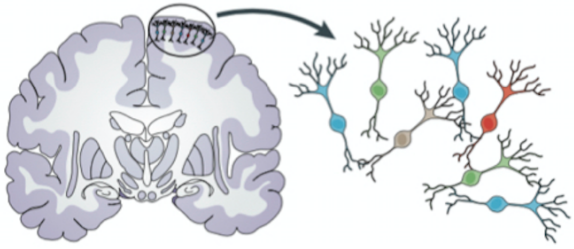
Constructing Simulation Data with Dependence Structure for Unreliable Single-Cell RNA-sequencing Data using Copulas

M. Sc. Cornelia Fuetterer
Institut für Statistik, Ludwig-Maximilians Universität München
Dr. Georg Schollmeyer,
Institut für Statistik, Ludwig-Maximilians Universität München
Prof. Dr. Thomas Augustin,
Institut für Statistik, Ludwig-Maximilians Universität München

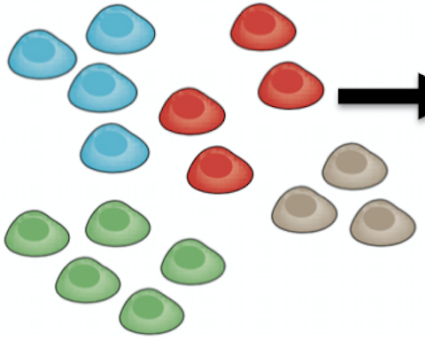
Working group

Biological application

Tissue (e.g. tumor)



Isolate and sequence individual cells

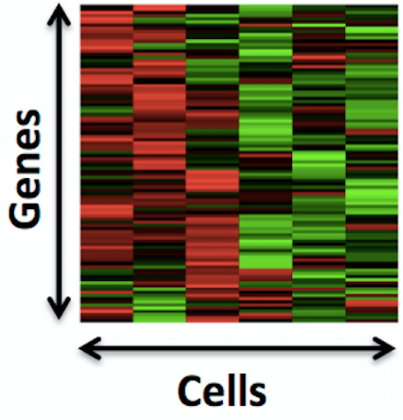


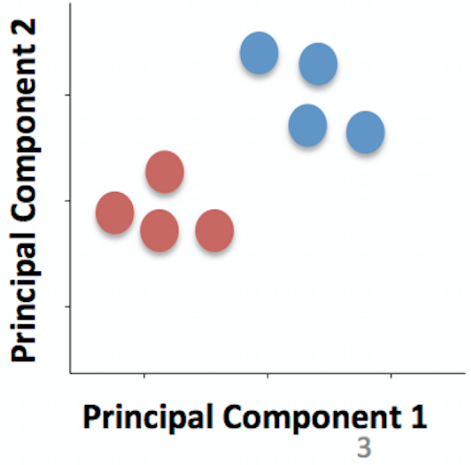
Gene 1
Cell 1

Read Counts

	Cell 1	Cell 2	…
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
…			

Compare gene expression profiles of single cells





Constructing Simulation Data with Dependence Structure for Unreliable Single-Cell RNA-sequencing Data using Copulas

1. Construction of Simulation Data
2. Incorporation of Dependence Structure
3. Consequences with regard to Application

Outline

Distribution Approximation of the Distribution of Read Counts

Best distribution approximation of read counts:
Zero Inflated Negative Binomial (ZINB) Zeileis et al. (2008), Wagner et al. (2013) and Kleiber and Zeileis (2016):

Zero Inflated Negative Binomial (ZINB):

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

Generalisation of the negative binomial distribution:
Mixture of Poisson distributions with a gamma distributed poisson rate

$$f_{NB}(X_j = x) = \frac{\Gamma(x+\phi)}{\Gamma(\phi) \cdot x!} \cdot \frac{\mu^x \cdot \phi^\phi}{(\mu+\phi)^{x+\phi}} \cdot \mathbf{I}_{\mathbb{N}}(x)$$

Different Degrees of Heterogeneity

Basis of the Simulation Design:
Quantiles of the estimated parameters
Based on the 7225 genes of the real data set Kolodziejczyk et al. (2015)

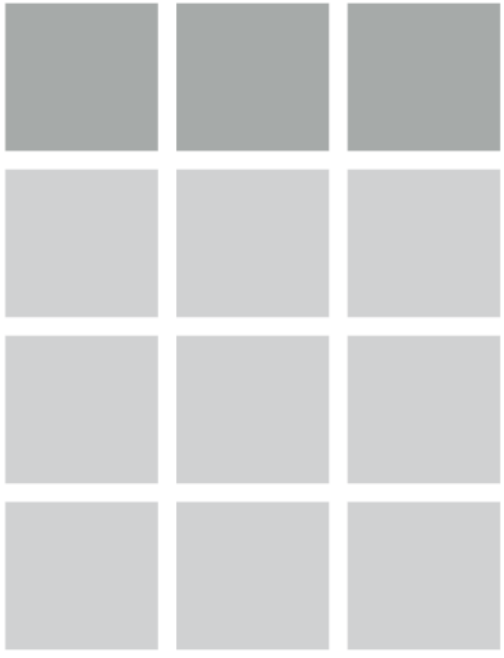
Scenario 1 Most homogeneous scenario
$\Rightarrow$ Narrowest parameter interval
Scenario 3 Most heterogeneous scenario
$\Rightarrow$ Broadest parameter interval

	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: 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.

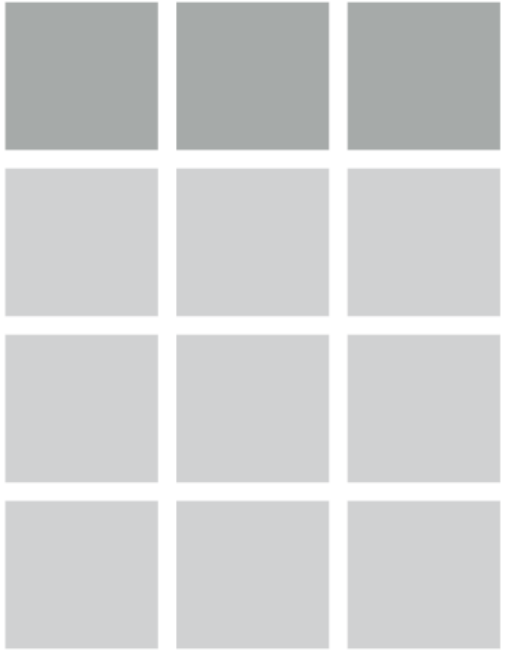
Undistorted Simulation Data - No dependence structure

Scenario 1:
Homogenous



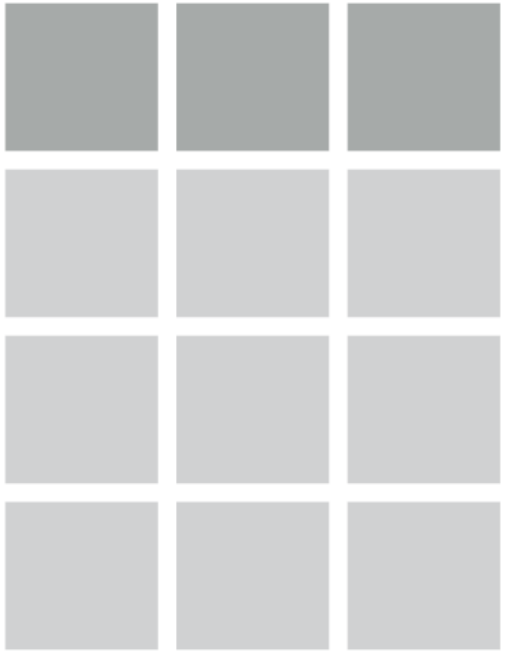
$(n^{(1)} + n^{(2)}) \times m$

Scenario 2:
Transition



$(n^{(1)} + n^{(2)}) \times m$

Scenario 3:
Heterogeneous



$(n^{(1)} + n^{(2)}) \times m$

Constructing Distorted Data via Lower and Upper Distribution Functions

Upper distribution function: Measuring tendencially decreased read counts
Lower distribution function: Measuring tendencially increased read counts

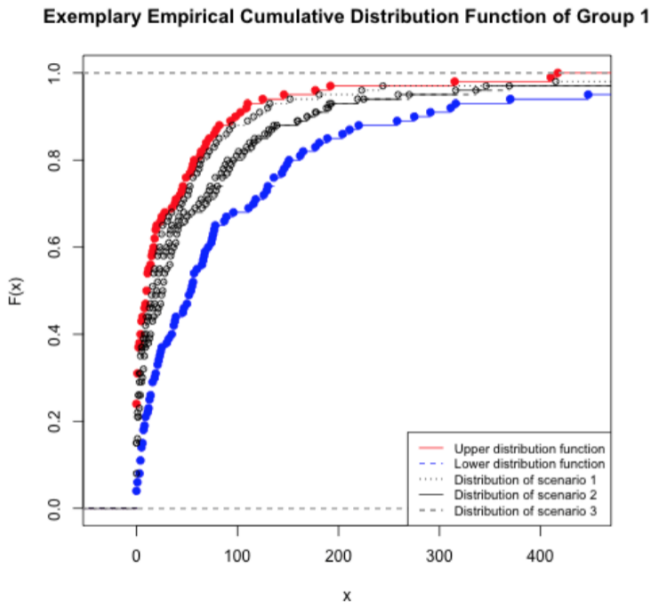


Figure: Lower and upper cumulative distribution function of simulated gene 3 for group 1 using the statistical software R of the R Core Team (2014).

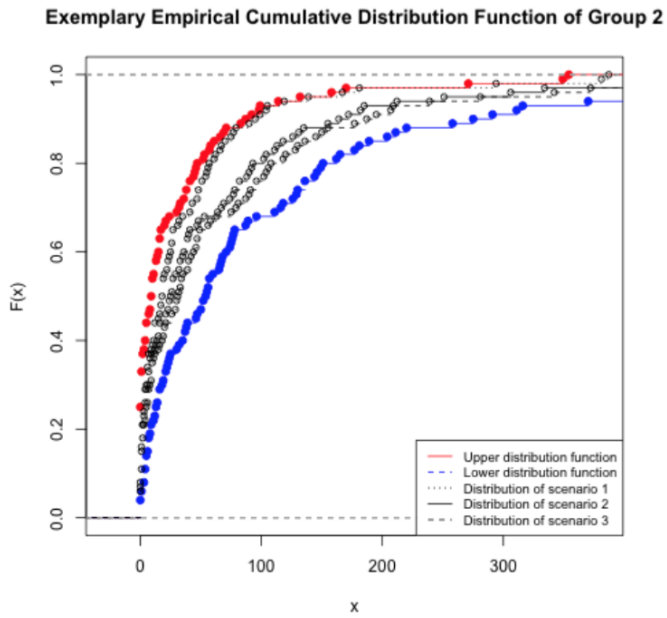
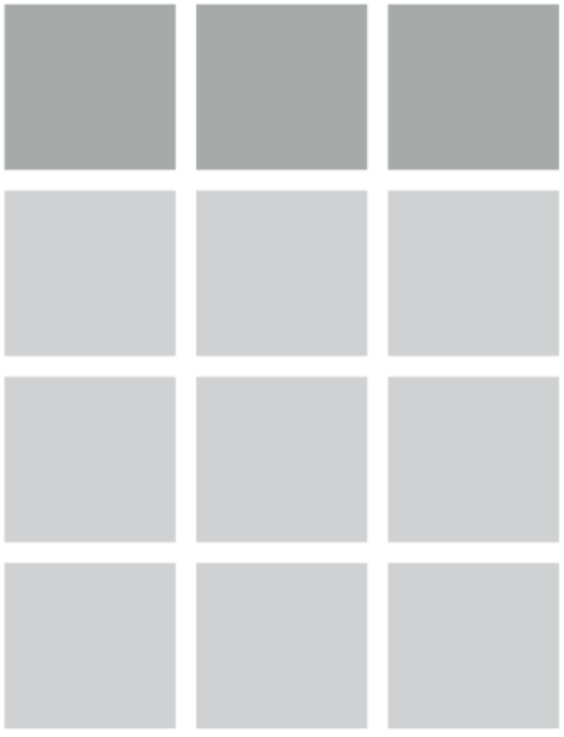


Figure: Lower and upper cumulative distribution function of simulated gene 3 for group 2 using the statistical software R of the R Core Team (2014).

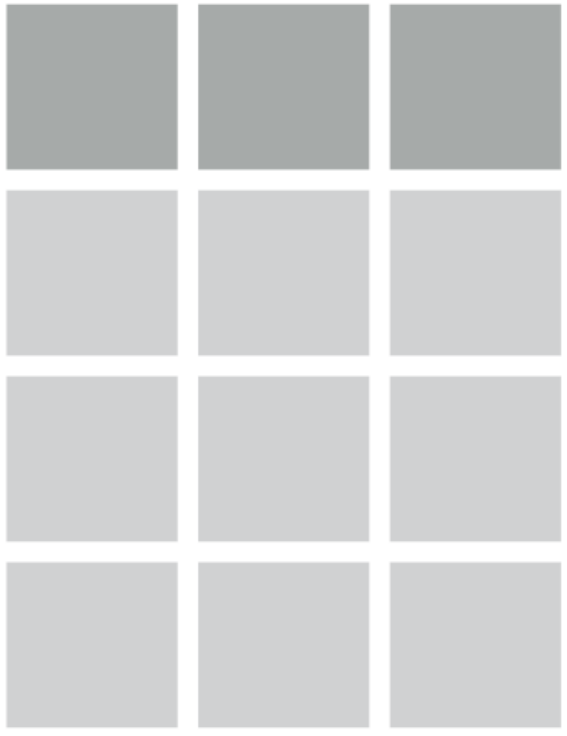
Distorted Simulation Data - No dependence structure

Upper Distribution:



$(n^{(1)} + n^{(2)})$ x m

Lower Distribution:



$(n^{(1)} + n^{(2)})$ x m

Outline

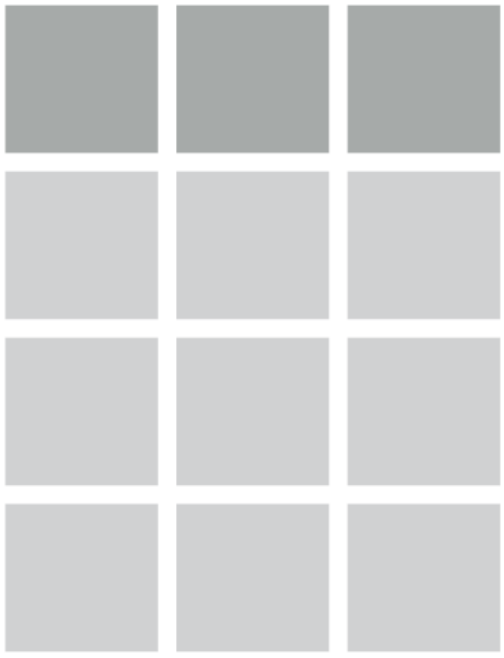
Dependence Structure using Copulas

Sklar (1959) 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:

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

Undistorted Simulation Data - With dependence structure

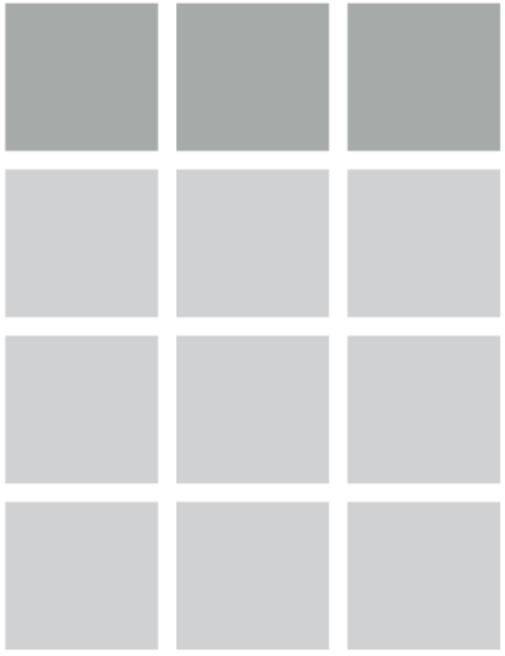
Scenario 1:
Homogenous



$(n^{(1)} + n^{(2)}) \times m$

Gaussian Copula
Clayton Copula
Frank Copula

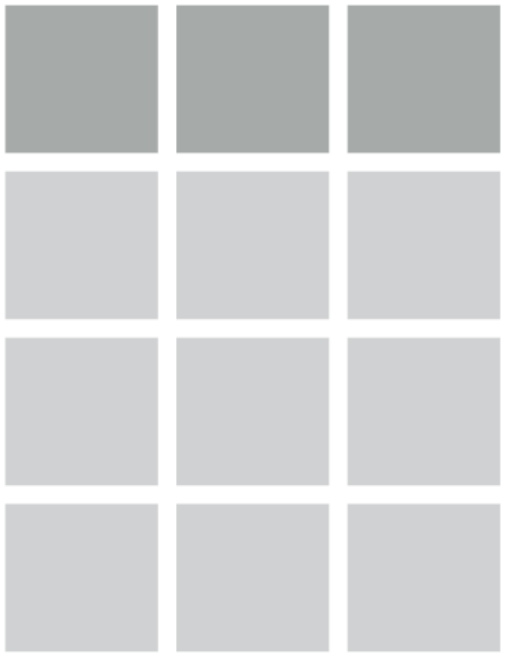
Scenario 2:
Transition



$(n^{(1)} + n^{(2)}) \times m$

Gaussian Copula
Clayton Copula
Frank Copula

Scenario 3:
Heterogeneous



$(n^{(1)} + n^{(2)}) \times m$

Gaussian Copula
Clayton Copula
Frank Copula

Distorted Data with Dependence Structure

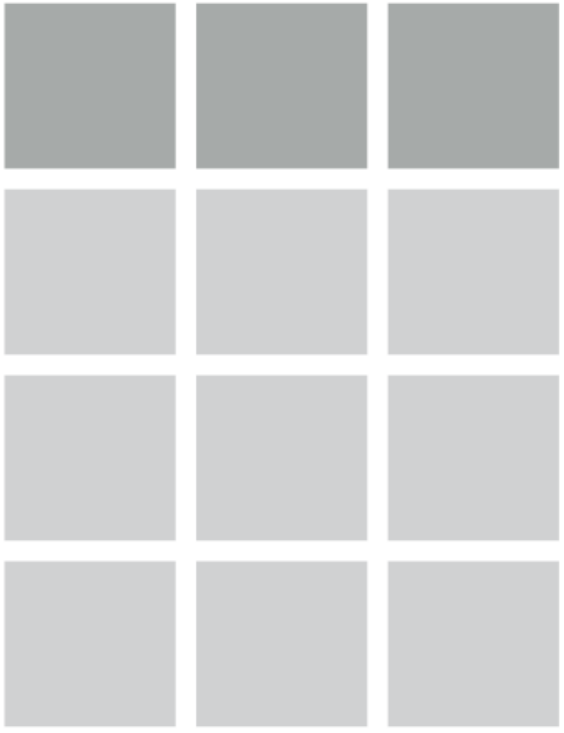
Distorted data are no longer ZINB distributed:

$\Rightarrow$ No parametric marginals anymore
$\Rightarrow$ Computation of upper and lower cumulative distribution function in order to sample from the joint distribution, keeping the same marginals:

$$\underline{\hat{F}}^{(g)}_{\mathbf{X}}(x_1, ..., x_m) = C_v(\underline{\hat{F}_1}^{(g)}(x_1), \underline{\hat{F}_2}^{(g)}(x_2), ..., \underline{\hat{F}_m}^{(g)}(x_m))$$
$$\overline{\hat{F}}^{(g)}_{\mathbf{X}}(x_1, ..., x_m) = C_v(\overline{\hat{F}_1}^{(g)}(x_1), \overline{\hat{F}_2}^{(g)}(x_2), ..., \overline{\hat{F}_m}^{(g)}(x_m))$$

Distorted Simulation Data - With dependence structure

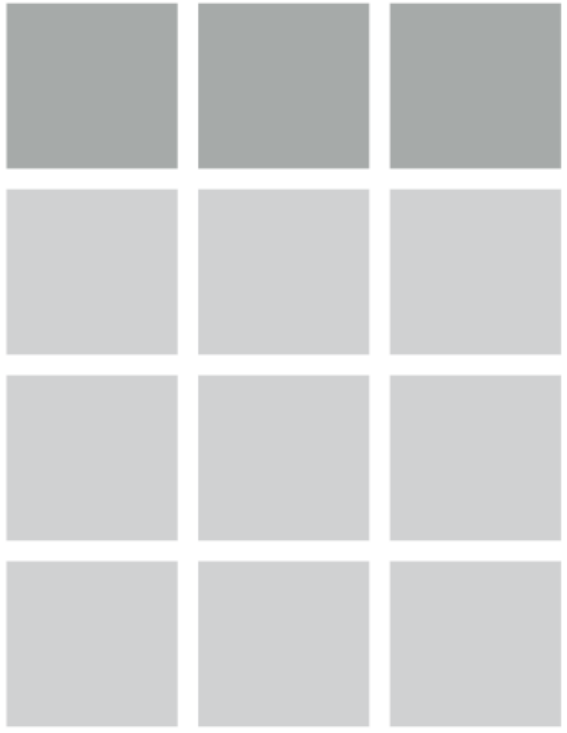
Upper Distribution:



$(n^{(1)} + n^{(2)})$ x m

Gaussian Copula
Clayton Copula
Frank Copula

Lower Distribution:



$(n^{(1)} + n^{(2)})$ x m

Gaussian Copula
Clayton Copula
Frank Copula

Outline

Results of the application

Undistorted data:

- Classification improvement with a higher number of genes

Distorted data:

- Upwards distorted (Lower Distribution):
  A lot of variation possible due to ($\mathbb{W} \in [0, \infty)$)
  $\Rightarrow$ Easier distinctions of the target groups
- Downwards distorted (Upper Distribution):
  Less variation possible due to $\mathbb{W} \in [0, \infty)$
  $\Rightarrow$ Difficult distinctions of the target groups
- Upwards distortion results in better accuracy than downwards distortion

Discussion

Intention of simulation data:

- Reflection of measurement error of an instrument
- Allowance for calibration of measuring instruments in the appropriate direction (Current state-of-the-art: tends to miss low read counts)

References

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

Kolodziejczyk, A. A., J. K. Kim, J. C. Tsang, T. Ilicic, J. Henriksson, K. N. Natarajan, A. C. Tuck, X. Gao, M. Bühler, P. Liu, J. C. Marioni, and S. A. Teichmann (2015). Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell 17*, 471–85.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sklar, A. (1959). Fonctions de Répartition à n Dimensions Et Leurs Marges. *Publications de l'Institut Statistique de l'Université de Paris 8*, 229–231.

Wagner, G. P., K. Kin, and V. J. Lynch (2013). A model based criterion for gene expression calls using RNA-seq data. *Theory in Biosciences 132*, 48–66.

Zeileis, A., C. Kleiber, and S. Jackman (2008). Regression models for count data in r. *Journal of Statistical Software 27 (8)*.