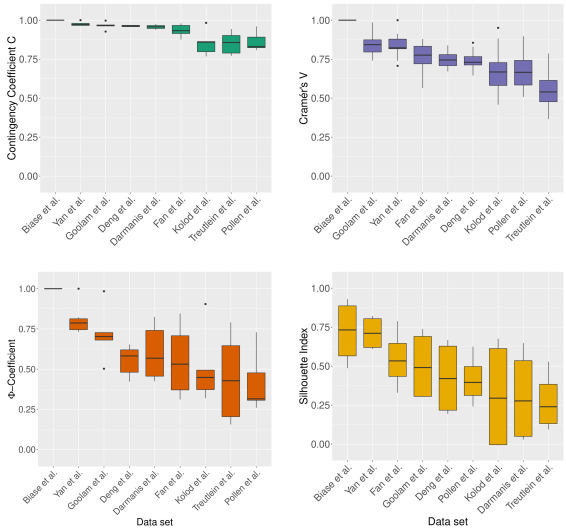

Fig. 4: Association values of the contingency coefficient C (green), Cramér's V (violet) and the adjusted Φ-coefficient (orange) as well as the silhouette index (yellow) for the methods of the benchmark data.

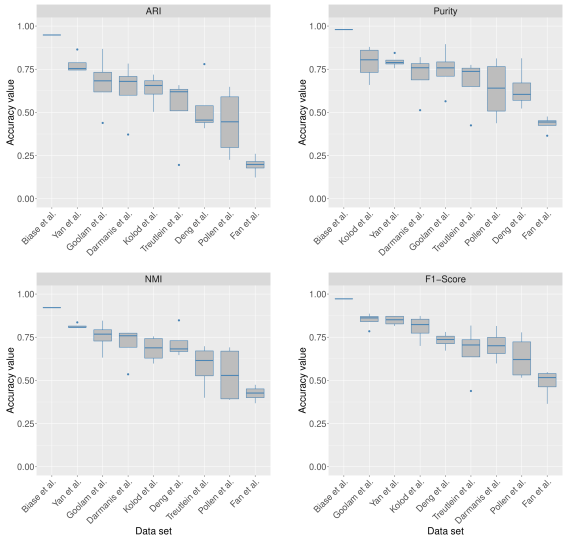

Fig. 5: Accuracy values of clustering methods of the benchmark data, evaluated by the adjusted Rand index (ARI), the Normalized Mutual Information (NMI), the F1-score, and the purity.

A. Association of the Clustering Results of Different Methods

In Fig. 4, the benchmark data are ordered according to their descending medians of the respective internal validation values. We can see that all four measures agree in ranking the Biase et al. data set first, as this data set shows its highest median among all the remaining benchmark data sets. All method-association-measures assess the corresponding clustering results of this data set as perfectly associated. Although, the Biase et al. data set is ranked first, the median of the silhouette indices of each method is approximately 0.75 and does by far not reach its maximal value of 1. We can state that for all single-cell RNA-sequencing data, the boxplot of the silhouette index indicates always considerably higher variation among the clustering results compared to the method-association-measures.

Concerning the following rankings, the contingency coefficient C proposes exactly the same order of data sets as the Φ-coefficient. Compared to the Φ-coefficient, the contingency coefficient C always shows higher association values with a (considerably) lower standard deviation. The ranking based on Cramér's V is only marginally different. With exception of the data sets of Fan et al. and Deng et al., Cramér's V differs in the order by maximally one rank compared to the remaining association measures. In case of Cramér's V, one could even argue to place the boxplot of Fan et al. after the Deng data set due to its lower whisker, respecting the 25% quantile in the total ranking. The resulting change of the order would make Cramér's V more comparable to the other method-association-measures. The rankings of the silhouette index differ the most in comparison to all the other considered measures. In general, the observation-based association measures (contingency coefficient C and Cramér's V) have the tendency of indicating higher association values with less deviations. The pair-wise-based measure of the Φ-coefficient assesses the association with generally lower association values. Overall, the silhouette index shows the lowest values.

Measuring the correlation of the rankings with Spearman's rank correlation coefficient based on the medians of each internal validation measure leads to the left part of Table III. We can see, in accordance with the description of the above association, that the contingency coefficient C and the Φ-coefficient are perfectly correlated to each other. With a Spearman correlation coefficient of 0.9, Cramér's V is also highly correlated to these measures. The silhouette index is correlated with approximately 0.8 to each of the other association measures.

B. Accuracy

For the evaluation of our association accuracy heuristic, the investigation of accuracy (see Fig. 5) is mandatory. The clustering methods in case of the Biase et al. data set show very high accuracy for all four external validation measures. With exception to the data set of Deng et al. and Treutlein et al., the ARI and the NMI propose exactly the same descending order for all remaining benchmark data sets. Also, the rankings of the overall F1-score show quite similar rankings of the data sets, whereas the purity differs the most to the remaining accuracy measures. For the data sets of Treutlein et al., Pollen et al. and Fan et al., the adjusted Rand index indicates a poor performance for some methods. In addition to that, in case of the Pollen et al. data set, all measures indicate a very high

TABLE III: Spearman's rank correlation coefficients of the association measures and the accuracy measures for the considered benchmark data sets

	Association				Accuracy			
	contingency coefficient C	*Cramér's V*	*Φ-coefficient*	*silhouette index*	*ARI*	*NMI*	*purity*	*F1-score*
contingency coefficient C	1.00	0.90	1.00	0.75	0.77	0.83	0.47	0.75
Cramér's V	0.90	1.00	0.90	0.83	0.65	0.70	0.38	0.60
Φ-coefficient	1.00	0.90	1.00	0.75	0.77	0.83	0.47	0.75
Silhouette	0.75	0.83	0.75	1.00	0.38	0.45	0.18	0.47
ARI	0.77	0.65	0.77	0.38	1.00	0.98	0.87	0.87
NMI	0.83	0.70	0.83	0.45	0.98	1.00	0.83	0.88
purity	0.47	0.38	0.47	0.18	0.87	0.83	1.00	0.73
F1-score	0.75	0.60	0.75	0.47	0.87	0.88	0.73	1

TABLE IV: Benchmark data, showing the averaged internal validation measures (and its standard deviations) over the clustering results of each data set for differing number of clusters K.

Data set	Internal validation measure	2	3	4	5	6	7	8	9	10
Biase	*contingency coefficient C*	0.69 (0.22)	**1.00*** (0.00)	0.95 (0.02)	0.94 (0.03)	0.93 (0.03)	0.95 (0.03)	0.96 (0.02)	0.97 (0.01)	0.97 (0.02)
	Cramér's V	0.59 (0.26)	**1.00*** (0.00)	0.85 (0.06)	0.78 (0.11)	0.73 (0.10)	0.78 (0.10)	0.77 (0.09)	0.80 (0.07)	0.80 (0.09)
	Φ-coefficient	0.31 (0.40)	**1.00*** (0.00)	0.77 (0.13)	0.61 (0.21)	0.58 (0.22)	0.64 (0.24)	0.59 (0.25)	0.65 (0.19)	0.58 (0.17)
	silhouette index	0.61 (0.23)	**0.72*** (0.21)	0.62 (0.11)	0.61 (0.07)	0.59 (0.12)	0.49 (0.18)	0.50 (0.12)	0.47 (0.07)	0.45 (0.10)
Darmanis	*contingency coefficient C*	0.68 (0.34)	0.88 (0.08)	0.94 (0.03)	0.94 (0.03)	0.95 (0.02)	**0.96** (0.02)	**0.96** (0.02)	**0.96*** (0.01)	**0.96** (0.02)
	Cramér's V	0.62 (0.39)	0.74 (0.16)	**0.81** (0.07)	0.78 (0.07)	0.77 (0.05)	0.78 (0.06)	0.77 (0.07)	0.74* (0.06)	0.74 (0.08)
	Φ-coefficient	0.50 (0.51)	0.59 (0.26)	**0.70** (0.14)	0.65 (0.22)	0.64 (0.19)	0.69 (0.21)	0.67 (0.21)	0.60* (0.17)	0.55 (0.12)
	silhouette index	0.28 (0.31)	0.31 (0.34)	0.32 (0.33)	0.32 (0.33)	0.30 (0.31)	**0.33** (0.34)	0.31 (0.31)	0.31* (0.31)	0.27 (0.27)
Deng	*contingency coefficient C*	0.70 (0.32)	0.81 (0.15)	0.91 (0.02)	0.92 (0.03)	0.94 (0.03)	0.93 (0.03)	0.96 (0.02)	**0.97** (0.01)	0.96* (0.01)
	Cramér's V	0.65 (0.37)	0.65 (0.19)	0.74 (0.05)	0.74 (0.08)	0.74 (0.09)	0.69 (0.08)	0.76 (0.09)	**0.78** (0.03)	0.75* (0.02)
	Φ-coefficient	0.50 (0.51)	0.54 (0.31)	**0.73** (0.11)	**0.73** (0.18)	0.70 (0.14)	0.60 (0.12)	0.55 (0.18)	0.55 (0.10)	0.55* (0.09)
	silhouette index	0.41 (0.38)	**0.44** (0.38)	0.38 (0.32)	0.40 (0.30)	0.42 (0.31)	0.42 (0.30)	**0.44** (0.28)	0.43 (0.25)	0.43* (0.25)
Deng	*contingency coefficient C*	0.70 (0.14)	0.81 (0.00)	0.91 (0.03)	0.92 (0.02)	0.94 (0.03)	0.93 (0.04)	0.96 (0.04)	**0.97** (0.02)	0.96* (0.01)
	Cramér's V	0.65 (0.20)	0.65 (0.01)	0.74 (0.07)	0.74 (0.06)	0.74 (0.08)	0.69 (0.12)	0.76 (0.14)	**0.78** (0.08)	0.75* (0.06)
	Φ-coefficient	0.50 (0.36)	0.54 (0.03)	**0.73** (0.13)	**0.73** (0.18)	0.70 (0.17)	0.60 (0.22)	0.55 (0.21)	0.55 (0.16)	0.55* (0.16)
	silhouette index	0.41 (0.27)	**0.44** (0.26)	0.38 (0.19)	0.40 (0.21)	0.42 (0.17)	0.42 (0.20)	**0.44** (0.17)	0.43 (0.20)	0.43* (0.20)
Fan	*contingency coefficient C*	0.80 (0.34)	**0.98** (0.06)	0.94 (0.03)	0.92 (0.02)	0.93 (0.03)	0.93* (0.02)	0.95 (0.02)	0.96 (0.02)	0.97 (0.02)
	Cramér's V	0.71 (0.39)	**0.94** (0.14)	0.80 (0.07)	0.74 (0.08)	0.74 (0.10)	0.72* (0.09)	0.76 (0.09)	0.76 (0.07)	0.77 (0.11)
	Φ-coefficient	0.48 (0.58)	**0.88** (0.24)	0.68 (0.14)	0.62 (0.15)	0.58 (0.20)	0.55* (0.21)	0.56 (0.20)	0.59 (0.18)	0.54 (0.21)
	silhouette index	0.47 (0.33)	0.50 (0.35)	0.52 (0.26)	0.53 (0.23)	0.58 (0.18)	0.55* (0.14)	**0.58** (0.15)	0.54 (0.16)	0.49 (0.15)
Goolam	*contingency coefficient C*	0.69 (0.33)	0.94 (0.08)	0.93 (0.03)	**0.96*** (0.02)	0.95 (0.02)	0.94 (0.02)	0.95 (0.02)	0.95 (0.01)	0.95 (0.01)
	Cramér's V	0.64 (0.38)	**0.87** (0.14)	0.79 (0.09)	0.86* (0.06)	0.79 (0.08)	0.74 (0.07)	0.74 (0.09)	0.73 (0.07)	0.72 (0.06)
	Φ-coefficient	0.47 (0.50)	**0.78** (0.21)	0.64 (0.15)	0.72* (0.09)	0.61 (0.07)	0.57 (0.09)	0.55 (0.13)	0.49 (0.13)	0.46 (0.11)
	silhouette index	0.49 (0.31)	**0.53** (0.37)	0.50 (0.38)	0.51* (0.36)	**0.53** (0.37)	0.51 (0.37)	0.48 (0.36)	0.47 (0.41)	0.47 (0.40)
Kolod	*contingency coefficient C*	0.69 (0.13)	0.85 (0.02)	0.93* (0.07)	0.95 (0.07)	0.95 (0.05)	0.96 (0.03)	**0.98** (0.02)	0.97 (0.02)	**0.98** (0.02)
	Cramér's V	0.64 (0.17)	0.70 (0.04)	0.80* (0.12)	0.81 (0.18)	0.78 (0.14)	0.80 (0.10)	**0.85** (0.08)	0.81 (0.08)	0.82 (0.07)
	Φ-coefficient	0.53 (0.27)	0.50 (0.04)	0.67* (0.20)	0.72 (0.29)	0.67 (0.31)	0.70 (0.23)	**0.76** (0.20)	0.73 (0.16)	**0.76** (0.12)
	silhouette index	0.29 (0.27)	**0.31** (0.29)	0.29* (0.28)	0.26 (0.31)	0.29 (0.32)	0.29 (0.32)	0.30 (0.28)	0.28 (0.30)	0.28 (0.29)
Pollen	*contingency coefficient C*	0.85 (0.10)	0.87 (0.07)	0.86* (0.06)	0.88 (0.05)	0.89 (0.04)	0.90 (0.03)	0.91 (0.04)	**0.93** (0.02)	**0.93** (0.01)
	Cramér's V	**0.77** (0.15)	0.73 (0.13)	0.66* (0.12)	0.65 (0.09)	0.64 (0.09)	0.62 (0.06)	0.63 (0.11)	0.65 (0.08)	0.64 (0.04)
	Φ-coefficient	**0.59** (0.26)	0.50 (0.25)	0.41* (0.18)	0.42 (0.15)	0.42 (0.19)	0.39 (0.13)	0.42 (0.22)	0.38 (0.15)	0.37 (0.09)
	silhouette index	**0.41** (0.23)	**0.41** (0.19)	**0.41*** (0.17)	0.39 (0.17)	0.36 (0.18)	0.33 (0.20)	0.33 (0.17)	0.31 (0.18)	0.31 (0.21)
Treutlein	*contingency coefficient C*	0.78 (0.17)	0.86 (0.07)	0.85 (0.08)	0.85* (0.07)	0.90 (0.03)	0.91 (0.04)	0.91 (0.04)	0.91 (0.04)	**0.92** (0.04)
	Cramér's V	**0.54** (0.21)	**0.54** (0.11)	0.52 (0.16)	0.44* (0.12)	0.36 (0.08)	0.33 (0.10)	0.31 (0.10)	0.28 (0.09)	0.25 (0.11)
	Φ-coefficient	0.47 (0.27)	**0.78** (0.19)	0.64 (0.26)	0.72* (0.28)	0.61 (0.11)	0.57 (0.13)	0.55 (0.11)	0.49 (0.14)	0.46 (0.12)
	silhouette index	**0.43** (0.20)	0.39 (0.19)	0.31 (0.25)	0.28* (0.20)	0.28 (0.19)	0.28 (0.17)	0.28 (0.20)	0.26 (0.20)	0.24 (0.20)
Yan	*contingency coefficient C*	**1.00** (0.00)	0.95 (0.04)	0.95 (0.02)	0.97 (0.02)	0.97 (0.01)	0.98* (0.01)	0.97 (0.01)	0.98 (0.01)	0.98 (0.01)
	Cramér's V	**1.00** (0.00)	0.87 (0.09)	0.83 (0.05)	0.88 (0.07)	0.87 (0.05)	0.87* (0.07)	0.83 (0.06)	0.84 (0.06)	0.86 (0.05)
	Φ-coefficient	**1.00** (0.00)	0.74 (0.17)	0.78 (0.08)	0.84 (0.10)	0.83 (0.08)	0.81* (0.10)	0.73 (0.14)	0.69 (0.13)	0.69 (0.12)
	silhouette index	0.53 (0.36)	0.58 (0.30)	0.60 (0.24)	0.64 (0.16)	0.68 (0.16)	**0.71*** (0.11)	0.67 (0.09)	0.64 (0.12)	0.63 (0.12)

variation of accuracy concerning the different methods. With regard to the Spearman's rank coefficient in Table III, we can see that all accuracy measures are very highly correlated to each other, based on the rankings of their median accuracy value. Furthermore, the ARI and NMI are almost perfectly correlated, and also highly correlated to the overall F1-score and to the purity.

C. Detection of the underlying K

In order to find out whether the corresponding association measures could also be a measure for choosing the correct K, we investigate the internal validation measures for $K = 2, \cdots, 10$ (see Table IV). We consider both the average values for each method as well its standard deviation, provided in brackets. In bold we highlight the highest value of each

measure or the different data sets. The asterisk indicates the true underlying number of groups of the benchmark data. We can see that among our association measures, the contingency coefficient C most often detects the true underlying number of cell types (3 times out of 9 data sets). This is exactly the same frequency as for the silhouette index. Both measures show their highest averaged values for the Biase et al. data set, whereas the other data sets differ in their reached maximum values of the two measures, selecting the correct number of underlying cell types. Cramér's V and the Φ-coefficient detect the true underlying number of groups only in the case of the Biase et al. data set.

D. On the Relationship between Association and Accuracy

In the description of association and accuracy, we have seen some similar rankings in case of the the contingency coefficient C and the Φ-coefficient, as well as for the ARI and NMI. Relating association and accuracy to each other, we can observe that the value range of the Φ-coefficient shows a tendency of related NMI values. Especially the clustering methods applied to the data sets Treutlein et al. and Pollen et al. are both poorly associated and have a poor performance. In case of the Daramanis et al. data set, a high variation of the Φ-coefficient can be observed. This might give a hint that at least one method performs differently compared to the others. With regard to the corresponding accuracy values, we can see that indeed one method leads to considerably lower performance. For the same association measure a high variation is indicated for the lower ranked data sets of Fan et al. and Treutlein et al., which results in really bad performance of some methods.

For a further validation of our association accuracy heuristic, we take the Spearman's rank correlation coefficients of Table III into account. Both the contingency coefficient C as well as the Φ-coefficient are highly correlated to the ARI and to the NMI. Cramér's V is less correlated to these accuracy measures. The silhouette index only shows poor or fair correlation with regard to purity, ARI and NMI.

VI. Conclusion

In this paper, we propose a heuristic that is especially useful for high-dimensional settings, where user inspection easily becomes infeasible. We demonstrated that our association accuracy heuristic works perfectly for the constructed simulation data under systematically varied dependence and allows a trustworthy ranking in case of the single-cell RNA-sequencing benchmark data sets. In the simulation study, Cramér's V seemed to work out best, whereas in case of the real data, the Φ-coefficient seems best for relating association with accuracy. However, we emphasize that all method-association-measures deliver very reliable rankings that have the tendency to reflect well the order of the overall accuracy of the different data sets.

With the Spearman correlation we have been able to show that highly associated methods are indeed highly correlated with their accuracy. We therefore see our heuristic as validated in both the simulated data and real data, and state that it is definitively worth more investigating. Furthermore, we see a big advantage in case of the pair-wise constructed Φ-coefficient as it is nicely interpretable. In addition to that, the Φ-coefficient brings along the benefit that the association of groupings with differing K can be analyzed as its construction is based on considering pair-wise (dis)similar groupings. This could bring the advantage of incorporating e.g. the context specific clustering method Seurat, which is also often used for determining new cell types.

Our heuristic provides the user a quite powerful tool for internal validation in situations where ground truth is not easily available. In situations, where several context related state-of-the-art methods are lowly associated, no automatic determination of cell types is recommended. This might prevent the blind trust into one single method.

As our heuristic is easily applicable, no background knowledge of the methods is requested, which could bring along the drawback that different methods could be applied without questioning its suitability in the specific application. We do not recommend the silhouette index as it only shows a low correlation to accuracy. Unfortunately, we cannot give a general recommendation of the best performing method-association-measure. But we can recommend all the proposed method-association-measures as they show a high correlation to accuracy.

Of course, further studies are required, on benchmark data sets as well as on refined simulation settings, also including more than two simulated sub-populations. We claim that our heuristic is not only limited to clustering methods. Furthermore, the investigation of the relationship between association and accuracy could be even more interesting with a higher number of methods, an aspect that has not been investigated in detail but clearly is of further research interest.

For determining the correct number of underlying groups, the contingency coefficient only succeeded in very highly associated situations. This could be an argument that our method-association-measures might not deliver the correct number of underlying groups but provides access to stable partitions over different methods. As we only have access to ground truth, nobody knows the reality. This might be an argument for sticking to the agreement of methods as we do it in case of expert decisions.

Acknowledgment

We are very grateful for the inspiring comments of several reviewers.

References

[1] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui *et al.*, "mRNA-Seq whole-transcriptome analysis of a single cell," *Nature Methods*, vol. 6, no. 5, pp. 377–382, 2009.

[2] L. Zappia, B. Phipson, and A. Oshlack, "Splatter: simulation of single-cell rna sequencing data," *Genome Biology*, vol. 18, no. 1, pp. 1–15, 2017.

[3] M. B. Pouyan, V. Jindal, J. Birjandtalab, and M. Nourani, "Single and multi-subject clustering of flow cytometry data for cell-type identification and anomaly detection," *BMC Medical Genomics*, vol. 9, no. 2, pp. 99–110, 2016.

[4] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, "Spatial reconstruction of single-cell gene expression data," *Nature Biotechnology*, vol. 33, no. 5, pp. 495–502, 2015.

[5] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green *et al.*, "SC3: consensus clustering of single-cell RNA-seq data," *Nature Methods*, vol. 14, no. 5, pp. 483–486, 2017.

[6] A. Duò, M. D. Robinson, and C. Soneson, "A systematic performance evaluation of clustering methods for single-cell RNA-seq data," *F1000Research*, vol. 7, 2018.

[7] T. Ullmann, C. Hennig, and A.-L. Boulesteix, "Validation of cluster analysis results on validation data: A systematic framework," *arXiv preprint arXiv:2103.01281*, 2021.

[8] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, 2005.

[9] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[10] C. Fuetterer, T. Augustin, and C. Fuchs, "Adapted single-cell consensus clustering (adasc3)," *Advances in Data Analysis and Classification*, vol. 14, pp. 885–896, 2020.

[11] K. R. Moon, D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. van den Elzen, M. J. Hirn, R. R. Coifman, N. B. Ivanova, G. Wolf, and S. Krishnaswamy, "Visualizing structure and transitions in high-dimensional biological data," *Nature Biotechnology*, vol. 37, no. 12, pp. 1482–1492, 2019.

[12] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *The Journal of Open Source Software*, 2018.

[13] T. Hastie, R. Tibshirani, and J. Friedman, "Unsupervised learning," in *The Elements of Statistical Learning*. Springer, 2009, pp. 485–585.

[14] M. Hassani and T. Seidl, "Using internal evaluation measures to validate the quality of diverse stream clustering algorithms," *Vietnam Journal of Computer Science*, vol. 4, no. 3, pp. 171–183, 2017.

[15] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *2010 IEEE International Conference on Data Mining*. IEEE, 2010, pp. 911–916.

[16] D. Pfitzner, R. Leibbrandt, and D. Powers, "Characterization and evaluation of similarity measures for pairs of clusterings," *Knowledge and Information Systems*, vol. 19, no. 3, pp. 361–394, 2009.

[17] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data," *Nature Reviews Cancer*, vol. 8, no. 1, pp. 37–49, 2008.

[18] S. Balbi, "Beyond the curse of multidimensionality: high dimensional clustering in text mining," *Statistica Applicata – Italian Journal of Applied Statistics*, vol. 22, no. 1, pp. 53–63, 2010.

[19] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *International Conference on Database Theory*. Springer, 1999, pp. 217–235.

[20] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity*, 1998.

[21] Z. Gu, *cola: A Framework for Consensus Partitioning*, 2019. [Online]. Available: https://github.com/jokergoo/cola

[22] F. H. Biase, X. Cao, and S. Zhong, "Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing," *Genome Research*, vol. 24, no. 11, pp. 1787–1796, 2014.

[23] B. Treutlein, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, and S. R. Quake, "Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq," *Nature*, vol. 509, no. 7500, pp. 371–375, 2014.

[24] C. Heumann, M. Schomaker, and Shalabh, *Introduction to statistics and data analysis*. Springer, 2016.

[25] H. Cramer, *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[26] Y. Chan, "Biostatistics 104: correlational analysis," *Singapore Med J*, vol. 44, no. 12, pp. 614–619, 2003.

[27] C. P. Dancey and J. Reidy, *Statistics without maths for psychology*. Pearson Education, 2007.

[28] H. Akoglu, "User's guide to correlation coefficients," *Turkish Journal of Emergency Medicine*, vol. 18, no. 3, pp. 91–93, 2018.

[29] P. D. Ellis, *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press, 2010.

[30] K. D. M. E. Handayeni and B. S. P. Ariyani, "Commuters' travel behaviour and willingness to use park and ride in tangerang city," *IOP Conference Series: Earth and Environmental Science*, vol. 202, 2018.

[31] M. Goolam, A. Scialdone, S. J. Graham, I. C. Macaulay, A. Jedrusik, A. Hupalowska, T. Voet, J. C. Marioni, and M. Zernicka-Goetz, "Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos," *Cell*, vol. 165, no. 1, pp. 61–74, 2016.

[32] F. Stermann, "On the influence of dependence structures on clustering performance (in german)," *Bachelor thesis, Department of Statistics, LMU Munich*.

[33] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognition*, vol. 36, no. 2, pp. 451–461, 2003.

[34] A. Signorell, K. Aho, A. Alfons, N. Anderegg, T. Aragon, A. Arppe *et al.*, *DescTools: Tools for Descriptive Statistics*, 2020, R package version 0.99.36. [Online]. Available: https://cran.r-project.org/package=DescTools

[35] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[36] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik, *cluster: Cluster Analysis Basics and Extensions*, 2021, R package version 2.1.2 . [Online]. Available: https://CRAN.R-project.org/package=cluster

[37] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[38] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery, "mclust 5: clustering, classification and density estimation using Gaussian finite mixture models," *The R Journal*, vol. 8, no. 1, pp. 289–317, 2016. [Online]. Available: https://doi.org/10.32614/RJ-2016-021

[39] J. Chiquet, G. Rigaill, and M. Sundqvist, *aricode: Efficient Computations of Standard Clustering Comparison Measures*, 2020, R package version 1.0.0.

[40] Y. Yan, *MLmetrics: Machine Learning Evaluation Metrics*, 2016, R package version 1.1.1. [Online]. Available: https://CRAN.R-project.org/package=MLmetrics

[41] V. Lyubchich and Y. R. Gel, *funtimes: Functions for Time Series Analysis*, 2020, R package version 7.0. [Online]. Available: https://CRAN.R-project.org/package=funtimes

[42] S. Darmanis, S. A. Sloan, Y. Zhang, M. Enge, C. Caneda, L. M. Shuer, M. G. H. Gephart, B. A. Barres, and S. R. Quake, "A survey of human brain transcriptome diversity at the single cell level," *Proceedings of the National Academy of Sciences*, vol. 112, no. 23, pp. 7285–7290, 2015.

[43] X. Fan, X. Zhang, X. Wu, H. Guo, Y. Hu, F. Tang, and Y. Huang, "Single-cell rna-seq transcriptome analysis of linear and circular rnas in mouse preimplantation embryos," *Genome Biology*, vol. 16, no. 1, pp. 1–17, 2015.

[44] M. Goolam, A. Scialdone, S. J. Graham, I. C. Macaulay, A. Jedrusik, A. Hupalowska, T. Voet, J. C. Marioni, and M. Zernicka-Goetz, "Heterogeneity in oct4 and sox2 targets biases cell fate in 4-cell mouse embryos," *Cell*, vol. 165, no. 1, pp. 61–74, 2016.

[45] A. A. Kolodziejczyk, J. K. Kim, J. C. Tsang, T. Ilicic, J. Henriksson, K. N. Natarajan, A. C. Tuck, X. Gao, M. Bühler, P. Liu *et al.*, "Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation," *Cell Stem Cell*, vol. 17, no. 4, pp. 471–485, 2015.

[46] B. Treutlein, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, and S. R. Quake, "Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq," *Nature*, vol. 509, no. 7500, pp. 371–375, 2014.

[47] A. A. Pollen, T. J. Nowakowski, J. Shuga, X. Wang, A. A. Leyrat, J. H. Lui, N. Li, L. Szpankowski, B. Fowler, P. Chen *et al.*, "Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex," *Nature biotechnology*, vol. 32, no. 10, pp. 1053–1058, 2014.

[48] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan *et al.*, "Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells," *Nature Structural & Molecular Biology*, vol. 20, no. 9, p. 1131, 2013.

[49] Q. Deng, D. Ramsköld, B. Reinius, and R. Sandberg, "Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells," *Science*, vol. 343, no. 6167, pp. 193–196, 2014.

Proceedings of Machine Learning Research 103:216–224, 2019 ISIPTA 2019

Constructing Simulation Data with Dependence Structure for Unreliable Single-Cell RNA-Sequencing Data Using Copulas

Cornelia Fuetterer CORNELIA.FUETTERER@STAT.UNI-MUENCHEN.DE
Georg Schollmeyer GEORG.SCHOLLMEYER@STAT.UNI-MUENCHEN.DE
Thomas Augustin THOMAS.AUGUSTIN@STAT.UNI-MUENCHEN.DE
Institut für Statistik, Ludwig-Maximilians Universität München (LMU), Munich, Germany

Abstract

Simulation studies are becoming increasingly important for the evaluation of complex statistical methods. They tend to represent idealized situations. With our framework, which incorporates dependency structures using copulas, we propose multidimensional simulation data with marginals based on different degrees of heterogeneity, which are built on different ranges of distribution parameters of a zero-inflated negative binomial distribution. The obtained higher and lower variation of the simulation data allows to create lower and upper distribution functions lead to simulation data containing extreme points for each observation. Our approach aims at being closer to reality by considering data distortion. It is an approach of examining classification quality in case of measurement distortions in gene expression data and might propose specific instructions of calibrating measuring instruments.

Keywords: Simulation studies, Copula, Imprecise probabilities, Lower and upper distribution function, Distorted measurements, Classification, Single-cell RNA-sequencing data, Statistical genetics

1. Introduction

In the context of gene expression there are up to 30% of measurements with missing data, as Yang et al. [16] indicate. This phenomenon can be traced back to the failure of measuring low read counts and the stochastic nature of gene expression. But it is not only known that gene expression in the lower range of the count data is difficult to measure. Another property of the sequencing procedure, which is the process of measuring gene expression, is that the upper sequencing range of the gene expression is also more sensitive to outliers. Therefore, measurements of gene expression do not always reflect reality which justifies the motivation of incorporating distortion of measuring tendencially higher and lower values into simulation data.

In this paper we show how the extent of different degrees of heterogeneity as well as distorted measurements with and without dependence structure affect the quality of a typical procedure in single-cell genetics concerning the classification of two subpopulations.

Thus, we will create three different scenarios for each subpopulation which represent a homogeneous and a heterogeneous population as well as a mixture of both. The homogeneous population will be constructed containing the smallest range of possible gene expression, whereas the heterogeneous population allows for a higher variability of possible gene expression. The mixture of both populations allows values with a range lying in between these populations. The pointwise lower and upper distribution functions were formed over the simulation data of the three scenarios for each target group. These are inspired by imprecise probability theory and should express the situations that compared to the real data situation, higher and lower ribonucleic acid (RNA) values were measured during the sequencing procedure.

Each of the created simulation situations based on the three scenarios as well as the distorted data will be analysed assuming that the genes are independent of each other, but also assuming the same dependence structure as the one given by the scRNA-sequencing data set provided by the authors Kolodziejczyk et al. [6]. The generation of simulation data allows keeping the dependence structure between genes as well as the marginal distributions. For the choice of the marginal distribution we decided to use the zero-inflated negative binomial distribution (ZINB) as it approximates best the measurement of gene expression in the context of single-cells (= read counts) [see 15]. If the dependence structure was not taken into account but simulated under independence, these high-dimensional data would lead to dependence structures of individual genes that cannot be controlled. This might have an influence on the classification results. With our approach it can be ensured that each of both target groups have the same dependence structures between the individual genes as in the used real data. This approach allows to set the focus explicitly on the simulated values. Thus, it is possible to examine the influence of distorted measurements in detail.

For each simulation study with and without dependence structure containing different numbers of genes we want to evaluate the classification quality that a single-cell is correctly assigned to the respective subpopulation. This is done by taking the adjusted Rand index (ARI) [see e.g. 13], which is equal to 1 when the classification perfectly corresponds to the given single-cell populations and 0 in case of random assignment.

The paper is organized as follows. In Section 2, the construction of simulation data reflecting the different degrees of heterogeneity based on the marginal distributions of ZINB are described. Section 3 describes how we use the theory of lower and upper distribution functions to generate distorted data reflecting more or less reliable data based on the scenarios presented in Section 2. Taking the dependence structure of genes into account in the simulation data by using copulas can be found in Section 4, which also contains the notation and theory of copulas. The results of the final simulation data are summarized in Section 5, followed by the conclusion, discussion and outlook in Section 6.

All the conducted steps presented below are based on appropriate packages of the R program (version 3.5.1) or were implemented in R by the first author.

2. Situations Reflecting Different Degrees of Heterogeneity

The aim of this section is to determine the influence of unreliable meausurements on the classification quality in the view of two subpopulations. We introduce a new framework of creating simulation data by defining three different scenarios for each subpopulation respectively, representing a homogeneous (Scenario 1) and a heterogeneous population (Scenario 3) as well as a transition scenario of those (Scenario 2). This leads total to three simulation data (Scenario 1, Scenario 2, Scenario 3) containing two subpopulations $n^{(1)}$= 250 and $n^{(2)}$= 250.

2.1. Use of Reference Data for Different Degrees of Heterogeneity

The original single-cell data set of Kolodziejczyk et al. [6] that was used as reference contains 295 single-cells of single-cell population 1 and 250 single-cells of single-cell population 2. Based on the gene expression of each of these subpopulations, the target groups of the simulation data were constructed. The sample size was chosen close to the publicly available, real single-cell RNA-seq data set of Kolodziejczyk et al. [6] to represent realistic scenarios in our simulations. The simulation data were also inspired by the quantiles of the estimated parameters of the original genes following a zero-inflated negative binomial distribution for the underlying structure of our scenarios.

The choice of the zero-inflated negative binomial distribution is based on recent research that states that the marginal distribution of gene expression can be approximated best by the zero-inflated negative binomial distribution following Wagner et al. [15]. Therefore, the parameters describing a zero-inflated negative binomial distribution were respectively estimated from the real data based on the single-cells belonging to each of the two single-cell populations. The zero-inflated negative binomial distribution is a mixture of a point mass at zero and the negative binomial distribution as count distribution. This allows an inflation of observing a zero read count, which is represented by the first summand. The second summand stands for the negative binomial distribution, e.g. Kleiber and Zeileis [5], [17]:

$$f_{ZINB}(X_j = x) = \begin{cases} \pi_j + (1-\pi_j) f_{NB}(0) & \text{if } x = 0 \\ (1-\pi_j) f_{NB}(x) & \text{if } x \in \mathbb{N} \end{cases}$$

with

X_j:	Random variable describing the counts of the j-th gene (j = 1, ..., m)
π_j:	Weight of the zero-inflation
x:	Observed read count
μ:	Mean
ϕ:	Shape parameter

For the generation of the simulation data, a generalization of the negative binomial distribution was used which is a mixture of Poisson distributions with a gamma distributed Poisson rate. The corresponding probability density function is the following:

$$f_{NB}(x) = f(x|\mu,\phi) = \frac{\Gamma(x+\phi)}{\Gamma(\phi)\cdot x!} \cdot \frac{\mu^x \cdot \phi^\phi}{(\mu+\phi)^{x+\phi}}$$

This generalization of the negative binomial distribution allows ϕ to be continuous. In the implementation we use the parameters $\mu \in \mathbb{R}^+$, describing the expectation of the negative binomial distribution and its dispersion parameter $\phi \in \mathbb{R}^+$. The parameter π will describe the fraction of zero-inflation as introduced above.

For our simulation data, we focused on genes that follow a zero-inflated-negative binomial distribution in both subpopulation 1 and subpopulation 2. We excluded genes with a proportion of 80 % or more zeros and with read counts never exceeding the value 2 over all measured single-cells. Genes not having a zero-inflation of their measurements are fitted to a negative binomial distribution. Applying these calculations to the originally 30 200

available genes, 26 856 genes are in compliance with these criteria, which leads to 26 856 estimates of the parameter vector for the negative binomial or zero-inflated negative binomial distribution per target group using the R package *emdbook* [1]. The construction of this simulation study is based on all the 7225 genes that fulfilled the criteria above following a zero-inflated negative binomial distribution in both subpopulations of the reference data.

2.2. Undistorted Simulation Data

In order to simulate from an imprecise setting we consider different scenarios with different interval widths, which are determined by the different parameter intervals of μ,ϕ and π for each scenario in target group 1 (Group 1) and target group 2 (Group 2).

The simulation design based on the quantiles of the estimated parameters of the 7225 genes will generate simulation data that are ZINB distributed. Scenario 1 describes the most homogeneous scenario, which is the reason for the determination of the narrowest parameter interval which leads to the smallest difference in the range of values in the subsequent sampling process. Accordingly, Scenario 3 is constructed as the broadest parameter interval, since it is intended to represent the most heterogeneous scenario. The transition Scenario 2 lies in between Scenario 1 and Scenario 2. As shown in Table 1 the difference in quantiles for both target groups increases for each scenario of parameter μ (Sc. 1: 45%, Sc. 2: 60%, Sc. 3: 70%) as well as for ϕ and π (Sc. 1: 10%, Sc. 2: 20%, Sc. 3: 30%).

	μ		ϕ	π
Sc.	Group 1	Group 2	Group 1, Group 2	Group 1, Group 2
1	[35%-80%]	[15%-60%]	[45%-55%]	[45%-55%]
2	[25%-85%]	[10%-70%]	[40%-60%]	[40%-60%]
3	[20%-90%]	[5%-75%]	[35%-65%]	[35%-65%]

Table 1: Quantiles of the estimated ZINB parameters of the reference data that are used for the construction for each scenario of target group 1 and target group 2.

Based on simulation studies we investigated the influence of the different parameters towards clustering quality and came to the result that the parameter μ has the highest influence on the clustering quality, which was the reason for allowing a broader range for Scenario 1-3. This means more variation for this parameter during the sampling process as well as a higher range of Scenario 2 and 3 compared to the remaining parameters. In order to facilitate the detection of a difference between the two target groups based on a lower number of genes ($m = 50, 100, 500$) as in the real setting, target group 2 was constructed with lower values as target group 1. The remaining parameters were based on the same quantiles for each target group as they do not play a decisive role with regard to the classification result.

Based on the determined quantile ranges of the parameters μ, ϕ and π, we construct the corresponding parameter intervals from the reference data for group 1 (see values Table 2) and group 2 (Table 3):

Sc.	μ_1	ϕ_1	π_1
1	[45, 293]	[0.27, 0.47]	[5.30*10^{-7}, 0.01]
2	[27, 397]	[0.24, 0.55]	[3.65*10^{-7}, 0.04]
3	[19, 576]	[0.18, 0.78]	[2.28*10^{-7}, 0.08]

Table 2: Constructed intervals of the ZINB parameters of each scenario describing group 1.

Sc.	μ_2	ϕ_2	π_2
1	[12, 112]	[0.27, 0.47]	[4.85*10^{-7}, 2.11*10^{-5}]
2	[6, 171]	[0.23, 0.55]	[3.26*10^{-7}, 2.91*10^{-2}]
3	[2, 217]	[0.17, 0.82]	[2.18*10^{-7}, 6.11*10^{-2}]

Table 3: Constructed intervals of the ZINB parameters of each scenario describing group 2.

For both subpopulations, the parameters describing the marginal distribution (ZINB) of each gene for target group 1 and group 2 are obtained by drawing out of the possible ranges for each parameter, assuming a discrete uniform distribution. The described procedure (see Table 2 and Table 3) is conducted for each of the three scenarios.

This leads to parameter set for group 1:

$$\theta^{(1)} = \{\mu_1^{(1)}, \phi_1^{(1)}, \pi_1^{(1)}, \mu_2^{(1)}, \phi_2^{(1)}, \pi_2^{(1)}, \mu_3^{(1)}, \phi_3^{(1)}, \pi_3^{(1)}\},$$

and equivalent for group 2:

$$\theta^{(2)} = \{\mu_1^{(2)}, \phi_1^{(2)}, \pi_1^{(2)}, \mu_2^{(2)}, \phi_2^{(2)}, \pi_2^{(2)}, \mu_3^{(2)}, \phi_3^{(2)}, \pi_3^{(2)}\}.$$

Based on the m sampled parameters $\theta_l^{(1)}$ for each scenario l of target group 1 and $\theta_l^{(2)}$ for target group 2, the simulation data are constructed by generating $n_1 = 250$ and $n_2 = 250$ random numbers out of a zero-inflated negative binomial distribution for m genes. As a final step, the individual subgroups are joined such that simulation data with the dimension ($(n_1 + n_2)$ x m) are created. This represents the situation of "No dependence structure" of the undistorted simulation data.

3. Constructing Distorted Data via Lower and Upper Distribution Functions

In this subsection the simulation data with distortion built on the constructed scenarios will be presented. These upwardly and downwardly distorted data are based on the gene-wise lower ($\underline{F_j}^{(g)}$) and upper ($\overline{F_j}^{(g)}$) distribution functions according to Montes et al. [7] for each target group g ($g = 1,2$). Therefore, we derive functions $\underline{F_j}^{(g)}$, $\overline{F_j}^{(g)}$: $\mathbb{R} \to [0,1]$, by

$$\underline{F_j}^{(g)}(x) = inf\{F_j^{(g)}(x) : F_j^{(g)} \in \mathscr{F}_j^{(g)}\},$$
$$\overline{F_j}^{(g)}(x) = sup\{F_j^{(g)}(x) : F_j^{(g)} \in \mathscr{F}_j^{(g)}\}.$$

The set of possible distribution functions of each gene of each target group ($\mathscr{F}_j^{(g)}$) is limited to the three different scenarios.

We will investigate simulation data being biased upwards as well as being biased downwards. Therefore, we determine $\hat{\overline{F}}_j$ and $\hat{\underline{F}}_j$ on the read counts x of the gene-wise upper and lower estimated distribution functions for each single-cell of the constructed simulation data set representing the different scenarios l for group g

$$\hat{\underline{F}}_j^{(g)}(x) = \inf_{l=1,2,3} \hat{F}_j^{(g)}(x \mid \theta_l^{(g)}),$$
$$\hat{\overline{F}}_j^{(g)}(x) = \sup_{l=1,2,3} \hat{F}_j^{(g)}(x \mid \theta_l^{(g)})$$

and consider the concatenation of the determined gene expression of all the single-cells over all m genes as distorted data.

This means, that in contrast to the classical imprecise probability definition of considering the set of all possible distribution functions constituting the lower and upper distribution function, we take the infimum and supremum distribution value of each single-cell for each gene over the three constructed scenarios. This means that the distorted data are generated according to the lower and upper distribution functions. This approach leads to gene-wise distribution functions that are no longer distributed to ZINB. The intention behind the construction of these distorted data is that we want to analyse the effects on the quality of clustering in case we obtained tendencially decreased read counts with the measuring instrument or increased read counts. It will be investigated how the distribution of these biased read counts is changed by taking the upper and lower distribution function. This is illustrated for target population 1 in Figure 1 and target population 2 in Figure 2 using the cumulative distribution function.

The lower distribution function (blue) reflects the situation of read counts being biased upwards for fictional gene 3. Given the instrument has a tendency to measure smaller values is represented by the upper distribution function (red) in the following two figures:

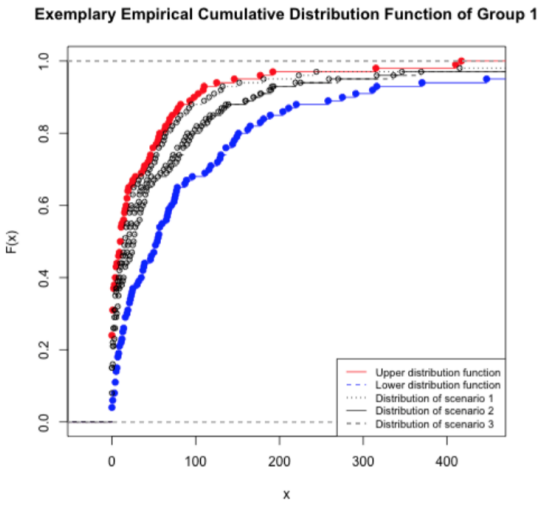

Figure 1: Lower and upper cumulative distribution function of simulated gene 3 for group 1 using the statistical software R [9].

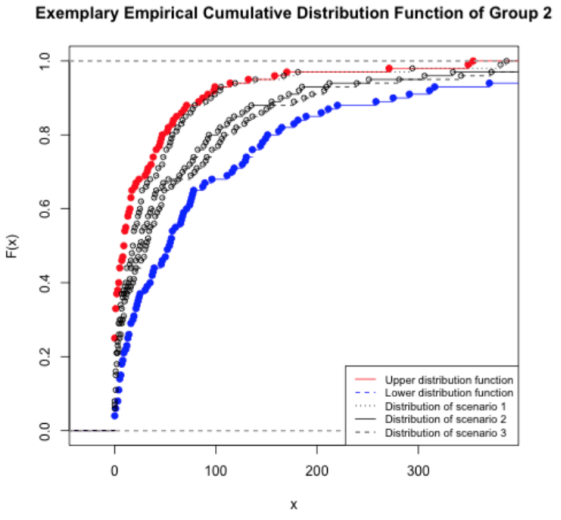

Figure 2: Lower and upper cumulative distribution function of simulated gene 3 for group 2 using the statistical software R [9].

Applying the described procedure of lower and upper distribution functions on m genes and combining them as described in the case of the undistorted data leads to the distorted simulation data with "No dependence structure".

With regard to these distortions in both directions, we will later analyse the classification results without dependence structure and with dependence structure. This brings us to the last extension of our simulation data, described in the next subsection of taking the dependence of genes into account.

4. Dependence Structure Using Copulas

Since the marginals in gene expression data have already been studied quite well and the dependence structure can be estimated on the basis of real data sets, the use of copulas for the construction of our simulation data is justified. Thus, the idea using copulas in a gene-based context in our simulation data leads to a construction of generating univariate marginal distributions F_j^g for each gene j keeping the underlying univariate marginal distributions F_j as well as keeping the same dependence structure as in the real data set for both target groups. Based on this motivation, the principle of copulas will be introduced in the first step based on the distribution function for two genes of group g. The described application will be extended towards distorted measurements, but first of all we would like to briefly recall on the concept of copulas.

Given a function C fulfills the following aspects (1)-(3) and allows a mapping of [0,1] x [0,1] $\rightarrow$ [0,1], then C can be well described as a copula, e.g. Nelsen [see 8] :

(1) $C(F_1^{(g)}, F_2^{(g)}) = C(0, F_2^{(g)}) = 0, \quad \forall\, F_1^{(g)}, F_2^{(g)} \in [0,1]$
(2) $C(F_1^{(g)},1) = F_1$ and $C(1,F_2^{(g)}) = F_2^{(g)}\ \forall\, F_1^{(g)}, F_2^{(g)} \in [0,1]$
(3) $C(F_1^{(g)}(x_2), F_2^{(g)}(x_2))$-$C(F_2^{(g)}(x_2), F_2^{(g)}(x_1)) -$
$C(F_1^{(g)}(x_1), F_2^{(g)}(x_2)) + C(F_1^{(g)}(x_1), F_2^{(g)}(x_1)) \geq 0,$
$\forall F_1^{(g)}(x_1) \leq F_1^{(g)}(x_2), F_2^{(g)}(x_1) \leq F_2^{(g)}(x_2)$

In order to obtain the joint distribution function $F_{\mathbf{X}}^{(g)}(x_1,...,x_m)$ in higher dimensions m for one target group, one can construct a copula function over all marginal distributions. Sklar [12] states that one can find a copula function of family v over all marginal distributions, which leads to the joint distribution function, that keeps the univariate marginal distributions:

$$F_{\mathbf{X}}^{(g)}(x_1,...,x_m) = C_v(F_1^{(g)}(x_1), F_2^{(g)}(x_2), ..., F_m^{(g)}(x_m))$$

This theorem will be later used for the creation of undistorted datasets respecting the dependence structure.

With the introduction of copulas it is possible to consider non-linear dependence structures [see 8]. Based on the fact that gene expression below a certain limit cannot be measured during the sequencing procedure, it is assumed, that genes tend to have a higher correlation in the low value range. There might also be a dependence in the higher value range as genes can contain outliers and extreme single-cells might tend to have genes with extremely high gene expression.

For example, it is possible that the Pearson correlation in the data is very low, but if one takes a closer look at a scatter plot of two genes, it could show a high dependence structure, as it is the case with the reference data. This observation can be explained by an underlying non-linear dependence in the data, which is considered using copulas.

4.1. Use of Reference Data for Dependence Structure

For the construction of the dependence structure using a copula, we assume the dependencies of m genes from the original count data of Kolodziejczyk et al. [6] as true. The built copula represents the joint distribution of the originally observed m genes and remains fixed for each simulation study (with fixed m). The dependence structure obtained by the real data, is based on both single-cell populations in order to prevent group specific effects.

With the use of the *VineCopula* R package of Schepsmeier et al. [10], the structure is generated by the R-vine tree which is maximized over the edges of the spanning tree with regard to the empirical Kendall's tau $\hat{\tau}_{ij}$:

$$\max \sum_{edges\ e_{ij} \in \text{ spanning tree}} | \hat{\tau}_{ij} |,$$

with a spanning tree as a tree which is based on all nodes.

In each simulation data set, the allowed copula families of constructing the tree are based only on the specified copula family for each target group using the same genes in the original data for both subpopulations. The structure selection algorithm of Dissmann et al. [2] constructs all possible pairwise copulas of the given copula family and chooses those parameters which correspond to the maximum likelihood estimation.

4.2. Simulation Data With Dependence Structure

For each simulation study, the situation of assuming the genes to be independent will be defined as "No dependence structure". With the use of the terms "Gaussian Cop", "Clayton Cop" and "Frank Cop", we designate the simulation data keeping the same marginals like in the "No

dependence structure" setting and sample out of the built copulas for respecting the same data structure using the Gaussian copula, the Clayton copula and the Frank copula.

The application of each copula with the defined dependence structure for each scenario as well as for the constructed distorted data sets, generates a common distribution function. For each of the scenarios one can generate the simulation data by applying the quantile function with the sampled parameters for each gene. In the case of the distorted data, we do not have the parametric marginals anymore as they are no longer zero-inflated negative binomially distributed. So we computed in accordance to the upper and lower cumulative distribution function, the lower and upper quantile function in order to sample from the joint distribution, keeping the same marginals.

In addition to the classical construction of copulas introduced above, the copulas will also be used for undistorted datasets, actually for downwardly distorted count data and for upwardly distorted count data. Following the fact, that $\underline{F_j}^{(g)}$ and $\overline{F_j}^{(g)}$ are again cumulative distribution functions, allows to determine the joint distribution over all m genes by using the following copula construction of family v [see 7, 14]:

$$C_v(\underline{F_1}^{(g)}, ..., \underline{F_m}^{(g)}) \text{ and } C_v(\overline{F_1}^{(g)}, ..., \overline{F_m}^{(g)})$$

5. Results

This lead to the final simulation data with and without dependence structure for distorted and undistorted data and for different numbers of genes m. Each of these combinations was analyzed on the basis of 50, 100 and 500 genes. All the simulation studies contained 500 single-cells with 250 single-cells representing each target group. For all simulation studies, we first classified the gene expression assuming there is no dependence structure between the genes. In addition, we studied the influence of different copulas (Gaussian, Clayton and Frank copula) fitted to the same original count data, given the same number of genes. Taking the same dependence structure over each target group as in the reference data, allowed a better comparison of the simulation studies as we focused on the marginal distributions and decided to keep the fitted structure fixed over each simulation design. This applied not only to the distorted data, but also to each scenario.

Before presenting the classification results, we want to point out the intention behind the construction of the different simulation datasets once again. The simulation data of each scenario represents different ranges of possible read counts. Scenario 1 allows the smallest range of parameters for the ZINB distribution and therefore represents the most homogeneous scenario. Scenario 3 contains the broadest range of possible parameters and therefore reflects the most heterogeneous data situation of all the scenarios, containing also the most homogeneous scenario (Scenario 1). As the range of the parameters for Scenario 2 lies in between the one of Scenario 1 and 3, one can state that Scenario 2 is a transition scenario from homogeneous to heterogeneous. The simulation data set which was created by the lower distribution function represents the data set situation of measuring tendencially higher read counts. With the construction of the upper distribution function, one aims to reconstruct read counts that are tendencially biased downwards.

In the following, a k-means clustering of the *mclust* R package of Scrucca et al. [11] is performed creating two clusters with and without using the dependence structures of the Gaussian, Clayton and Frank Copula. For evaluating the clustering quality, the adjusted Rand index is applied, which is also implemented in the R package *mclust*. In accordance to the undistorted data, the assumption that the single-cells of different target groups are independently distributed is still valid for distorted data. Therefore it does not cause any problem to simply merge the data sets constructed for each target group to obtain a whole data set containing both subpopulations for each simulation data set.

5.1. Results of the Undistorted Data

Based on the construction of the undistorted data, which are represented by the three scenarios, one can assume that detecting the different subpopulations might be easier in the third scenario compared to the second and first scenario. This assumption can be confirmed in the case of the independent settings for 50, 100 and 500 genes with regard to the adjusted Rand Index, which is displayed in Table 4 , 5, and 6. In case of considering dependence structures in the simulation data, this statement is only valid for the simulation data of all investigated numbers of genes using the Gaussian copula and for the Clayton copula in the dimension of using 500 genes. All in all, one can state that in the lowest dimension, the Gaussian copula performs best for scenarios tending to be more heterogeneous. In case of a very homogeneous data situation it seems as if the choice of the Frank copula was the best. With 100 and 500 genes, the Frank copula performs best in every scenario.

	Scenario 1	Scenario 2	Scenario 3
No dependence structure	0.32	0.49	0.55
Gaussian Cop	0.46	0.53	0.63
Clayton Cop	0. 42	0.41	0.38
Frank Cop	0. 60	0.47	0.53

Table 4: ARI for the simulation data for n_1= 250, n_2= 250, m= 50 (Simulation study 1 of undistorted data only).

	Scenario 1	Scenario 2	Scenario 3
No dependence structure	0.52	0.71	0.87
Gaussian Cop	0.68	0.70	0.70
Clayton Cop	0. 42	0.41	0.38
Frank Cop	0. 92	0.80	0.91

Table 5: ARI for the simulation data for n_1= 250, n_2= 250, m= 100 (Simulation study 2 of undistorted data only).

	Scenario 1	Scenario 2	Scenario 3
No dependence structure	0.65	0.85	0.98
Gaussian Cop	0.88	0.88	0.93
Clayton Cop	0.49	0.49	0.51
Frank Cop	1	1	0.99

Table 6: ARI for the simulation data for n_1= 250, n_2= 250, m= 500 (Simulation study 3 of undistorted data only).

To conclude at this stage, one has to pay attention to the choice of the right copula. Especially in the case of simulation data, one should not create independent simulation data as a simplification of reality. One should rather pay attention to the right choice of copulas which can achieve better results compared to an independence structure.

5.2. Results of the Distorted Data

In the following, we describe the classification results of the distorted data, which can be found in Table 7, Table 8, and Table 9. The clustering performance of the distorted data was always better in case of using the lower distribution function compared to the upper distribution function in the setting of an independence structure as well as in the setting of the Gaussian, Clayton and Frank copula. In addition, one can state that with the use of the lower distribution functions the clustering performance gets better with an increase of the dimension. The only exception is the clustering performance of the Frank copula using 100 genes instead of 50 genes, which leads to a decrease of the adjusted Rand index from 0.63 to 0.61. In case of the lower distribution function, the Clayton copula performs the worst. Choosing the best performance in using 50 and 100 genes, one obtains the best classification result using an independence structure. In the highest dimension of 500 genes, the Frank copula performs best.

	Lower Distribution	Upper Distribution
No dependence structure	0.80	0.14
Gaussian Cop	0.49	0.41
Clayton Cop	0.29	0.24
Frank Cop	0.63	0.20

Table 7: ARI for the simulation data for n_1= 250, n_2= 250, m= 50 (Simulation study 1 of distorted and undistorted data).

	Lower Distribution	Upper Distribution
No dependence structure	0.90	0.18
Gaussian Cop	0.49	0.35
Clayton Cop	0.34	0.24
Frank Cop	0.61	0.17

Table 8: ARI for the simulation data for n_1= 250, n_2= 250, m= 100 (Simulation study 2 of distorted and undistorted data).

	Lower Distribution	Upper Distribution
No dependence structure	0.93	0.30
Gaussian Cop	0.75	0.27
Clayton Cop	0.47	0.14
Frank Cop	0.97	0.01

Table 9: ARI for the simulation data for n_1= 250, n_2= 250, m= 500 (Simulation study 3 of distorted and undistorted data).

The performance of clustering having tendencially lower read counts is not going to be interpreted because the results are quite bad and can almost be compared to a random assignment of observations to the target groups.

6. Conclusions, Discussion and Outlook

6.1. Conclusions

With the construction of the upwards and downwards distorted data of the three scenarios it was possible to generate distorted simulation data. The values of the upper distribution functions reflect the situations containing lower gene expression, whereas the lower distribution functions contain upper distortions of the simulated values of each scenario. Due to the fact that only positive values (including zero) can be generated out of the ZINB means that the deviations in the upper measuring range can vary distinctively more than in the lower range of values. In connection with the measured gene expression, the

immense outliers are often also addressed in the analysis of real scRNA-seq data sets. This is another indication that the simulation data might represent well the real data situation of single-cells.

One can state for the classical simulation studies that choosing the right copula can improve the clustering performance. Specifying the effect of different copulas on distorted data requires further analysis.

The phenomenon of the natural ranges of the lower distribution and upper distribution simulation data might be the explanation for the bad performance of the simulation data of the upper distribution. This leads to the conclusion that in the extreme case of measuring always the highest value one allows higher variation of gene expression which leads to an easier distinction of the target groups. Whereas in the case of measuring tendentially always the lowest value only brings little variation of gene expression and leads to less adequate classification results. In accordance to this statement, we have seen that the clustering performance tends to be better, the more heterogeneous the data are. We can further conclude that the clustering behaviour of the undistorted data improves, the more genes are used. This fact can also be observed in the case of using lower distributions but that does not apply to the distortion based on the upper distributions for the reasons already mentioned.

The proposed approach has been a first step to provide simulations showing consequences of distorted measurements towards the ability of assigning single cells to the right group membership. The approach has been designed to represent the extreme cases of distorted data. For a more in depth investigation into each direction of distortion it might be appropriate to continue developing tools of determining distortion based on well defined scenarios.

6.2. Discussion and Outlook

With the decision of creating simulation data based on quantiles, we set the focus on genes with a tendency of a homogeneous gene structure without outliers since imprecise measurement might play a higher role in these situations. Therefore, the range of obtained results might nicely reflect the imprecision of the real measurements of gene expression. In case of using the lower (upper) distribution function, the tendency of measuring always higher (lower) gene expression than the real one, might reflect the measurement error of an instrument that has the tendency of measuring higher (lower) gene expression.

The construction of distorted simulation data might nicely correspond to the idea that the measured gene expression can be distorted into both directions. Especially the case of having strong outliers can have a high impact on the classification result. With our simulation studies, we investigated the clustering behaviour based on maximal 500 genes, but in reality there are several thousands of genes to analyse. Choosing the lower and upper distribution function, constructed by the infimum and supremum of different distribution functions, might not be a valid choice in a higher dimension setting anymore. Given we would generate the lower and upper distribution functions in even higher-dimensional settings and given we still have the three defined scenarios, then the proportion of those read counts, which are located at the respective boundaries of the value range, would increase. Thus, the final clustering would take place increasingly on read counts with very little gene expression or on genes with very strong outliers, depending on the construction of the respective scenarios.

Further research should focus more on the role of lower and upper distribution functions in the context of p-boxes [see 3, 4], describing a whole set of scenarios and on decision procedures relying on the whole induced credal set. Thus, for a future project, it would be interesting how a construction of a less clear scenario would affect the clustering performance. Another point that could be discussed, is how to improve the sampling procedure underlying the simulation, in order to use simulations closer to the idea of truly interval-valued probability, but this is a general topic that clearly goes far beyond the scope of this paper.

Regarding the dependence structure, one could further determine the influence of the used copula families using vine copulas, especially in a distorted setting. As a further step, it would also be of interest to look at the defined scenarios with the help of imprecise copulas [see 7].

Concerning the application of the obtained results, one imaginable conclusion of this simulation study would be whether it might be worth to calibrate measuring instruments further down or being more precise in the higher value range of count data. As extreme outliers often occur during the measurement of single-cell RNA gene data, it is not a surprise, that this tends to have an impact on the clustering result. Our tool might help to analyze the consequences of distorted measurements and might help to give assessments of how distorted measurements could affect the quality of the classification result. In addition, with a more precise investigation of the impact of outliers on the classification results it can be studied whether these outliers are useful for classification or not.

In accordance with our classification results, measuring a tendency of lower read counts than reality does not result in worse clustering performance at least in a low dimensional

context. So, the current state-of-the-art, which tends to miss low read counts, has a lower impact than misspecifying high read counts. Based on our new findings, we question the current approach of calibrating measuring instruments in the low sequencing ranges and demand further analyses that also take distortions in the higher measuring range into account.

Acknowledgments

We would like to thank the Hemberg Group of the Sanger Institute for providing the reference data publicly available. We also appreciate a lot the LMUMentoring program, supporting young researchers by providing financial support for the first and second author to travel to this conference. Last but not least, we are very grateful and want to thank the three anonymous referees for their stimulating comments.

References

[1] Ben Bolker. *emdbook: Ecological Models and Data in R*, 2019. R package version 1.3.11.

[2] Jeffrey Dissmann, Eike Christian Brechmann, Claudia Czado, and Dorota Kurowicka. Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics and Data Analysis*, 59 (1):52–69, 2013.

[3] Scott Ferson and W. Troy Tucker. Sensitivity analysis using probability bounding. *Reliability Engineering and System Safety*, 91:1435–1442, 2006.

[4] Scott Ferson, Vladik Kreinovich, Lev Ginzburg, Davis Myers, and Kari Sentz. Constructing probability boxes and Dempster-Shafer structures. *Sandia National Laboratories Technical Reports*, SAND2002-4015, 2003. URL https://digital.library.unt.edu/ark:/67531/metadc737049/. last access: 2019-05-15.

[5] Christian Kleiber and Achim Zeileis. Visualizing count data regressions using rootograms. *The American Statistician*, 70(3):296–303, 2016.

[6] Aleksandra A. Kolodziejczyk, Jong Kyoung Kim, Jason C.H. Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N. Natarajan, Alex C. Tuck, Xuefei Gao, Marc Bühler, Pentao Liu, John C. Marioni, and Sarah A. Teichmann. Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, 17:471–85, 2015.

[7] Ignacio Montes, Enrique Miranda, Renato Pelessoni, and Paolo Vicig. Sklar's theorem in an imprecise setting. *Fuzzy Sets and Systems*, 278:48–66, 2015.

[8] Roger B. Nelsen. *An Introduction to Copulas*. Springer, 2006.

[9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org.

[10] Ulf Schepsmeier, Jakob Stoeber, Eike Christian Brechmann, Benedikt Graeler, Thomas Nagler, Tobias Erhardt, Carlos Almeida, Aleksey Min, Claudia Czado, Mathias Hofmann, Matthias Killiches, Harry Joe, and Thibault Vatter. *VineCopula: Statistical Inference of Vine Copulas*, 2018. URL https://CRAN.R-project.org/package=VineCopula. R package version 2.1.8.

[11] Luca Scrucca, Michael Fop, Brendan T. Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):205–233, 2016.

[12] Abe Sklar. Fonctions de Répartition à n Dimensions Et Leurs Marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231, 1959.

[13] Nguyen Xuan Vin, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

[14] Damjan Škulj. Imprecise copulas constructed from shock models (Slides of a talk at the 11th Workshop on Principles and Methods of Statistical Inference with Interval Probability (WPMSIIP)), 2018. URL bellman.ciencias.uniovi.es/~ssipta18/ScheduleWPMSIIP.html. last access: 2019-05-15.

[15] Günter P. Wagner, Koryu Kin, and Vincent J. Lynch. A model based criterion for gene expression calls using RNA-seq data. *Theory in Biosciences*, 132: 48–66, 2013.

[16] Mary Qu Yang, Sherman M. Weissman, Yang William, Jialing Zhang, Allon Canaann, and Renchu Guan. MISC: missing imputation for single-cell RNA sequencing data. *BMC Systems Biology*, 12: 114, 2018.

[17] Achim Zeileis, Christian Kleiber, and Simon Jackman. Regression models for count data in r. *Journal of Statistical Software*, 27 (8), 2008.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Cornelia Fütterer, Malte Nalenz and Thomas Augustin

Discriminative Power Lasso – Incorporating Discriminative Power of Genes into Regularization-Based Variable Selection

Technical Report Number 239, 2021
Department of Statistics
University of Munich

http://www.statistik.uni-muenchen.de

Discriminative Power Lasso – Incorporating Discriminative Power of Genes into Regularization-Based Variable Selection

Cornelia Fuetterer[+*], Malte Nalenz[+*], and Thomas Augustin[+]

Abstract—In precision medicine, it is known that specific genes are decisive for the development of different cell types. In drug development it is therefore of high relevance to identify biomarkers that allow to distinguish cell-subtypes that are connected to a disease. The main goal is to find a sparse set of genes that can be used for prediction. For standard classification methods the high dimensionality of gene expression data poses a severe challenge. Common approaches address this problem by excluding genes during preprocessing. As an alternative, L1-regularized regression (Lasso) can be used in order to identify the most impactful genes.
We argue to use an adaptive penalization scheme, based on the biological insight that decisive genes are expressed differently among the cell types. The differences in gene expression are measured as their *discriminitive power* (DP), which is based on the univariate compactness within classes and separation between classes. ANOVA based measures, as well as measures coming from clustering theory, are applied to construct the covariate specific DP.
The resulting model, that we call *Discriminative Power Lasso* (DP-Lasso), incorporates the DP as covariate specific penalization into the Lasso. Genes with a higher DP are penalized less heavily and have a higher chance for being part of the final model. With that the model can be guided towards more promising and trustworthy genes, while the coefficients of uninformative genes can be shrunken to zero more reliably.
We test our method on single-cell RNA-sequencing data as well as on simulated data. On average, DP-Lasso leads to significantly sparser solutions compared to competing Lasso-based regularization approaches, while it is competitive in terms of accuracy.

Keywords—Penalized Regression, Variable Selection, Clustering validation metrics, scRNA-sequencing data.

I. Introduction

In personalized medicine, it is important to identify genes, which can be used to accurately predict the individual outcomes. For the development of biomarkers, a lower number of covariates means less effort in its subsequent clinical testing. As in high-dimensional settings many genes are often noise, the challenge is to select only the covariates that are relevant in terms of prognostic, predictive or biological impact on the drug or the disease [19]. In case of non-small cell lung cancer (NSCLC), the detection of the biomarker EML4-ALK fusion gene [27] led to the development of the drug crizotinib, which is used for patients carrying an ALK-fusion. In contrast to the earlier low response, crizotinib dramatically raised the response rate in NSCLC [19].

[+]Ludwig-Maximilians-University, Munich. Department of Statistics.
[*]These authors contributed equally to this work.

In general, the transition of healthy cells into cancerous cells affects changes in gene expression that can be measured. It is therefore common practice to investigate single-cell RNA sequencing data, introduced by [30], which allows insights into the different types of single cells. In the case of a cell cycle, the cell passes from the DNA synthesis (S-phase) to the mitosis (M-phase), including the gap phases (G1 and G2) in between. These different phases can be distinguished by its measured gene expression of a synchronized cell population. For example, a high score at the G2M checkpoint can be an indicator of a metastasis tumor [21]. Testing whether genes are differentially expressed among different cell types might therefore lead to valuable insights.

From a biological point of view, it is therefore of relevance to extract a sparse set of genes that can be used to classify and characterize the subpopulations [11]. One common approach is to use penalized regression models, such as the Lasso [31] that find a trade-off between model fit and model complexity. The advantage of the Lasso is that it provides variable selection, by setting coefficients exactly to zero. An extension is the adaptive Lasso [36] which uses covariate specific penalization terms. The penalization terms are inversely proportional to the ordinary least square (OLS) estimates from a multivariate regression model.

In this article, we combine the concepts of regularized regression with the biological background of differentially expressed genes. Genes that differ univariately with respect to the target, should be penalized less heavily.
We therefore introduce the term discriminative power (DP), which allows a covariate specific evaluation of compactness and separation with regard to the outcome. Discriminative power is measured by means of clustering indices [3], as well as by the classic concept of analysis of variance (ANOVA) [12].

The discriminative power is directly incorporated into the adaptive Lasso as covariate specific penalization, resulting in our approach Discriminative Power Lasso (DP-Lasso). Using the DP as penalization weights in a L1-regularized model can be seen as a soft filtering as we do not exclude any covariates before performing regression, but favour genes with good univariate properties. The idea is to give a higher penalty to covariates with low univariate DP and a reduced penalty to the more promising covariates.

This paper is structured as follows. In Section II we introduce notations and give an overview over commonly used regularization based methods. Section III introduces the DP-Lasso model. In Section IV and Section V we test the performance of DP-Lasso on scRNA-sequencing datasets as benchmark datasets, and on simulated data. Section VI concludes and provides an outlook.

II. Methods

In supervised learning, the goal is to estimate the underlying function that maps the p-dimensional covariate space to the outcome. As training data, we are given a matrix X, composed of p covariate vectors each containing the values of the N observations. This leads to the covariate matrix $X = (x_1, \cdots, x_p), j = 1, \cdots, p$, and the vector y containing the N outcomes. x_{ij} denotes the value of observation i for covariate j, x_j the N values of covariate j, and $x_{i\cdot}$ the p dimensional observation vector for observation i. Given that the outcome is continuous, a common approach is to estimate the linear model

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \quad , \tag{1}$$

where β is the p-dimensional vector of regression coefficients. In the following, categorical outcomes $y \in \{1, \cdots, K\}$ are considered. In this case a generalized linear model (glm) is appropriate, which uses a linear structure as in Equation (1) and connects it to the target through a link function [10]. Thus, for binary outcomes $y \in \{0, 1\}$ logistic regression is used and for $K > 2$ classes the multinomial-logit model. However, for ease of notation in the following the linear model is used to describe the methods.

In high dimensional data and especially $p >> N$ generalized linear models can not be estimated reliably, due to the problems of multicollinearity and perfect separation [1, 14]. Also glms can not deal efficiently with irrelevant predictors, as no variable selection is performed. It is therefore common practice to reduce the number of genes before analysis.

For this purpose, the univariate filtering approach selects covariates based on (adjusted) p-values of univariate tests or biological reasoning. The final result highly depends on the researcher's choice, because a threshold or number of genes kept for the analysis has to be specified.

Alternatively, one can use regularized regression models, that find a trade-off between model fit and model complexity for parameter estimation. Regularized regression models also lead to more stable solutions for β coefficients in $p >> N$, as extreme behavior is penalized [15]. This allows to find a unique solution in situations where glms might fail, such as perfect separability and multicollinearity.

In regularized regression models, the overall loss function is decomposed into the discrepancy of the observed target and the model prediction and a penalty term that controls the complexity of the model. In case of the classical Lasso, the penalty is equal to the L1-norm of the coefficients β, leading to the overall loss function [31]

$$L(y, X, \beta, \lambda, w) = \underbrace{\sum_{i=1}^{N} (y_i - x_{i\cdot}\beta)^2}_{\text{SSE}} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{\text{Penalty Term}}, \tag{2}$$

for linear regression. The degree of shrinkage and sparsity is controlled by a global shrinkage parameter λ, which is usually chosen via cross-validation.

Lasso regression allows to exactly shrink coefficients to zero, which leads to a covariate selection. Lasso has efficient solvers available, making it a good choice for high dimensional datasets. However, the Lasso has the known deficiency of over-shrinkage: To remove a large number of uninformative covariates, a high penalty parameter needs to be chosen. This in return will also shrink the coefficients of informative predictors to some extent. To counteract, the Lasso will take in correlated predictors, to substitute for the over-shrinkage [35]. This makes the interpretation of covariates left in the final model somewhat dubious, as it is unclear if the covariate itself is important or just as a substitute for the overshrinkage of another covariate.

If predictive performance is the primary objective, Ridge regression (L2-penalty) is a popular alternative. L2-penalty limits the influence of individual covariates, by penalizing high β's strongly, but shrinks no coefficient exactly to zero [15].

The Elastic Net [37] uses a mixture of the L1-norm (Lasso) and the L2-norm (Ridge). The loss function of the Elastic Net can be written as

$$L(y, X, \beta, \lambda, w) = \sum_{i=1}^{N} (y_i - x_{i\cdot}\beta)^2 + \alpha \sum_{j=1}^{p} \lambda_j |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \lambda_j \beta_j^2, \tag{3}$$

where α is a mixing parameter that controls the proportion of L1 and L2-penalty that is put on the coefficients.
Elastic Net often shows better predictive performance than Lasso, while also being able to set coefficients exactly to zero.

To reduce the amount of over-shrinkage and improve variable selection consistency, the adaptive Lasso [36] was proposed. Instead of using the same global shrinkage λ on every coefficient, the adaptive Lasso uses a covariate specific shrinkage parameter λ_j, which allows a separate penalty for each covariate. This leads to the loss function of the adaptive Lasso [36]

$$L(y, X, \beta, \lambda, w) = \sum_{i=1}^{N} (y_i - x_{i\cdot}\beta)^2 + \sum_{j}^{p} \lambda_j |\beta_j|, \tag{4}$$

where $\lambda_j = \lambda w_j$ is the covariate specific penalty and w_j are discount factors that increase or decrease the amount of penalization for covariate j. In the original adaptive Lasso, w_j is calculated as the inverse of the parameter estimates of the ordinary least squares (OLS) regression, hence $w_j = 1/\hat{\beta}_j^{(OLS)}$. For this approach it can be shown that it improves the model selection consistency under certain assumptions [36]. More concretely this results in less penalization of important covariates with high $\hat{\beta}_j^{(OLS)}$, which allows the final coefficients to become large, mitigating the over-shrinkage effect. In case of $p >> N$, the covariate specific weighting can be obtained by a ridge regression instead of the OLS estimates.

Several other extensions of the Lasso have been proposed, such as the fused Lasso [32], group Lasso [20], Bayesian Lasso [22] and Bayesian shrinkage priors [2].

Another commonly used approach for gene selection is the usage of tree ensembles, such as random forests [8]. Random forests [4], that combine several decision trees, are a popular choice for genetic classification data, as they have a strong predictive performance and do not require further assumptions. Measures, such as (unbiased) variable importance [29] and SHAP values [17] can be used to assess the importance of individual covariates, to rank covariates and to identify the most impactful genes.

III. Discriminative Power Lasso

In $p >> N$ situations, in which the number of covariates exceeds the number of observations, there always exists an infinite amount of solutions for the regression hyperplane defined by the regression coefficients. While regularization helps to promote sparsity and limits extreme behavior, we argue that additional information can guide the model towards more robust and reliable solutions. In contrast to the original adaptive Lasso, we want to limit the impact of covariates that only work well in a multivariate model, but are not discriminative univariately. If enough data is available, such interplay between different covariates can be reliably estimated. However, with limited training data, the chance of over-fitting on spurious relationships is high when learning multivariate models. Therefore, we suggest to promote instead genes that decompose the data into 'natural' groups, measured by the univariate discriminative power based on the conditional distribution $f(X_j|Y), j = 1, ..., p$.

The construction of the DP can be motivated by the concept of analysis of variance that measures the impact of a grouping variable on a numeric outcome by the differences of the group means. Therefore, for the construction of the DP we use the dependent variable y as independent variable that we condition on to explain the differences in X. This change in perspective adds new information that is unavailable in a purely supervised regression approach. Secondly, cluster validation measures that have been developed in unsupervised clustering theory can be applied. Instead of using the outputted cluster labels as groups, as it is usually done in unsupervised learning, we directly use the target labels y as grouping. The discriminative power therefore measures how well a covariate decomposes the underlying groups in terms of compactness and separation.

A. Target Adaptive Regularization

We implement the preference towards covariates with high discriminative power by discounting their penalty, similar to the adaptive Lasso. The overall loss function of DP-Lasso can be written as

$$L(y, X, \beta, \lambda, w) = \mathcal{E}(\hat{y}, y, \beta) + \sum_{j=1}^{p} \lambda_j |\beta_j|, \qquad (5)$$

where $\mathcal{E}$ is an appropriate loss function measuring the deviation of the fitted response vector $\hat{y}$ form the true values y, using a suitable link function. For logistic regression deviance or log-loss are common choices for $\mathcal{E}$. In case of a linear model the model takes the form of Equation (4). We propose to choose the covariate-specific penalty as $\lambda_j^{(DP)} := \lambda w_j^{(DP)}$ and $w_j^{(DP)} = 1/DP_j$, where DP_j is the discriminative power of gene j. This gives the model a gentle push towards covariates that appear more natural and reliable, based on their DP. Note that both the calculation of DP and the following regularized regression model are based on N observations of the training data.

Combining the DP with the supervised approach enriches the regression model with new information. Covariates with high DP are more likely to be selected in the final model, whereas covariates, that only work well in a multivariate model, but have a low individual DP are more likely to be removed. The adaptive shrinkage parameter also counteracts the over-shrinkage. Coefficients of covariates that work well in the multivariate model and also appear as good candidates, based on their DP, will be penalized less heavily and will be allowed to become large. On the other hand, clearly uninformative covariates with a low DP will receive an even higher penalty and can be removed more easily in the regularization step. Lastly, if several solutions to Equation (5) are similarly good, our approach gives a gentle push towards covariates that appear more trustworthy.

B. Characterization of natural groupings

This section motivates the construction of our DP measures. In general, we assume covariates X_j as more promising for which the underlying groups y are homogeneous and well separated from the other groups . This reflects the idea that relevant genes should express differently among the K classes. Figure 1 shows the distribution of two example genes from the below used single-cell RNA-sequencing dataset EMTAB2805 of [5]. For the gene on the left side, we can see that the two underlying classes show clear differences in their distribution. Also the two groups are relatively compact and their group-means well separated. For the gene on the right side, the two groups show a stronger overlap, and they are less separated.

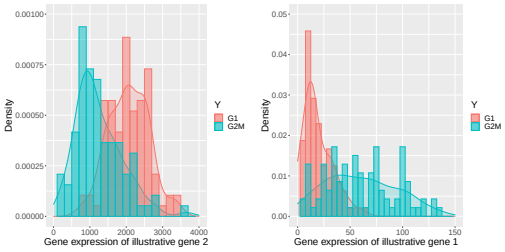

Fig. 1. Univariate distributions of two genes. The colors indicate the two groups. Left side: the two classes show clear differences in their distribution. Right side: the distributions are strongly overlapping with no clear difference.

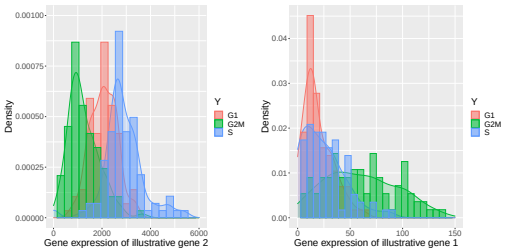

Fig. 2. Univariate distributions of two genes. The colors indicate the three groups. Left side: the three classes show clear differences in their distribution. Right side: the distributions are strongly overlapping with no clear difference.

Therefore, the gene on the left side appears to be a more natural candidate for a decisive gene and should have a higher chance of being selected. The same rationale can be used for $K > 2$. Figure 2 shows the univariate distributions for three classes on the same genes, which can be used to assess the compactness and separation.
Therefore, the idea of DP-Lasso is to prefer genes that decompose nicely into the underlying classes with regard to compactness and separation. We call this concept of 'natural grouping' the discriminative power DP. Genes with a high discriminative power will be favored in the regularization step (see Section III-A).

When using for example a logistic regression model, compactness of the groups (as an indication of naturality of the group) is not directly evaluated. The same goes for the distance between groups (or their means): As long as the groups are perfectly separable by a hyperplane, as is the case in $p >> N$, the margin to the discrimination plane is typically not considered in the loss function. Figure 3 shows two simulated covariates with a similar slope from a logistic regression model. While the two classes can be separated similarly good in both covariates, we would intuitively prefer the covariate shown at the right side, due to its distribution. Here the two classes express differently and the two groups are both compact and well separated, whereas the distribution at the left side appears more likely to be random. These descriptive illustrations aim to motivate the inclusion of additional information into the penalization by the discriminative power, which is described in the following.

The natural decomposition can be formalized by the concepts of compactness and separation with respect to the response.

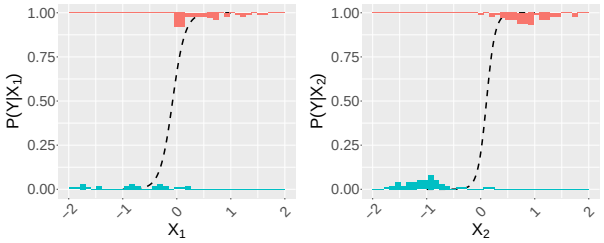

Fig. 3. The graph shows simulated genes, that can similarly well discriminated by a logistic regression. Left side: the clusters appear unnatural. Right side: compact groups with well separated group means.

C. Measures of discriminative power (DP)

In the following we describe three interesting options to measure the discriminative power. The goal is to capture information about the compactness and separation between classes in each gene. The discriminative power is therefore calculated univariately over each covariate j using the target variable y as grouping. In the following

$$x_j^{(k)} = \{x_{ij} : y_i = k\}_{i=1}^{N} \tag{6}$$

denotes the set of values of covariate j that belong to observations with the target class k, and $x_{hj}^{(k)}$ denotes the covariate values of the h'th observation in class k.
There exist a large number of quality criteria that are commonly used in unsupervised learning to evaluate clustering solutions. Also the idea of discriminative power can be interpreted as a classical test problem. The following describes three ways to measure DP, based on these principles.

1) ANOVA-approach: One classical way to test for differences in group means is the analysis of variance (ANOVA) [12]. Intuitively, the ANOVA expresses how much of the sample variance can be explained by the grouping. More concretely, the ANOVA tests whether there is a difference in the means of K groups based on its F-statistic.
Let $\bar{x}_j^{(k)} = \frac{1}{n_k} \sum_{h=1}^{n_k} x_{hj}^{(k)}$ denote the class mean of covariate j in target class k, where n_k is the number of observations belonging to class k and $\bar{x}_j$ denotes the overall mean over N observations. The according test statistic F_j measures the ratio of between-group variability and within-group variablity of covariate j via

$$F_j = \frac{(N-K)}{(K-1)} \frac{\sum_{k=1}^{K} n_k (\bar{x}_j^{(k)} - \bar{x}_j)^2}{\sum_{k=1}^{K} \sum_{h=1}^{n_k} (x_{hj}^{(k)} - \bar{x}_j^{(k)})^2}. \tag{7}$$

The value of the F-statistic is large in case that the distances between the groups are considerably higher than the distances within the groups. The higher the F-statistic, the higher the proportion of variance explained by the grouping, indicating significant differences in class means. We thus use the value of the F-statistic as one possibility for the measurement of discriminative power and determine the discount

factor $w_j^{(DP)}$ for the penalization in the subsequent step with $w_j^{(ANOVA)} = 1/F_j$. As $1/F_j$ can become quite large we use a logarithmic transform to attenuate the differences in DP between the genes and to avoid numerical instabilities.

2) Davies-Bouldin Index: The Davies-Bouldin index *DB* was developed for validating the clustering quality based on compactness and separation of the clusters [6]. As mentioned before, instead of evaluating a cluster solution, the K classes are evaluated. The DB index relates the compactness within the groups to the separation between the classes. The compactness of class k is measured by the root mean square error of observations from class k to the class mean $\bar{x}_j^{(k)}$ of class k in covariate j, leading to

$$\Delta_j^{DB}(k) = \sqrt{\frac{1}{n_k}\sum_{h=1}^{n_k}(x_{hj}^{(k)} - \bar{x}_j^{(k)})^2},$$

which in the univariate case simplifies to the standard deviation of observations in group k. The separation between the groups k and l is measured via the Euclidian distance of their respective class means $\bar{x}_j^{(k)}$ and $\bar{x}_j^{(l)}$, which in the univariate case simplifies to

$$\delta_j^{DB}(k,l) = |\bar{x}_j^{(k)} - \bar{x}_j^{(l)}|.$$

The overall DB index is then given as

$$DB_j = \frac{1}{K}\sum_{k=1}^{K}\max_{l\neq k}\left\{\frac{\Delta_j^{DB}(k) + \Delta_j^{DB}(l)}{\delta_j^{DB}(k,l)}\right\}, \qquad (8)$$

which compares each class to its closest class, as a more pessimistic measure. The better the groups are separated and compact, the lower the DB index and as a consequence this covariate should be less penalized. Therefore, the discount factor is taken as $w_j^{(DB)} = DB_j$.

3) Silhouette Index: The silhouette index S_j [24] considers the compactness and separation evaluated on the individual level. For the construction of the 'silhouette width' s_{ij} the closeness of observation i to all observations within its group $k = y_i$ is measured via

$$\Delta_j^{Sil}(i,k) = \frac{1}{(n_k - 1)}\sum_{h:y_h=k, h\neq i}|x_{ij} - x_{hj}^{(k)}|, \qquad (9)$$

which is similar to the compactness measure in the DB index. However, Δ_j^{Sil} takes the closeness to each individual observation into account, instead of measuring the deviation from the mean.
Separation between the groups is measured via,

$$\delta_j^{Sil}(i,k) = \min_{l\neq k}\left\{\frac{1}{n_l}\sum_{h=1}^{n_l}|x_{ij} - x_{hj}^{(l)}|\right\}, \qquad (10)$$

which takes the minimum average distance to the members of any other class. The silhouette width s_{ij} combines compactness and separation which leads to

$$s_{ij} = \frac{\delta_j^{Sil}(i,k) - \Delta_j^{Sil}(i,k)}{\max\{\Delta_j^{Sil}(i,k), \delta_j^{Sil}(i,k)\}}. \qquad (11)$$

As a last step, the silhouette index S_j is calculated by averaging over the silhouette width s_{ij} of all N individuals,

$$S_j = \frac{1}{N}\sum_{i=1}^{N} s_{ij} \quad \in \quad [-1,1]. \qquad (12)$$

S_j which can be used as a global measure of clustering quality given the covariate j and the target classes.
The absolute silhouette index takes values close to 1, if all observations are compact within their groups and well separated from the other groups. The more the silhouette index S_j approaches 0, the less compact the observations are within their groups and the less separated among covariate j. In this case the groupings are not nicely decomposed, and therefore this covariate is considered as less decisive.
The higher the absolute value of the silhouette index of covariate j, the better the distinction of the two underlying groups. Covariates with a high absolute silhouette index should be penalized less, therefore we set $w_j^{(Sil)} = 1/|S_j|$.

IV. Empirical Comparison

In this section we first present the scRNA-sequencing benchmark data and test the performance of DP-Lasso with different choices of the DP against competing methods. For both the binary classification, described in Section IV-C and the multiclass classification, described in Section IV-D, we perform a $5-$times repeated $10-$fold cross validation. In contrast to unsupervised clustering models we can only predict the number of underlying classes that are part of the training data set as the supervised model is based on the classes present in the training data.

A. Single-cell RNA-sequencing data (ScRNA-Seq data)

Based on the paper of [28], we use the same single-cell RNA-sequencing datasets as [16]. As proposed by [28], we only include genes into our analysis with read counts higher than 1 transcript per million mapped reads (TPM) in more than 25% of the considered cells. This leads to a differing number of covariates p in case of the binary classification and the multiclass classification task, as shown in Table I. For the choice of cell types, we use the same selection as [16]. In case of the binary response, two selected cell types will be analyzed (left side of Table I). In case of the multiclass classification task (right side of Table I), we analyze K cell populations. The underlying numbers of cells in case of the binary response ($K = 2$) are N_1 and N_2, and for the multiclass response ($K > 2$) the respective cell populations are denoted with $N_1, \cdots, N_K$.
In accordance with the paper of [16], we consider their proposed binary classification tasks. However, instead of their approach of all pairwise combinations, we use a multinomial model for the $K > 2$ cases, which means one model per dataset. In the following, the cell types of the analyzed single-cell RNA-sequencing datasets are described. The EMTAB2805 data of [5] contain the cell cycle stages *G1, S, G2M* of the mouse embryonic stem cell (mESC). For the dataset GSE45719 [7] we include the different states of transition

TABLE I
BENCHMARK DATA, SHOWING THE NUMBER OF COVARIATES p, NUMBER OF OBSERVATIONS N, AND THE OBSERVATIONS PER CLASS N_1 VS. N_2 IN THE BINARY CLASSIFICATION TASK AND N_1 VS. N_2 VS. $\cdots$ VS. N_K IN THE MULTICLASS CLASSIFICATION TASK

	Binary Response				Multiclass Response			
	EMTAB2805	GSE45719	GSE48968	GSE74596	EMTAB2805	GSE45719	GSE48968	GSE74596
p	13,110	10,851	7,987	6,748	12,849	11,065	7,831	7,329
Subpopulation 1	*G1*	*mid blastocyst*	BMDC 1h LPS	NKT0	*G1*	*mid blastocyst*	BMDC 1h LPS	NKT0
N_1	96	60	96	45	96	60	96	45
Subpopulation 2	*G2M*	*16-cell stage embryo*	BMDC 4h LPS	NKT17	*G2M*	*16-cell stage embryo*	BMDC 4h LPS	NKT17
N_2	96	50	191	44	96	50	191	44
Subpopulation 3	-	-	-	-	*S*	*8-cell stage embryo*	*BMDC 6h LPS*	*NKT1*
N_3	-	-	-	-	96	37	191	46
Subpopulation 4	-	-	-	-		-	-	NKT2
N_4	-	-	-	-	-	-	-	68

of *mid blastocyst*, *8-cell stage embryo* as well *16-cell stage embryo*. In case of the single-cell RNA-sequencing data of GSE48968 bone marrow-derived dendritic cells (BMDCs) were stimulated with three different pathogenic components, analyzing the different responses for the dataset [25]. We will analyze only the component Lipopolysaccharides (LPS) at different timepoints (*1h, 4h, 6h*) after incubation. The data set GSE74596 contains different types of Natural killer T (NKT) cells extracted from the thymus. The cell types *NKT1, NKT2 and NKT17* are subtypes of the helper T cells [9]. The objective is to determine a supervised model that can classify the different cell types, given the expression profiles in these datasets. Also, as a second objective it is important to find a sparse solution to focus on the most important genes.

B. Competing Methods

The L1-regularized regression is carried out with the R package *glmnet* [13]. The λ values are found via the internal 10-fold CV approach and chosen as the value λ leading to the smallest estimated generalization error. For adaptive Lasso, the covariate specific penalty weights are determined with ridge regression $w_j = 1/\hat{\beta}_j^{Ridge}$ due to the $p >> N$ situation. We also compare our methods to the Elastic Net, as a baseline for good predictive performance. The Elastic Net is fit using *glmnet* and $\alpha = 0.5$, leading to an equal mixture of L1 and L2-penalization (cf. Section II).

For DP-Lasso the ANOVA based DP weights are implemented with the R package *stats* [23]. The Silhouette index is calculated with the R package *cluster* [18] and the Davies-Bouldin index with the package *clusterSim* [34]. The final DP-Lasso model is again fit using the *glmnet* procedure, with the covariate specific penalty weights derived from the DP.

C. Binary classification

In this section the results for the experiments on binary classification tasks are presented and analyzed.

1) Accuracy – Binary: Accuracy is measured in terms of the misclassfication rate, averaged over all folds. The results of the empirical comparison can be found in Table II. Overall, the Elastic Net shows the lowest misclassification rate, however the difference to the DP-Lasso models and the normal Lasso is only marginal. The only exception is the adaptive Lasso, which performs clearly worse compared to the other methods. This is likely due to the strong correlation present in the data.
The three proposed DP-Lasso models show only minor differences in terms of misclassification rate, with a slight advantage for DP-L$_{ANOVA}$. We conclude, that the accuracy of DP-Lasso is comparable to the competitors regardless of the choice of the discriminative power.

2) Number of Coefficients – Binary: If the primary objective is to identify biomarkers, it is very important to find sparse solutions, as the cost of follow up studies can be high. Next, we therefore analyze the number of covariates selected by each method, which is the number of non-zero coefficients left in the regularized model. Of all methods, the Elastic Net (Enet), as expected selects the highest number of covariates due to its use of the L2-penalty.
In all binary classification tasks, all DP-Lasso models select significantly fewer covariates than the competing methods. Often the difference is quite large. For example on the GSE74596 dataset DP-L$_{ANOVA}$ selects only 4 covariates, whereas Lasso selects 18. A likely explanation is the over-shrinkage effect in Lasso regression, which takes in irrelevant predictors (cf. Section II). On the other hand, DP-L$_{ANOVA}$ is able to reduce the penalty on the important covariates and reaches a very sparse solution.
From the class of DP-Lasso models, DP-L$_{ANOVA}$ is the most selective and finds the sparsest solutions. However, DP-L$_{DB}$ and DP-L$_{Sil}$ also produce smaller model sizes compared to the competing methods on all binary classification tasks.

D. Multiclass Classification

DP-Lasso can also be applied to multiclass ($K > 2$) classification. Note, that in case of $K > 2$ and the multinomial-logit model $K - 1$ coefficient vectors β are fit for the different categories, whereas one category is used as reference category. For each covariate, DP is measured as before leading to an equal penalization for each of the outcome categories.

TABLE II
THE MISCLASSIFICATION RATE AND ITS STANDARD DEVIATION IN BRACKETS FOR BINARY AND MULTICLASS CLASSIFICATION ON THE FOUR BENCHMARK DATASETS. THE BEST RESULT ON EACH DATASET (LOWEST NUMBER) IS MARKED IN BOLD.

	Binary				Multiclass			
	EMTAB2805	GSE45719	GSE48968	GSE74596	EMTAB2805	GSE45719	GSE48968	GSE74596
Lasso	0.05 (0.006)	**0.01** (0.000)	**0.02** (0.003)	**0.00** (0.000)	**0.06** (0.010)	0.03 (0.009)	0.19 (0.100)	**0.01** (0.003)
Elastic Net	**0.04** (0.006)	**0.01** (0.000)	**0.02** (0.000)	**0.00** (0.000)	**0.06** (0.007)	**0.02** (0.005)	0.18 (0.008)	**0.01** (0.004)
adaptive Lasso	0.11 (0.008)	0.02 (0.000)	0.07 (0.007)	0.15 (0.031)	0.17 (0.006)	0.10 (0.013)	0.26 (0.010)	0.28 (0.015)
DP-L$_{ANOVA}$	0.05 (0.006)	**0.01** (0.000)	**0.02** (0.004)	**0.00** (0.000)	**0.06** (0.009)	0.11 (0.017)	**0.17** (0.009)	0.03 (0.006)
DP-L$_{DB}$	0.05 (0.009)	**0.01** (0.000)	**0.02** (0.004)	0.01 (0.001)	0.08 (0.007)	0.07 (0.016)	0.20 (0.014)	0.03 (0.006)
DP-L$_{Sil}$	**0.04** (0.006)	**0.01** (0.000)	0.04 (0.004)	**0.00** (0.006)	0.18 (0.018)	0.06 (0.008)	0.24 (0.011)	0.06 (0.013)

TABLE III
THE NUMBER OF SELECTED COEFFICIENTS AND ITS STANDARD DEVIATION IN BRACKETS FOR BINARY AND MULTICLASS CLASSIFICATION ON THE FOUR BENCHMARK DATASETS. THE BEST RESULT (LOWEST NUMBER) ON EACH DATASET IS MARKED IN BOLD.

	Binary				Multiclass			
	EMTAB2805	GSE45719	GSE48968	GSE74596	EMTAB2805	GSE45719	GSE48968	GSE74596
Lasso	58 (1.9)	20 (0.4)	55 (0.9)	18 (0.6)	127 (3.5)	67 (1.0)	163 (5.5)	72(1.7)
Elastic Net	142 (1.8)	103 (1.1)	125 (1.2)	66 (0.5)	250 (13.1)	199 (1.5)	276 (10.2)	197 (1.9)
adaptive Lasso	38 (2.1)	13 (0.6)	48 (0.8)	27 (0.7)	65 (1.6)	36 (0.3)	84 (4.8)	52 (3.0)
DP-L$_{ANOVA}$	**17** (0.4)	**5** (0.1)	**19** (0.4)	**4** (0.2)	**45** (0.6)	**23** (1.2)	**70** (1.1)	**17** (0.5)
DP-L$_{DB}$	25 (0.9)	9 (0.1)	30 (0.3)	7 (0.1)	71 (1.3)	39 (0.8)	125 (1.6)	37 (0.3)
DP-L$_{Sil}$	22 (0.5)	9 (0.3)	36 (0.6)	8 (0.4)	181 (2.2)	32 (0.8)	172 (1.8)	90 (3.3)

In contrast to the binary case, the adaptive Lasso uses a different penalization weight for each covariate and outcome category again resulting from the ridge estimator.

1) Accuracy – Multiclass: Accuracy is again measured as misclassification rate. The results can be found in Table II.
Of all methods the Elastic Net shows the strongest predictive performance, followed by the Lasso. The adaptive Lasso again performs clearly worse on all datasets in terms of accuracy.
From the DP-Lasso models, DP-L$_{DB}$ is competitive on most datasets, and DP-L$_{ANOVA}$ remains competitive on three of the datasets showing significantly worse performance on the GSE45719 data. DP-L$_{Sil}$ performs worse overall in the multinomial setting, but still notably better than the adaptive Lasso.

2) Number of Coefficients – Multiclass: In terms of model size, DP-L$_{ANOVA}$ again uniformly produces the sparsest solutions on all datasets. Lasso and Elastic Net keep around 3 to 10 times more non-zero coefficients in the respective models.
DP-L$_{DB}$ also produces relatively small models, on par with the adaptive Lasso, whereas DP-L$_{Sil}$ clearly struggles on the EMTAB2805, GSE48968 and GSE74596 datasets.

E. Empirical Results Summary

The empirical comparison on benchmark data indicates that DP-Lasso is able to maintain a high accuracy. At the same time DP-Lasso finds significantly smaller models, often by a factor of 3 to 10 compared to Lasso and Elastic Net. This is due to the incorporation of the DP into the penalization scheme, which helps to remove uninformative genes and focus instead on the relevant ones.
To summarise, DP-Lasso and especially DP-L$_{ANOVA}$ produces significantly smaller model sizes, while being able to maintain accuracy on par with current state-of-the-art regularized regression approaches.

V. SIMULATION STUDY

In this section, we test our method on simulated data. The setup is as follows. $X_1, ..., X_{10}$ are drawn from a normal distribution $\mathcal{N}(-1, \sigma)$, for observations of class 1, and from $\mathcal{N}(1, \sigma)$ for observations of class 2. This reflects the assumption that relevant genes express differently between the target groups. All additional covariates $X_{11}, ..., X_p$ are drawn from $\mathcal{N}(0, \sigma)$ and can therefore be considered as irrelevant. We test the values $p \in \{100, 1000, 5000\}$ and $\sigma^2 \in \{1, 2, 3\}$ and draw $N = 100$ observations in each setting. With increasing σ the groups become more overlapping and we expect learning to become increasingly difficult. Note that the covariates are drawn independently, implying $X \sim \mathcal{N}_p(\mu, \sigma^2 \mathcal{I}_p)$, where $\mathcal{I}$ is the identity matrix, making it an ideal situation for all methods. Each experiment is repeated 10 times and the results are averaged.
As in this experiment the relevant covariates are known, we measure the method's capabilities to identify the decisive covariates. To this end, we measure the Precision as

$$\text{Precision} = \frac{||\hat{\beta}_{true}||^0}{||\hat{\beta}||^0}, \tag{13}$$

TABLE IV
THE PRECISION AND RECALL ON THE DIFFERENT SIMULATION SETTINGS, AVERAGED OVER 10 RUNS. RESULTS ARE PRESENTED AS PRECISION / RECALL. FOR EACH SETTING THE METHOD WITH THE HIGHEST PRECISION IS MARKED IN BOLD.

	$\sigma^2 = 1$			$\sigma^2 = 2$			$\sigma^2 = 3$		
	$p = 100$	$p = 1000$	$p = 5000$	$p = 100$	$p = 1000$	$p = 5000$	$p = 100$	$p = 1000$	$p = 5000$
Lasso	0.86 / 0.99	0.60 / 0.99	0.53 / 0.98	0.45 / 0.96	0.32 / 0.96	0.23 / 0.93	0.48 / 0.95	0.37 / 0.84	0.28 / 0.84
Elastic Net	0.55 / 1.00	0.27 / 1.00	0.20 / 1.00	0.29 / 1.00	0.15 / 1.00	0.10 / 0.98	0.37 / 0.99	0.20 / 0.96	0.14 / 0.91
adaptive Lasso	0.99 / 0.97	0.97 / 0.98	0.94 / 0.95	0.88 / 0.98	0.58 / 0.96	0.37 / 0.91	0.71 / 0.93	0.35 / 0.82	0.28 / 0.85
DP-L$_{ANOVA}$	**1.00** / 0.87	**1.00** / 0.92	**1.00** / 0.85	**0.99** / 0.95	**0.88** / 0.93	**0.80** / 0.91	**0.82** / 0.87	**0.50** / 0.83	**0.38** / 0.85
DP-L$_{DB}$	**1.00** / 0.95	**1.00** / 0.94	**1.00** / 0.92	0.92 / 0.98	0.77 / 0.94	0.50 / 0.91	0.71 / 0.94	0.35 / 0.85	0.28 / 0.84
DP-L$_{Sil}$	**1.00** / 0.94	**1.00** / 0.94	**1.00** / 0.91	0.96 / 0.98	0.76 / 0.93	0.67 / 0.90	0.63 / 0.88	0.41 / 0.79	0.31 / 0.81

where $||\cdot||^0$ specifies the 0-norm, which counts up the non-zero entries and $\hat{\beta}_{true}$ denotes the first ten entries of the coefficient vector, which by design we know to be the correct effects. $\hat{\beta}$ denotes all coefficients obtained by the regularized model. This measure is useful as the number of potential covariates is high. However, if the model has a high Precision, the identified genes can be trusted.
Secondly, we measure the Recall

$$\text{Recall} = \frac{||\hat{\beta}_{true}||^0}{10}, \tag{14}$$

as the fraction of the relevant covariates that was discovered by the model.
The results are shown in Table IV. We can see that the DP-Lasso models show significantly higher Precision compared to Lasso and Elastic Net. The adaptive Lasso performs better than the Lasso in this ideal setting, in contrast to the results on the real data from the previous section. Overall DP-L$_{DB}$ and DP-L$_{ANOVA}$ show the highest Precision, even in very difficult data situations. For instance, with $N = 100, p = 5000, \sigma = 1$, DP-L$_{ANOVA}$, DP-L$_{DB}$ and DP-L$_{Sil}$ are able to maintain a 100% Precision and thus are very selective and able to find the correct covariates. DP-L$_{ANOVA}$ has the highest Precision in every setting.
It is also important to compare the Recall, as it reflects the fraction of true effects that are found by a model. Elastic Net shows the highest Recall, which is a result of the large number of coefficients that were kept in the model. On the other hand, all DP-Lasso models show a Recall which is typically slightly lower but still competitive with Lasso and adaptive Lasso. This again is due to the very selective nature of DP-Lasso.
Overall, we conclude that the non-zero coefficients found by the DP-Lasso can be trusted more to reflect true mechanisms, compared to its competitors. At the same time DP-Lasso is capable to maintain a competitive Recall.
It is reassuring to note that on average the accuracy of the methods measured by the area under the curve AUC is very similar, with a slight edge for the DP-L$_{DB}$, DP-L$_{ANOVA}$ and the Elastic Net.

VI. CONCLUSION

With DP-Lasso, we propose a novel regularization based approach for covariate selection in the context of gene expression data. Incorporating univariate measures of discriminative power that are based on the principles of separation and compactness enriches the model with additional information. Our approach can also be interpreted as soft filtering: instead of removing genes a-priori, more promising genes are simply promoted, freeing the modeller from ad-hoc choices, such as selecting the correct number of genes to remove. In a broader context we argue therefore that soft filtering instead of hard filtering also enhances reproducibility, as it reduces the 'researchers degrees of freedom' [26] involved in a study.

Empirically, we show that DP-Lasso is on par with the popular methods Lasso and Elastic Net in terms of accuracy, while it chooses significantly less genes. With a simulation study we confirm that DP-Lasso is capable of ignoring a large number of irrelevant predictors and instead focusses on the truly relevant ones – due to the double criteria of being relevant both univariately and in the multivariate model. This selectiveness is very desirable in the context of gene expression data, as both the number of candidate genes is high and follow-up studies are costly. Therefore, a short but confident list of very promising genes, as given by the DP-Lasso model, is preferred in this context.

As currently the discriminative power is calculated univariately, it does not explicitly take the correlation structure of the covariates into account. An interesting direction for future work would therefore be to extend the DP-Lasso approach by considering the correlation structure between covariates and adjust the penalization accordingly, similar to the approach in [33].

In this article, we focused on the application for genetic classification data, however DP-Lasso can also be applied in other domains. As long as the classes are expected to show differences in the univariate distribution of covariates, we expect DP-Lasso to deliver a good predictive performance coupled with a low number of selected covariates.

ACKNOWLEDGEMENTS

The first authors are very grateful for the support of the LMU mentoring program, connecting young researchers and providing mentors that give individual advice. In addition, we would like to thank Gerhard Tutz and Christian L. Müller for the very insightful and valuable discussion.

References

[1] Adelin Albert and John A Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

[2] Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, and Brandon Willard. Lasso meets horseshoe: A survey. *Statistical Science*, 34(3):405–427, 2019.

[3] Nadia Bolshakova and Francisco Azuaje. Cluster validation techniques for genome expression data. *Signal Processing*, 83(4):825–833, 2003.

[4] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[5] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160, 2015.

[6] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 224–227, 1979.

[7] Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196, 2014.

[8] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):1–13, 2006.

[9] Isaac Engel, Grégory Seumois, Lukas Chavez, Daniela Samaniego-Castruita, Brandie White, Ashu Chawla, Dennis Mock, Pandurangan Vijayanand, and Mitchell Kronenberg. Innate-like functions of natural killer T cell subsets result from highly divergent gene programs. *Nature Immunology*, 17(6):728–739, 2016.

[10] Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression*. Springer, 2007.

[11] Liang Fang and Cheng Su. *Statistical Methods in Biomarker and Early Clinical Development*. Springer, 2019.

[12] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in Statistics*, pages 66–70. Springer, 1992.

[13] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[14] Joyee Ghosh, Yingbo Li, and Robin Mitra. On the use of cauchy prior distributions for bayesian logistic regression. *Bayesian Analysis*, 13(2):359–383, 2018.

[15] Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The Elements of statistical learning: Data mining, inference, and prediction*. Springer, 2017.

[16] Beyrem Khalfaoui and Jean-Philippe Vert. DropLasso: A robust variant of Lasso for single cell RNA-seq data. *hal-01716704v2*, 2019.

[17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, 2017.

[18] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2019. R package version 2.1.0.

[19] M Man, TS Nguyen, C Battioui, and G Mi. Predictive subgroup/biomarker identification and machine learning methods. In *Statistical Methods in Biomarker and Early Clinical Development*, pages 1–22. Springer, 2019.

[20] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[21] Masanori Oshi, Hideo Takahashi, Yoshihisa Tokumaru, Li Yan, Omar M Rashid, Ryusei Matsuyama, Itaru Endo, and Kazuaki Takabe. G2m cell cycle pathway score as a prognostic biomarker of metastasis in estrogen receptor (er)-positive breast cancer. *International Journal of Molecular Sciences*, 21(8):2921, 2020.

[22] Trevor Park and George Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[24] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[25] Alex K Shalek, Rahul Satija, Joe Shuga, John J Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S Gertner, Jellert T Gaublomme, Nir Yosef, et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369, 2014.

[26] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.

[27] Manabu Soda, Young Lim Choi, Munehiro Enomoto, Shuji Takada, Yoshihiro Yamashita, Shunpei Ishikawa, Shin-ichiro Fujiwara, Hideki Watanabe, Kentaro Kurashina, Hisashi Hatanaka, et al. Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. *Nature*, 448(7153):561–566, 2007.

[28] Charlotte Soneson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261, 2018.

[29] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):1–11, 2008.

[30] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009.

[31] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[32] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[33] Gerhard Tutz and Jan Ulbricht. Penalized regression with correlation-based penalty. *Statistics and Computing*, 19(3):239–253, 2009.

[34] Marek Walesiak and Andrzej Dudek. The choice of variable normalization method in cluster analysis. In *Education Excellence and Innovation Management: A 2025 Vision to Sustain Economic Development During Global Challenges*, pages 325–340. International Business Information Management Association (IBIMA), 2020.

[35] Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[36] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[37] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, §8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 01.04.2022

Cornelia Sigrid Fütterer