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:

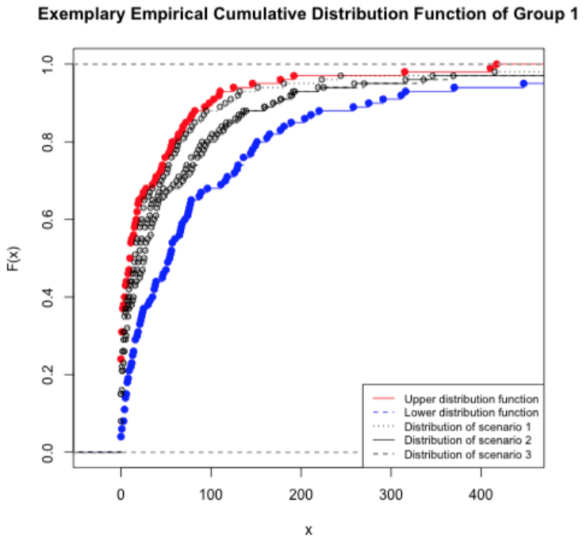

Figure 1: Lower and upper cumulative distribution function of simulated gene 3 for group 1 using the statistical software R [9].

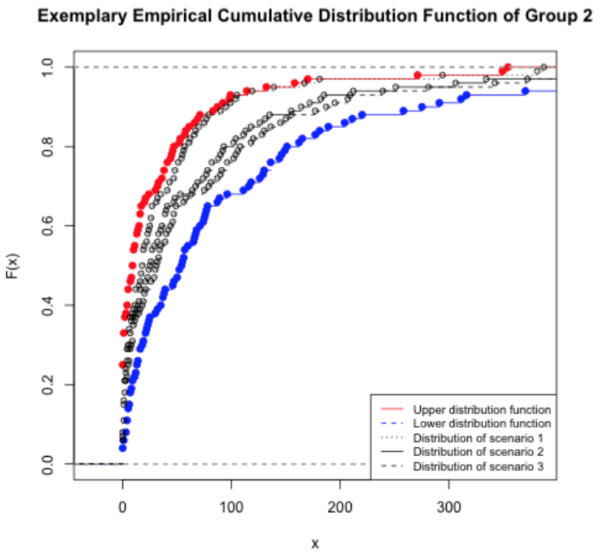

Figure 2: Lower and upper cumulative distribution function of simulated gene 3 for group 2 using the statistical software R [9].

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) $C(F_1^{(g)}, F_2^{(g)}) = C(0, F_2^{(g)}) = 0, \qquad \forall\, F_1^{(g)}, F_2^{(g)} \in [0,1]$
(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]:

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