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

Cornelia Fuetterer[1], Georg Schollmeyer[1], Thomas Augustin[1]

[1] *Department of Statistics, LMU Munich, Germany*

Cornelia.Fuetterer@stat.uni-muenchen.de

Research Question

Single-Cell RNA-Sequencing

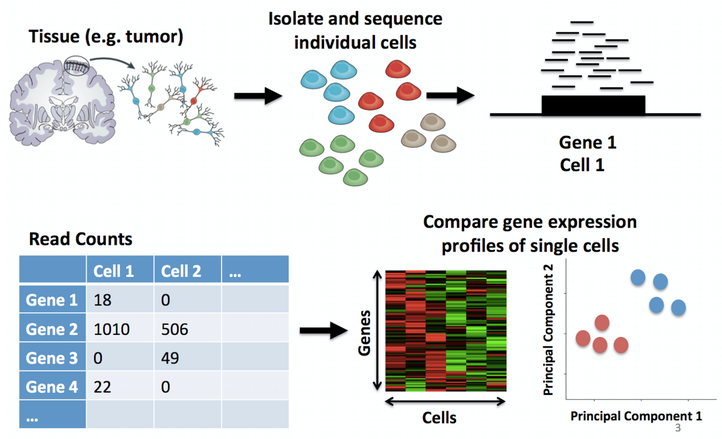


Figure 1: Single-cell RNA sequencing

Reference Data for Simulation data

Aim: Classification of modular transcriptional variation into its observed pluripotent states (Mouse Embryo Stem Cells)

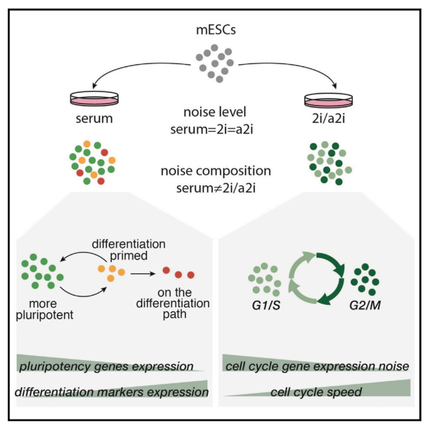


Figure 2: Inspiration for target groups

① Situations reflecting Different Degrees of Heterogeneity

Best approximation of read counts: Zero Inflated Negative Binomial (ZINB) distribution:

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

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

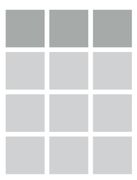
	μ		ϕ	π
Sc.	Group 1	Group 2	Group 1, Group 2	Group 1, Group 2
1	[35%-80%]	[15%-60%]	[45%-55%]	[45%-55%]
2	[25%-85%]	[10%-70%]	[40%-60%]	[40%-60%]
3	[20%-90%]	[5%-75%]	[35%-65%]	[35%-65%]

Table 1: Quantiles of the estimated ZINB parameters of the real data set

- Number of single-cells of target group 1: $n^{(1)} = 250$
- Number of single-cells of target group 2: $n^{(2)} = 250$
- Number of genes: m = 50, 100, 500

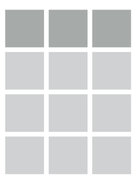
Basis of the simulation design:

Scenario 1:



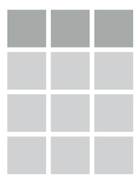
Most homogeneous (Narrowest parameter interval)

Scenario 2:



Transition from homogeneous to heterogeneous

Scenario 3:



Most heterogeneous (Largest parameter interval)

$\Rightarrow$ Parameter set for group g:

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

② Constructing distorted data $\underline{F_j}^{(g)}$, $\overline{F_j}^{(g)}$: $\mathbb{R} \to$[0,1]

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

$\mathcal{F}_j^{(g)}$: Set of possible distribution functions for each gene of each target group based on the three constructed scenarios

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

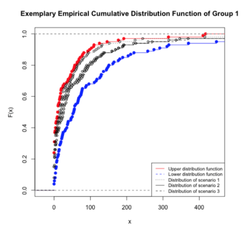


Figure 3: Lower and upper cumulative distribution function of simulated gene 3 for group 1

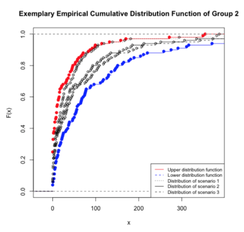


Figure 4: Lower and upper cumulative distribution function of simulated gene 3 for group 2

③ Constructing Distorted Data with dependence structure

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:

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

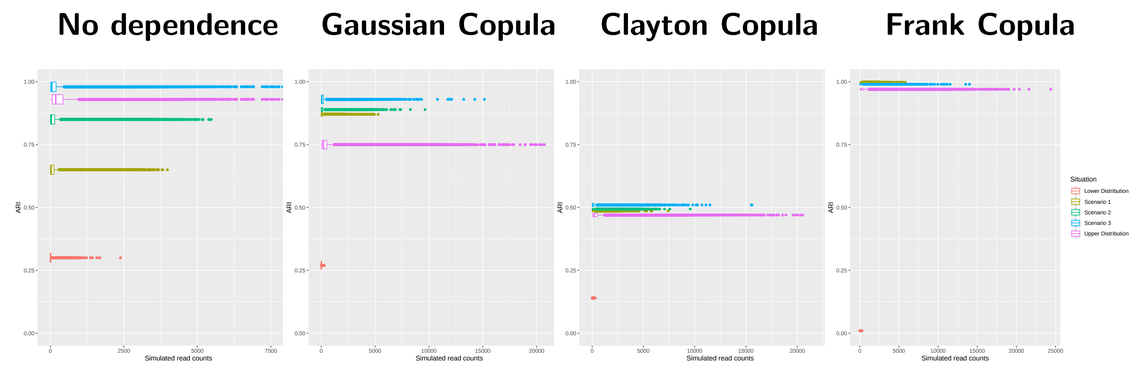
Distorted data are no longer ZINB distributed:

$\rightarrow$ No parametric marginals anymore

$\rightarrow$ Computation of lower and upper cumulative distribution function in order to sample from the joint distribution, keeping the same marginals:

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

④ Classification Results



⑤ Conclusion

Distorted data:

- Upwards distorted (Lower Distribution):
  A lot of variation possible due to $\mathbb{W}$
  $\Rightarrow$ Higher variation of gene expression
  $\Rightarrow$ Easier distinctions of the target groups
- Downwards distorted (Upper Distribution):
  Less variation possible due to $\mathbb{W}$
  $\Rightarrow$ Lower variation of gene expression
  $\Rightarrow$ Difficult distinctions of the target groups

References

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

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

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

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.

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.