Distribution-free complex hypothesis testing for single-cell RNA-seq differential expression analysis

Boris Hejblum

Univ. Bordeaux ISPED, Inserm BPH U1219, Inria BSO, SISTM, Bordeaux, France
Vaccine Research Institute, Créteil, France

25 Janvier 2021

Background

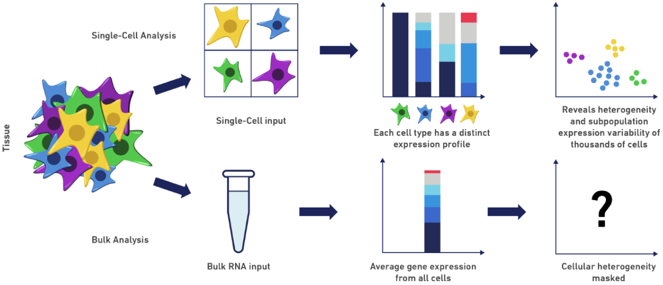
Bulk RNA-seq vs. Single-cell RNA-seq

- **bulk RNA-seq: average gene expression**

  ⇒ Mask signal coming from individual cells, ignoring tissue heterogeneity

- **single-cell RNA-seq: individual gene expression** from hundreds/thousands cells

  ⇒ Study biological processes that can only be observed at the cell level


Single-Cell Analysis
Tissue
Single-Cell input
Each cell type has a distinct expression profile
Reveals heterogeneity and subpopulation expression variability of thousands of cells
Bulk Analysis
Bulk RNA input
Average gene expression from all cells
?
Cellular heterogeneity masked

[*Source: 10xgenomics.com*]

New biological questions

This “new” technology allows to:

- detect different cell types
- characterize cellular heterogeneity
- perform cell maturation trajectory
- . . .

Associated statistical challenges

Many methodological challenges arise . . . **[Lähneman *et al.* *Genome Biology*, 2020]**

Associated statistical challenges

Many methodological challenges arise . . . **[Lähneman *et al.* *Genome Biology*, 2020]**

Differential Expression Analysis (DEA) from scRNA-seq data:

1. **Distribution** of gene expression across cells
   - **Sparsity:** large number of zeros ("dropouts")
     ⇒ Tiny amount of RNAs & low capture efficiency in a cell
   - **Heterogeneity:** multimodal and heterogeneous patterns
     ⇒ Different cell types, mRNA contents, cell states . . .

Associated statistical challenges

Many methodological challenges arise . . . [Lähneman *et al.* *Genome Biology*, 2020]

Differential Expression Analysis (DEA) from scRNA-seq data:

1. **Distribution** of gene expression across cells
   - **Sparsity:** large number of zeros ("dropouts")
     ⇒ Tiny amount of RNAs & low capture efficiency in a cell
   - **Heterogeneity:** multimodal and heterogeneous patterns
     ⇒ Different cell types, mRNA contents, cell states . . .
2. **Complex differential patterns**
   - difference in mode, in proportion, in both . . .

Associated statistical challenges

Many methodological challenges arise . . . **[Lähneman *et al.* *Genome Biology*, 2020]**

Differential Expression Analysis (DEA) from scRNA-seq data:

1. **Distribution** of gene expression across cells
   - **Sparsity:** large number of zeros ("dropouts")
     ⇒ Tiny amount of RNAs & low capture efficiency in a cell
   - **Heterogeneity:** multimodal and heterogeneous patterns
     ⇒ Different cell types, mRNA contents, cell states . . .
2. **Complex differential patterns**
   - difference in mode, in proportion, in both . . .
3. (in)dependent **multiple-sample** analysis
   - hierarchical observation levels

Associated statistical challenges

Many methodological challenges arise . . . [Lähneman *et al.* *Genome Biology*, 2020]

Differential Expression Analysis (DEA) from scRNA-seq data:

1. **Distribution** of gene expression across cells
   - **Sparsity:** large number of zeros ("dropouts")
     ⇒ Tiny amount of RNAs & low capture efficiency in a cell
   - **Heterogeneity:** multimodal and heterogeneous patterns
     ⇒ Different cell types, mRNA contents, cell states . . .
2. **Complex differential patterns**
   - difference in mode, in proportion, in both . . .
3. (in)dependent **multiple-sample** analysis
   - hierarchical observation levels

⇒ need for a new flexible method

State-of-the-art in DEA methods for scRNA-seq

- **Parametric methods**
  - scDD – Dirichlet process Gaussian mixture model + Bayes Factor **[Korthauer et al., 2016]**
  - MAST – 2 part glm **[Finak et al., 2015]**
  - SCDE – Bayesian mixture of Poisson & NB **[Kharchenko et al., 2014]**
  - DEsingle – ZINB + LRT **[Miao et al., 2018]**
- **Non-parametric methods**
  - EMDomics – Wassertein distance **[Nabavi et al., 2016]**
  - SigEMD – Wassertein distance + imputation **[Wang & Nabavi, 2018]**
  - D3E – Cramer-von Mises / Kolmogorov-Smirnov / Anderson-Darling test **[Delmans & Hemberg, 2016]**
  - scDD – Kolmogorov-Smirnov **[Korthauer et al., 2016]**
  - `distinct` – cdf comparison, requires biological replicates **[Tiberi et al., 2020]**

State-of-the-art in DEA methods for scRNA-seq

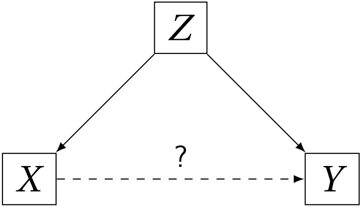
- **Parametric methods**
  - scDD – Dirichlet process Gaussian mixture model + Bayes Factor **[Korthauer et al., 2016]**
  - MAST – 2 part glm **[Finak et al., 2015]**
  - SCDE – Bayesian mixture of Poisson & NB **[Kharchenko et al., 2014]**
  - DEsingle – ZINB + LRT **[Miao et al., 2018]**
- **Non-parametric methods**
  - EMDomics – Wassertein distance **[Nabavi et al., 2016]**
  - SigEMD – Wassertein distance + imputation **[Wang & Nabavi, 2018]**
  - D3E – Cramer-von Mises / Kolmogorov-Smirnov / Anderson-Darling test **[Delmans & Hemberg, 2016]**
  - scDD – Kolmogorov-Smirnov **[Korthauer et al., 2016]**
  - `distinct` – cdf comparison, requires biological replicates **[Tiberi et al., 2020]**

Limitations

- strong **distributional** assumptions
- **2–group** comparisons only
- no **covariate adjustment** (except MAST & `distinct`)

Methods

DEA & Conditional independence test



Conditional dependence graph [Li *et al.* 2020]

Complex designs

- Y: scRNA-seq expression
- X: variable of interest (multi-dimensional, continuous and/or discrete)
- Z: covariates (multi-dimensional, continuous and/or discrete)

Using the cdf for DEA

DEA: Does the gene expression Y differs according to a (group of) factor(s) X ?

$$H_0 : Y \perp X$$

Using the cdf for DEA

DEA: Does the gene expression Y differs according to a (group of) factor(s) X ?

$$H_0 : Y \perp X$$

If a group of factors X is associated with the gene expression Y
⇒ conditional cdf of Y would be significantly ≠ from the marginal cdf:

$$H_0 : F_{Y|X}(y, x) = F_Y(y)$$

Using the cdf for DEA

DEA: Does the gene expression Y differs according to a (group of) factor(s) X, given Z ?

$$H_0 : Y \perp X \mid Z$$

If a group of factors X is associated with the gene expression Y, given Z
⇒ conditional cdf of Y would be significantly ≠ from the marginal cdf:

$$H_0 : F_{Y|X,Z}(y, x, z) = F_{Y|Z}(y, z)$$

Estimating the empirical CDF with linear regressions

The conditional CDF of Y given X and Z is:

$$F_{Y|X,Z}(y \,|\, x, z) = \mathbb{P}(Y \leq y \,|\, X = x, Z = z) = \mathbb{E}(\mathbb{1}_{\{Y \leq y\}} \,|\, X = x, Z = z)$$

Estimating the empirical CDF with linear regressions

The conditional CDF of Y given X and Z is:

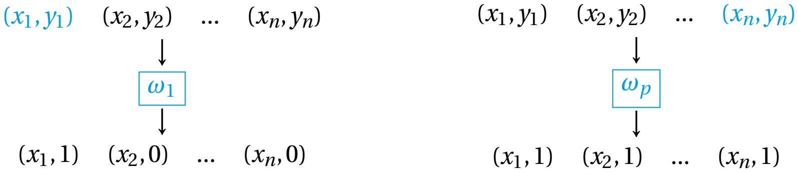
$$F_{Y|X,Z}(y\,|\,x,z) = \mathbb{P}(Y \leq y\,|\,X = x, Z = z) = \mathbb{E}(\mathbb{1}_{\{Y\leq y\}}\,|\,X = x, Z = z)$$

For a given gene g and for a sequence of p ordered thresholds $\omega_1, \dots, \omega_p$:

$$\mathbb{E}\left(\mathbb{1}_{\{y_i\leq\omega_j\}}|X = x_i, Z = z_i\right) = \beta_{0j} + \beta_{1j}x_i + \beta_{2j}z_i, \quad \forall i = 1, \dots, n$$

Estimating empirical CDFs with multiple linear regressions

CDF estimation with p linear regressions:



$(x_1, y_1) \quad (x_2, y_2) \quad \dots \quad (x_n, y_n)$

$\downarrow$

$\boxed{\omega_1}$

$\downarrow$

$(x_1, 1) \quad (x_2, 0) \quad \dots \quad (x_n, 0)$

$(x_1, y_1) \quad (x_2, y_2) \quad \dots \quad (x_n, y_n)$

$\downarrow$

$\boxed{\omega_p}$

$\downarrow$

$(x_1, 1) \quad (x_2, 1) \quad \dots \quad (x_n, 1)$

Estimating empirical CDFs with multiple linear regressions

CDF estimation with p linear regressions:

$$\begin{tikzcd}
(x_1, y_1) & (x_2, y_2) \arrow[d] & \ldots & (x_n, y_n) \\
 & \boxed{\omega_1} \arrow[d] & & \\
(x_1, 1) & (x_2, 0) \arrow[d] & \ldots & (x_n, 0) \\
 & \mathbb{E}(\mathbb{1}_{\{Y \leq \omega_1\}} \mid X) = \widehat{F}(\omega_1 \mid X) & &
\end{tikzcd}$$

$$\begin{tikzcd}
(x_1, y_1) & (x_2, y_2) \arrow[d] & \ldots & (x_n, y_n) \\
 & \boxed{\omega_p} \arrow[d] & & \\
(x_1, 1) & (x_2, 1) \arrow[d] & \ldots & (x_n, 1) \\
 & \mathbb{E}(\mathbb{1}_{\{Y \leq \omega_p\}} \mid X) = \widehat{F}(\omega_p \mid X) & &
\end{tikzcd}$$

Estimating empirical CDFs with multiple linear regressions

CDF estimation with p linear regressions:

$$\begin{tikzcd}
(x_1, y_1) \quad (x_2, y_2) \quad \dots \quad (x_n, y_n) \arrow[d] \\
\boxed{\omega_1} \arrow[d] \\
(x_1, 1) \quad (x_2, 0) \quad \dots \quad (x_n, 0) \arrow[d] \\
\mathbb{E}(\mathbb{1}_{\{Y \leq \omega_1\}} \mid X) = \widehat{F}(\omega_1 \mid X) \arrow[d] \\
\boxed{X = x_1} \arrow[d] \\
\mathbb{E}(\mathbb{1}_{\{Y \leq \omega_1\}} \mid X = x_1) = \widehat{F}(\omega_1 \mid x_1)
\end{tikzcd}$$

$$\begin{tikzcd}
(x_1, y_1) \quad (x_2, y_2) \quad \dots \quad (x_n, y_n) \arrow[d] \\
\boxed{\omega_p} \arrow[d] \\
(x_1, 1) \quad (x_2, 1) \quad \dots \quad (x_n, 1) \arrow[d] \\
\mathbb{E}(\mathbb{1}_{\{Y \leq \omega_p\}} \mid X) = \widehat{F}(\omega_p \mid X) \arrow[d] \\
\boxed{X = x_n} \arrow[d] \\
\mathbb{E}(\mathbb{1}_{\{Y \leq \omega_p\}} \mid X = x_n) = \widehat{F}(\omega_p \mid x_n)
\end{tikzcd}$$

Asymptotic test

$$H_0 : \beta_{1j} = 0, j = 1, \ldots, p$$

β_{1j}: coefficient for X in the CDF estimating regression in ω_j

Asymptotic test

$$H_0 : \beta_{1j} = 0, j = 1, ..., p$$

β_{1j}: coefficient for X in the CDF estimating regression in ω_j

Consider the following test statistic:

$$D_n = n \sum_{j=1}^{p} \beta_{1j}^2$$

It converges to a mixture of χ^2:

$$\widehat{D}_n \quad \xrightarrow[n \to +\infty]{} \quad \sum_{j=1}^{p} \widehat{a}_j \chi_1^2$$

⇒ Benjamini-Hochberg correction for multiple testing

Permutation test

Under H_0, observations of X are exchangeable for a given Y

1. **No covariates Z**

   B random permutations ⇒ B test statistics: $\mathscr{D} = \{D_1^*, ..., D_B^*\} \sim H_0$

   ⇒ p-value estimate: $\frac{1}{1+B}\left(1 + \sum_{b=1}^{B} \mathbb{1}_{\left\{\widehat{D} \leq D_b^*\right\}}\right)$ **[Phipson & Smyth, 2010]**

2. **With covariates Z**

   x_i exchangeable conditional on Z

   - Z discrete: stratification
   - Z continuous: conditional on distances between observations of Z

‘

⇒ Benjamini-Hochberg correction for multiple testing

Practical considerations for computational speed up

- **Adaptive permutations**
  1. Start by a **small** number of permutations (e.g. 100)
  2. **Increase the number of permutations** (e.g. to 250) **only for genes with low p-values** (e.g. < 0.1) for which additional numerical precision is needed
  3. Repeat step 2 with decreasing p-value threshold (e.g. 0.05 and then 0.01) to **reach large number of permutations only for a limited number of genes**

- **Spaced thresholds**

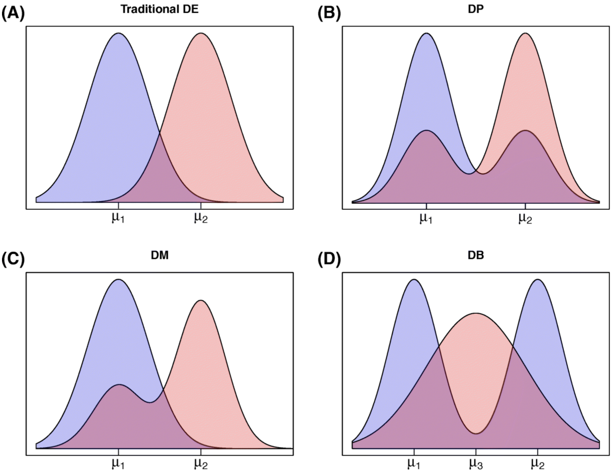
  ⇒ as many ω_j possible as unique values y_i
  less thresholds: speed vs numerical precision

- **OLS**

  ⇒ estimations of $\widehat{\beta}_{1j}$s

Results

2 group comparison benchmark with state-of-the-art

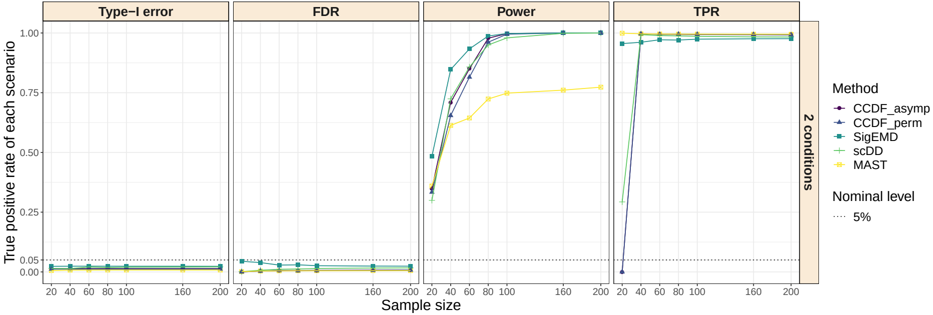


[source: Korthauer et al. (2016)]

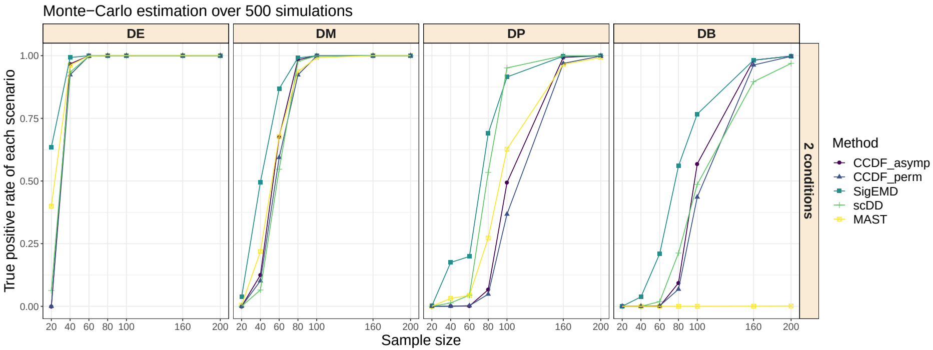
The two conditions case

Monte−Carlo estimation over 500 simulations

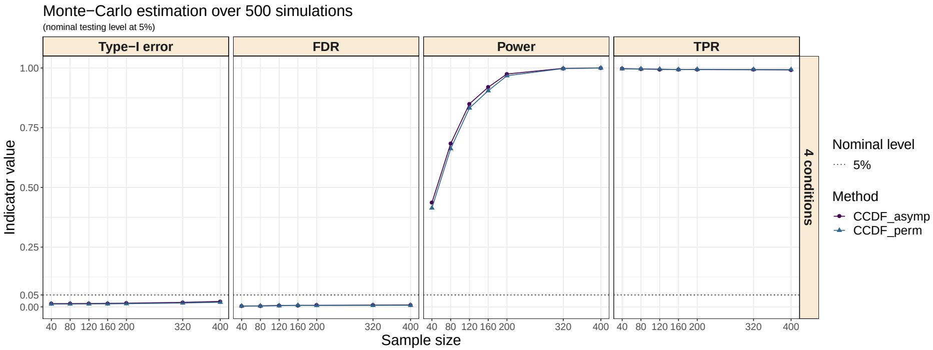
(nominal testing level at 5%)



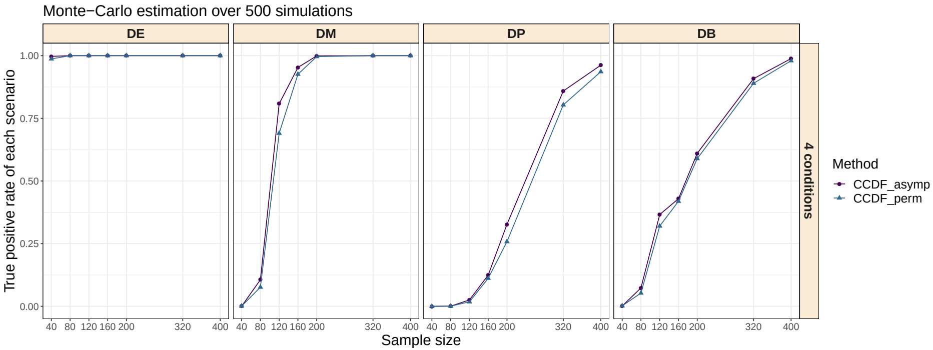
The two conditions case – DE genes breakdown



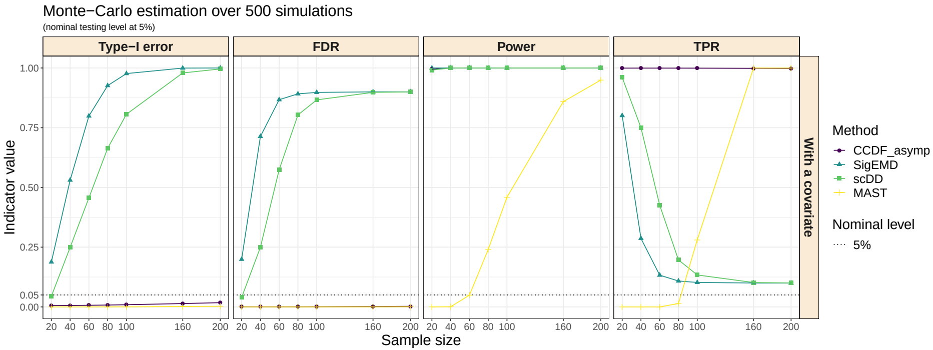
Multiple comparisons: 4 conditions



Multiple comparisons: 4 conditions – DE genes breakdown



Two conditions comparison given a confounding covariate Z



Positive control real dataset

Islam et al. (2011) dataset: 22,928 genes from 48 mouse embryonic stem cells and 44 mouse embryonic fibroblasts

⇒ Positive control dataset
⇒ Use of the already-published top 1,000 DE genes validated through qRT-PCR experiments as a **gold standard DE gene set**

Positive control real data with FDR of 0.05

Method	**Number of detected DE genes**	**True Positive Rate**
CCDF	7,345	0.696
SigEMD†	3,702	0.488
scDD†	2,638	0.351
MAST†	734	0.198

† results from Wang et al. (2019)

Negative control real dataset

Grün et al.(2014) dataset: 12,535 genes for 80 pool-and-split samples obtained under the same condition

⇒ Negative control data
⇒ Random sampling from the 80 sample to get 10 datasets
⇒ There should be no DE genes

Negative control real data with FDR of 0.05

Method	Number of detected DE genes	False Positive Rate
CCDF	0	0
scDD†	5	0.0007
MAST†	0	0
SigEMD†	50	0.007

† results from Wang et al. (2019)

Motivation: dendritic cells sub-populations characterization

11,985 genes measured for 2,914 single cells across 4 cell populations

Sub-population	**number of cells**
DC1	479
DC2 & DC3	1,526
pDC	297
preDC	612

Motivational data-set analysis results

- Which genes are significantly different according to DC sub-populations ?

  ⇒ 4651 DE genes

- Which genes are significantly associated with one specific biomarker gene expression, adjusted on DC sub-populations ?

  ⇒ 191 DE genes

- Which genes are significantly associated with one specific biomarker gene expression, when DC sub-populations are pooled ?

  ⇒ 619 DE genes

Discussion

Conclusion

Key features

- Competitive statistical power + unique capabilities
- Distribution-free
- Multiple comparisons, complex designs
- Estimate conditional eCDF with multiple regressions
- Asymptotic & permutation tests
- R package `ccdf` available on [`https://github.com/Mgauth/ccdf`]
- `distinct` philosophical proximity

Limits

- Computational burden (currently ∼a few minutes)
- Numerical approximations (OLS, ω_js, permutations, χ^2_1 mixture coefficients. . .)

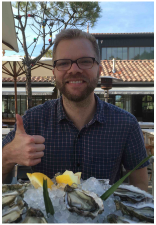
Future work

- Complete the benchmark with all applicable state-of-the-art methods
- Motivational study results biological interpretation
- Multi-sample extension
- Speed-up code
- Perturbations rather than permutations

Thank you for your attention ! – **Questions ?**

Marine Gauthier

Denis Agniel



Rodolphe Thiébaut

Véronique Godot

PhD & postdoc are welcomed !

👉 *boris.hejblum@u-bordeaux.fr*

Two conditions comparison given a confounding covariate Z

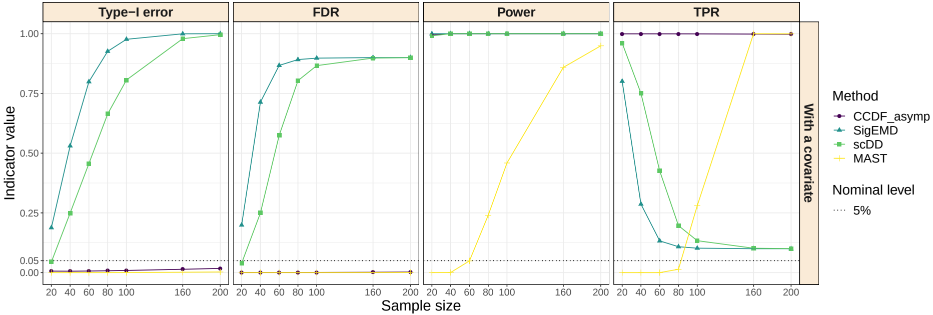
Confounding variable $Z \sim N(10,5)$ with:

$$X = \begin{cases} 1, & Z \leq Q_1 \quad \text{and} \quad Q_2 \leq Z \leq Q_3 \\ 2, & \text{otherwise} \end{cases} \qquad Y = \begin{cases} A * X + \varepsilon_1, & \text{DE gene} \\ 0.3 * Z + \varepsilon_2, & \text{non-DE gene} \end{cases}$$

where Q_p is the p^{th} quartile of Z, $A \sim N(5,1)$, $\varepsilon_1 \sim N(0,1)$, and $\varepsilon_2 \sim N(0,1)$

Monte−Carlo estimation over 500 simulations

(nominal testing level at 5%)



Multiple comparisons: 4 conditions – data generation details

- **multiple DE**: unimodal distributions and single component with a different mean in each condition
- **multiple DP**: bimodal distributions and two components in each condition with equal component means across conditions. The proportion in the first mode is 0.2 for condition 1, 0.4 for condition 2, 0.8 for condition 3, 0.6 for condition 4
- **multiple DM**:
  - distribution with 1 mode for condition 1
  - distribution with 2 modes for condition 2
  - distribution with 3 modes for condition 3
  - distribution with 4 modes for condition 4

  with respectively one, two and three overlapping component(s). Cells belonging to each mode are uniformly distributed.
- **multiple DB**:
  - distribution with 1 mode for condition 1
  - distribution with 2 modes for condition 2
  - distribution with 3 modes for condition 3
  - distribution with 4 modes for condition 4

  The means in condition 2, 3 and 4 are equal to the mean in condition 2. Cells belonging to each mode are uniformly distributed.