Université de Paris

Centre of Research in Epidemiology and Statistics Sorbonne Paris Cité
Inserm UMR 1153

institutCurie
Together, let's beat cancer.

Clustering with missing data : pooling multiple imputation results with consensus clustering

Lilith Faucheux, Matthieu Resche-Rigon, Emmanuel Curis, Vassili Soumelis, Sylvie Chevret

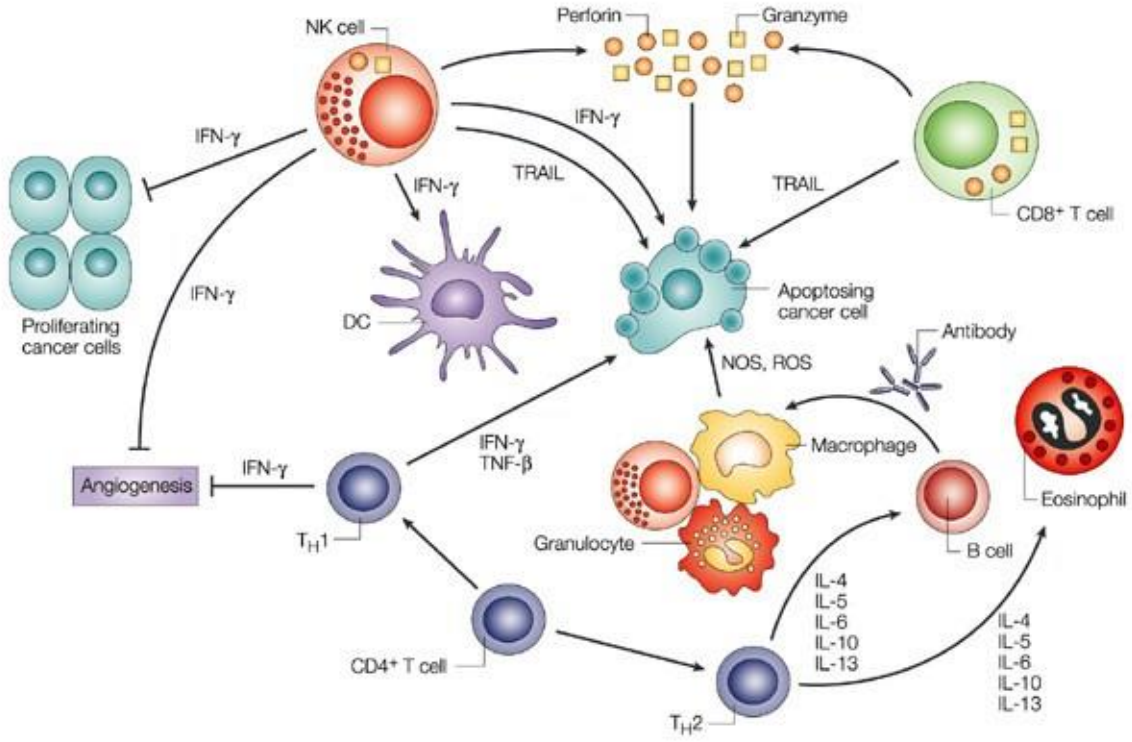
PhD student at ECSTRRA team, UMR1153, Paris, France

Journées Statistiques et Santé

Context

Classification of breast tumors according to their immune Tumor MicroEnvironment

Analysis of soluble proteins produced and identified by immune cells



Nature Reviews | Cancer

Dranoff G., *Nature Reviews Cancer,* 2004

Issue : Missing data and left-censored data

N = 420 patients

Objective

Evaluate performance of clustering procedures using multiple imputation, on datasets with missing and left-censored data

Objective

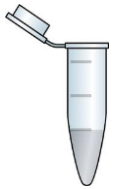
Evaluate performance of clustering procedures using multiple imputation, on datasets with missing and left-censored data

Outline

- Simulation framework
- Clustering procedures using MI
- Performances on simulations

Simulation framework

Experimental data

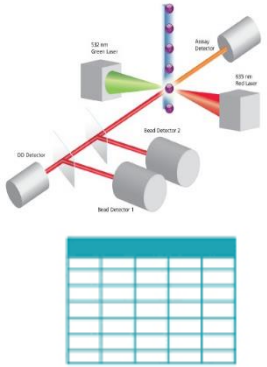


Concentration ?

Assays pipeline

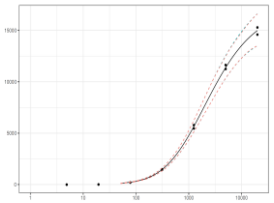
Standard curve fitting

C1
C2
…
C7



Fluorescence (FI)

$$FI = F(concentration)$$



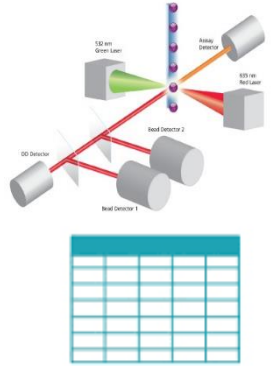
Parameter estimation

$\widehat{F}$

Sample concentration estimation

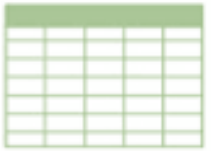
C ?

$\widehat{F}$



Fluorescence (FI)

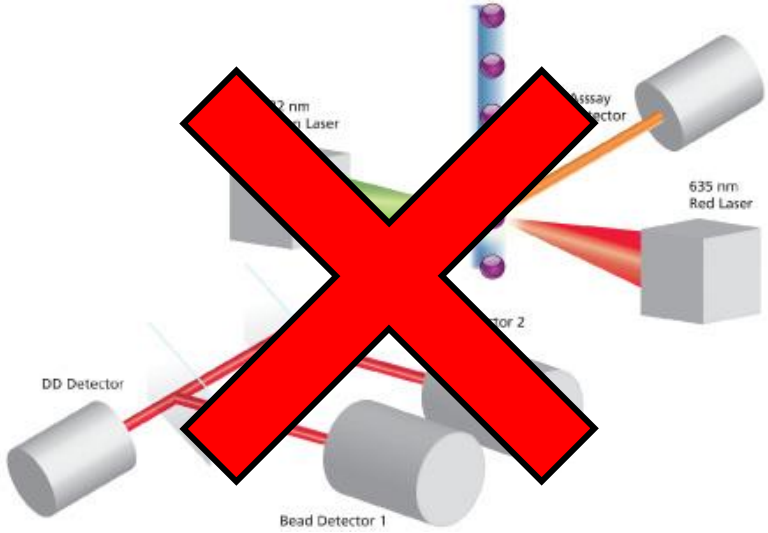
$$\widehat{Concentration} = \hat{F}^{-1}(FI)$$



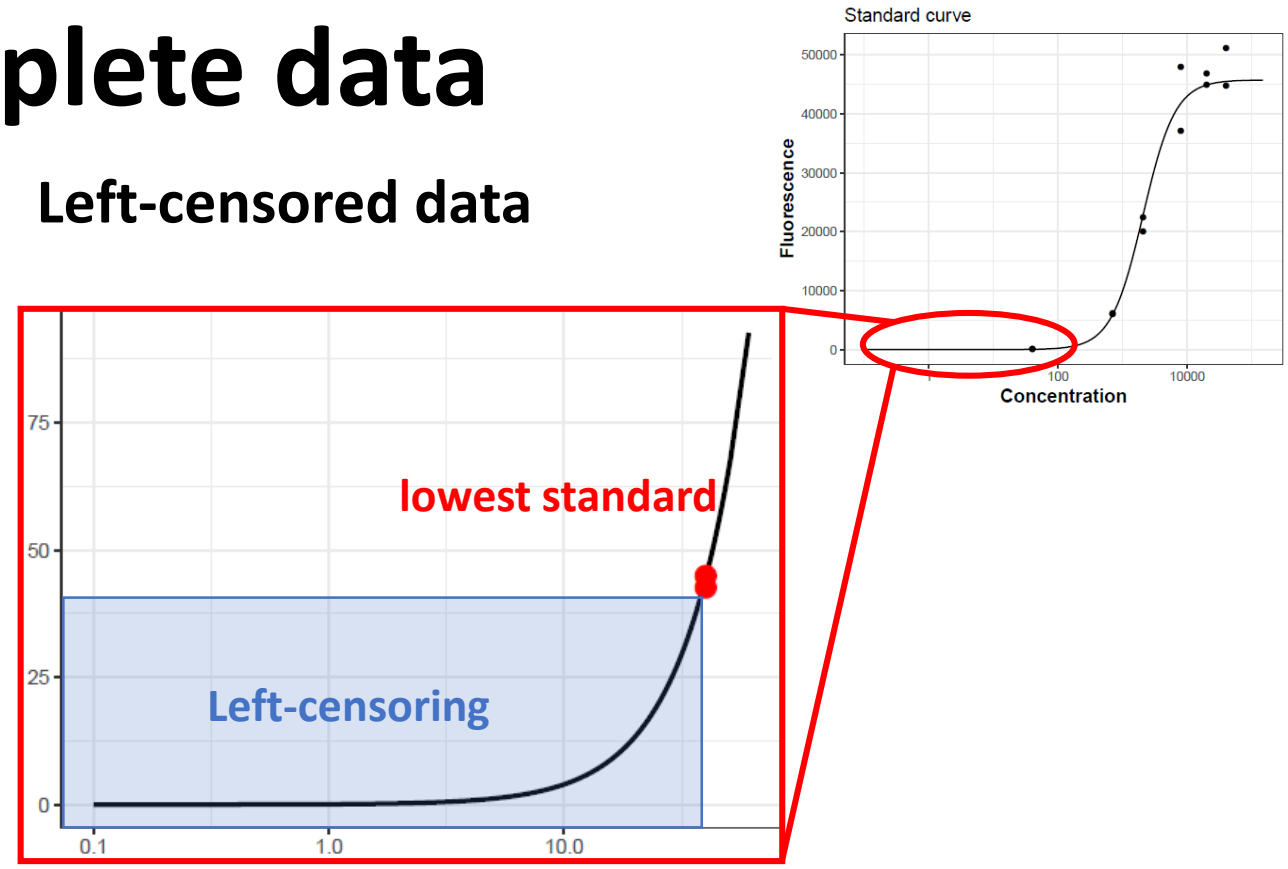
Concentration estimation

Experimental data: Incomplete data

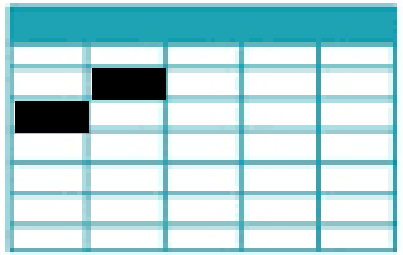
Missing data



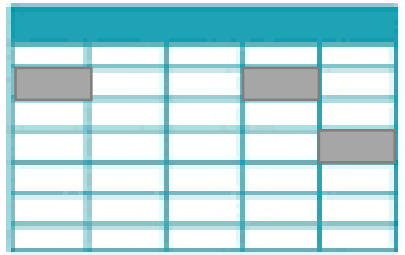
Left-censored data



Lack of fluorescence acquisition



Estimated concentration is lower than the lowest standard



Data simulation

Simulation of samples exact concentration

C

$Log\mathcal{N}(\mu, \Sigma)$

Exact concentration

F

$FI = F(concentration) \times \varepsilon,$
$\varepsilon \sim Log\mathcal{N}(0, 0.1)$

Fluorescence (FI)

Assays pipeline

Curve fitting

C1
C2
…
C7

$FI = F(concentration)$

$\widehat{F}$

Fluorescence (FI)

Parameter estimation

Sample concentration estimation

C ?

$\hat{F}$

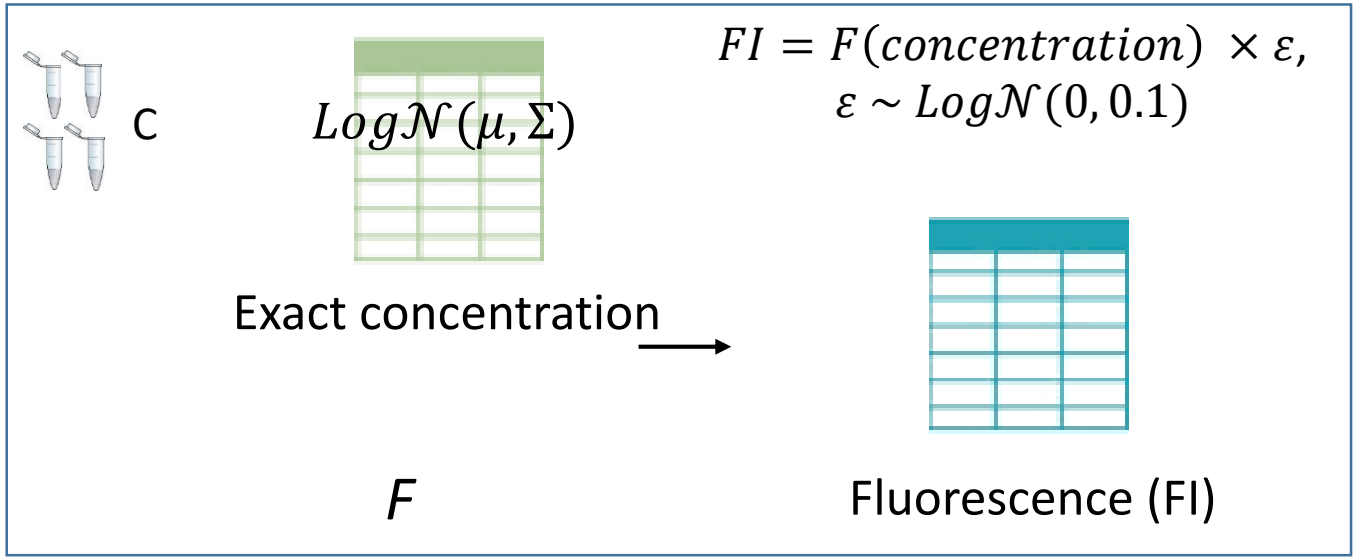
$\widehat{Concentration} = \hat{F}^{-1}(FI)$

Fluorescence (FI)

Concentration estimation

Data simulation

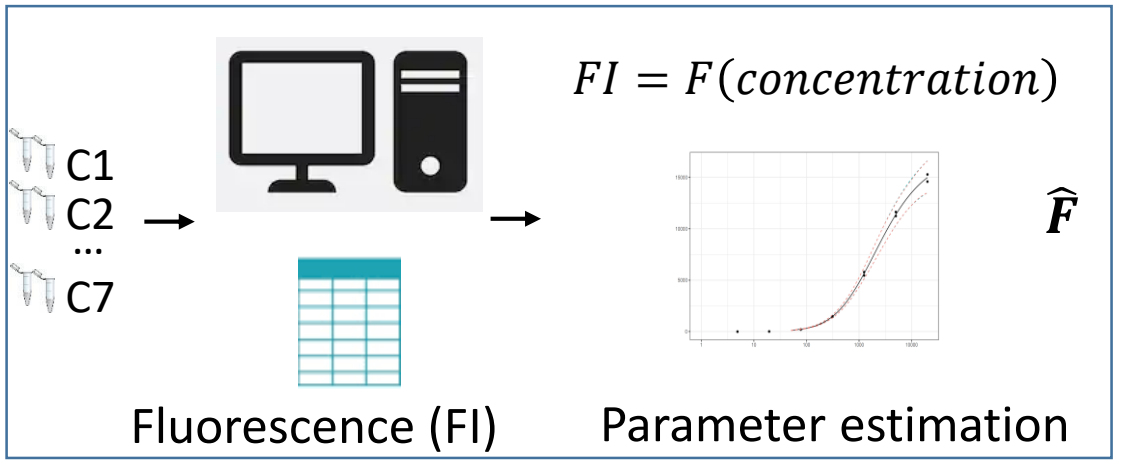
Simulation of samples exact concentration



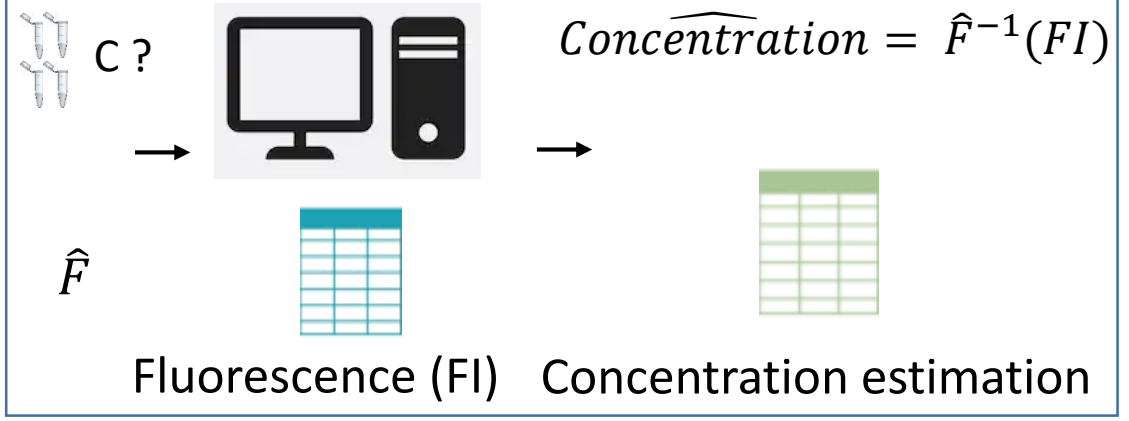
3 variables (X1, X2, X3)
500 observations

Assays pipeline

Curve fitting

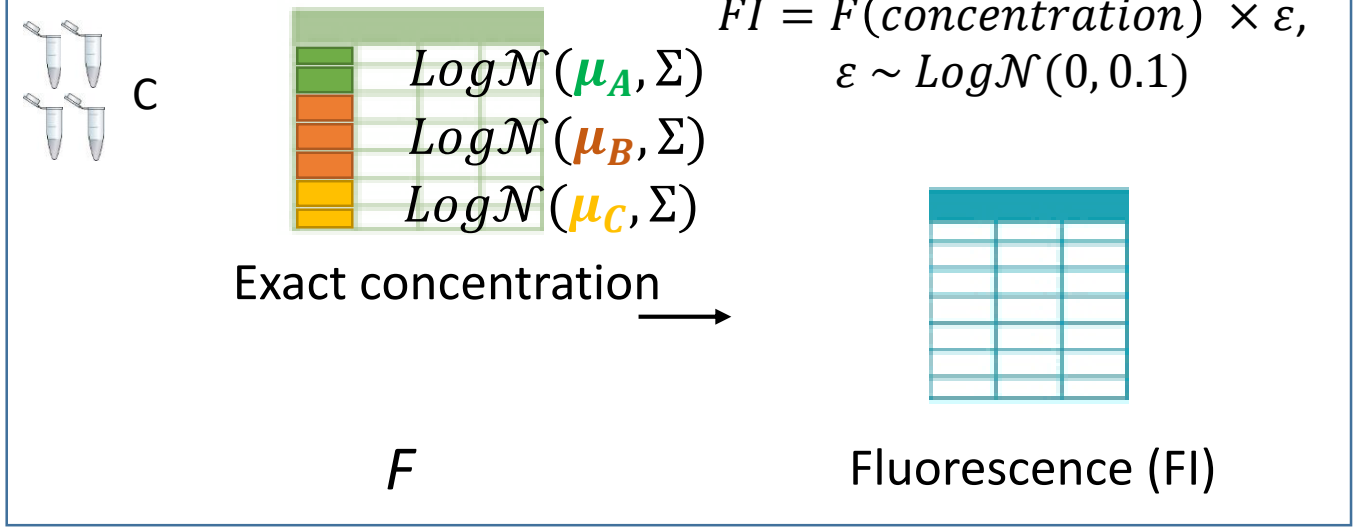


Sample concentration estimation



Data simulation

Simulation of samples exact concentration

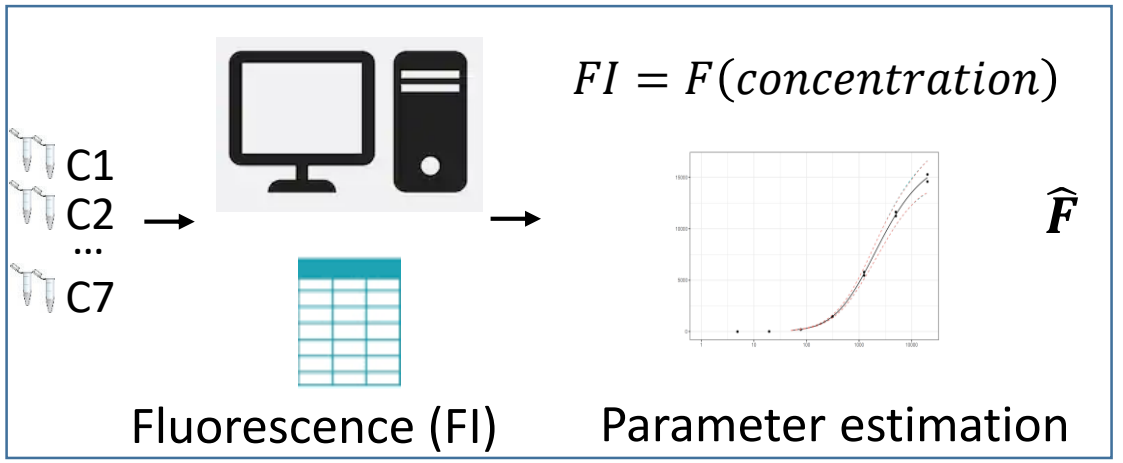


3 variables (X1, X2, X3)
500 observations
3 groups

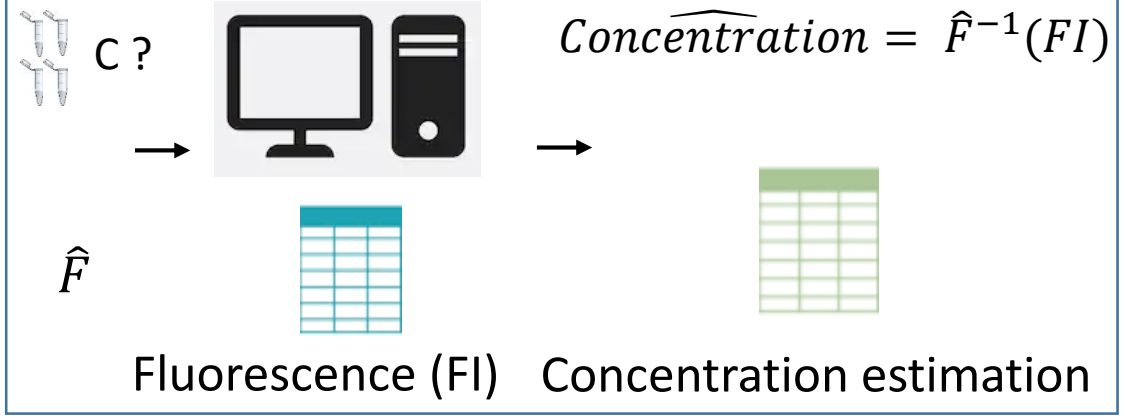
$\Sigma = I(\sigma^2)$

Assays pipeline

Curve fitting

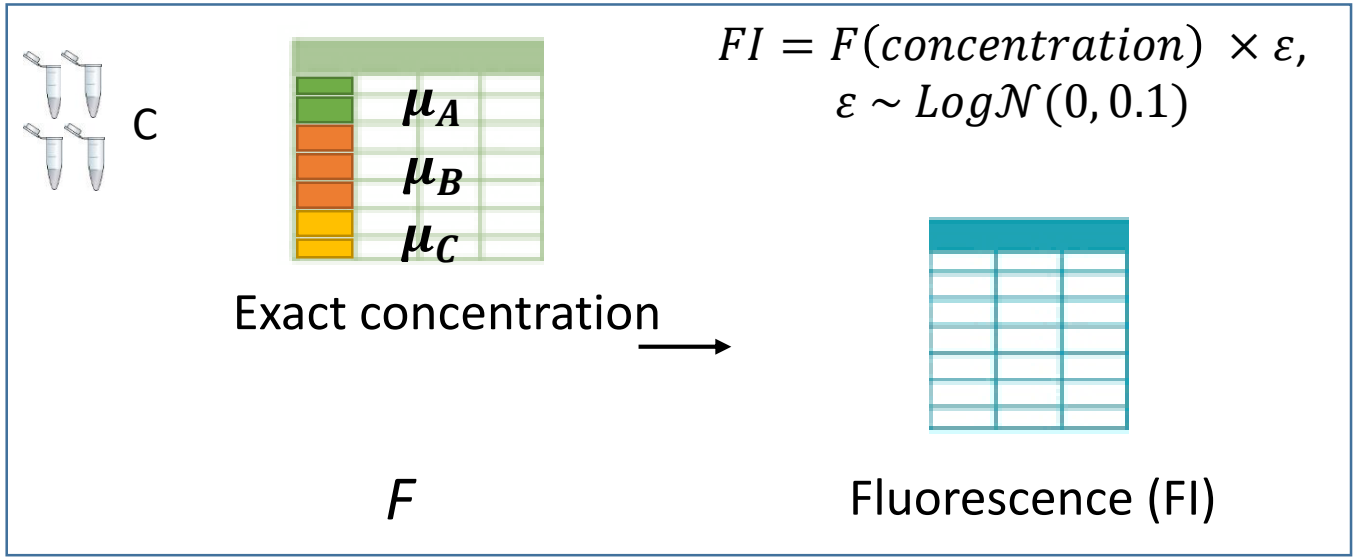


Sample concentration estimation



Data simulation

Simulation of samples exact concentration

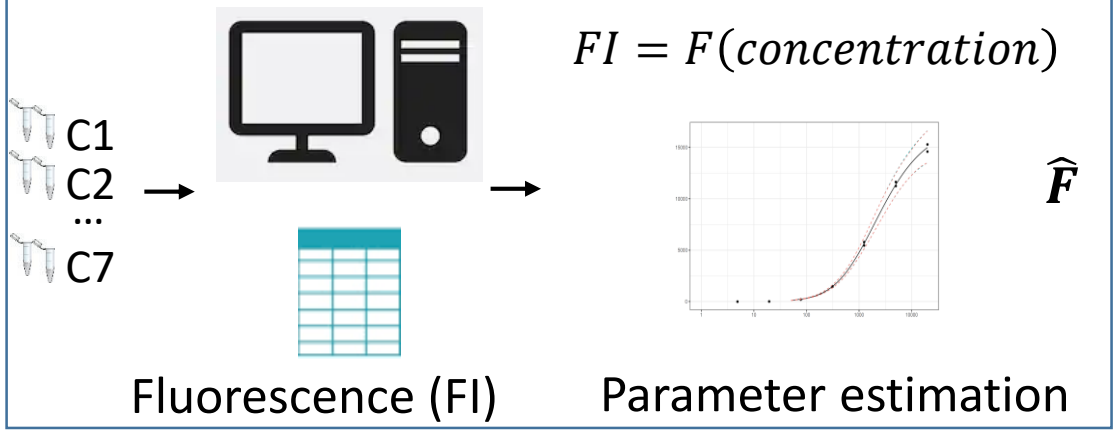


3 variables (X1, X2, X3)
500 observations
3 groups
30% of missing data **on X1**

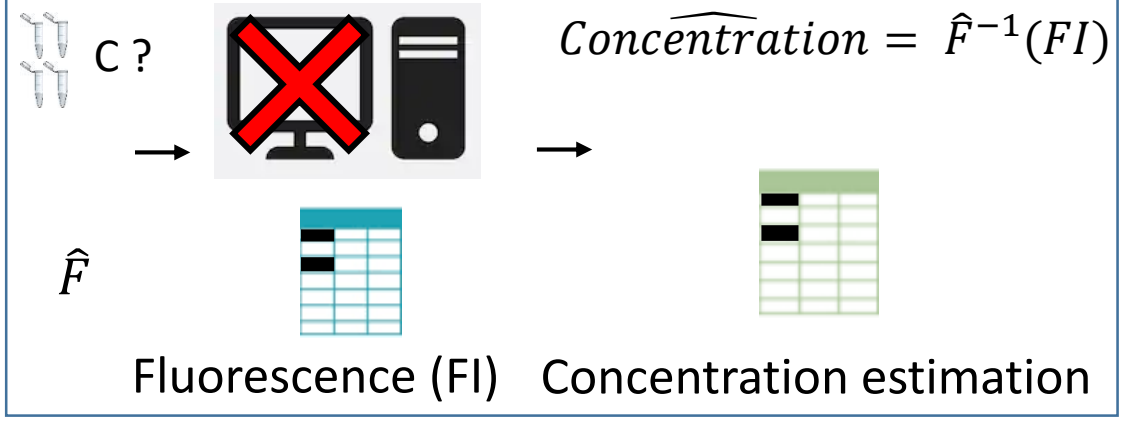
MCAR
MAR (X2)
MNAR (X1)

Assays pipeline

Curve fitting

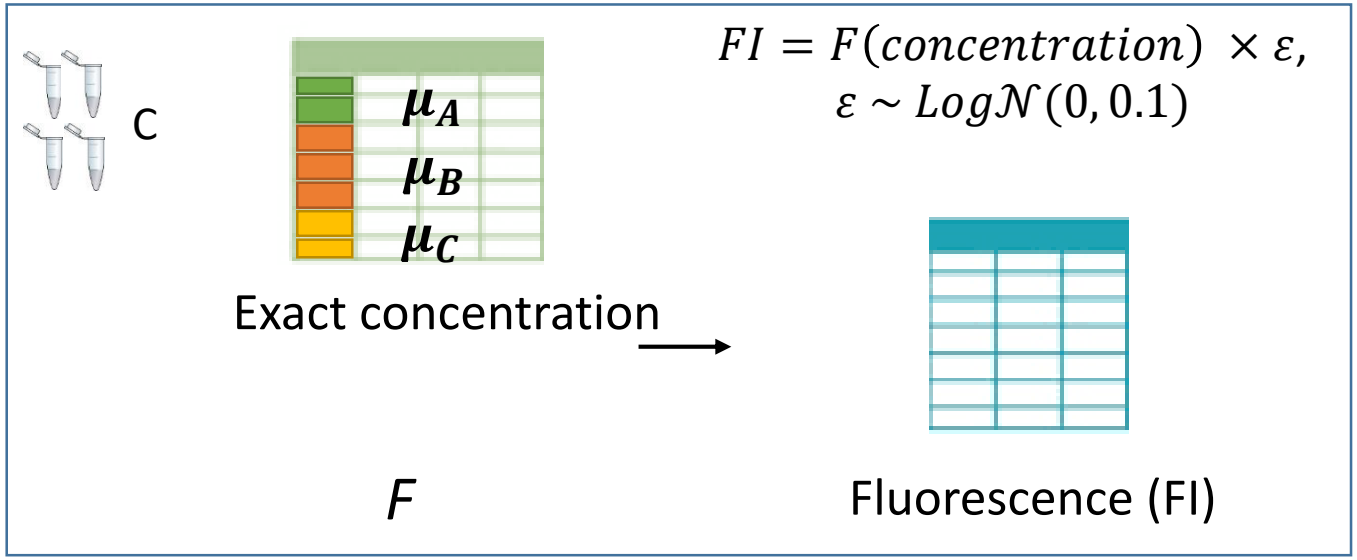


Sample concentration estimation



Data simulation

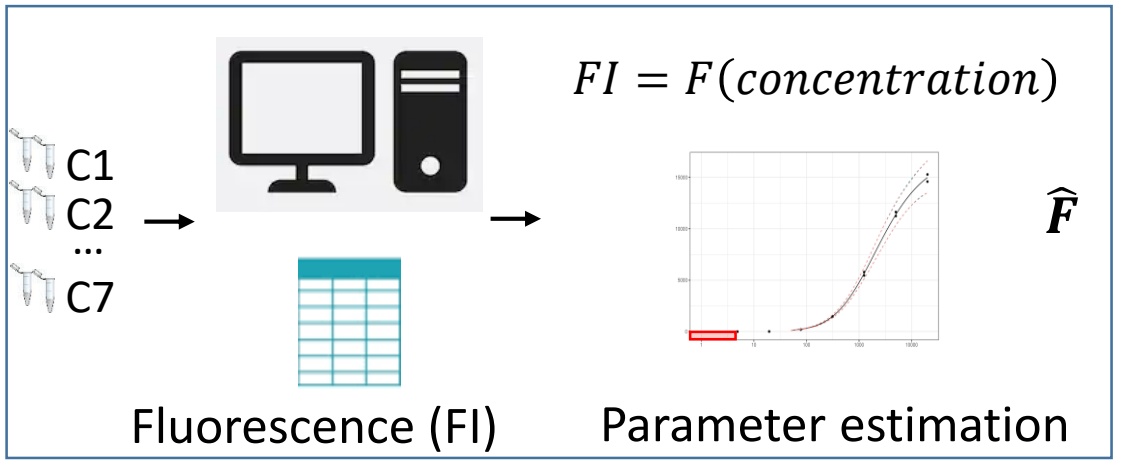
Simulation of samples exact concentration



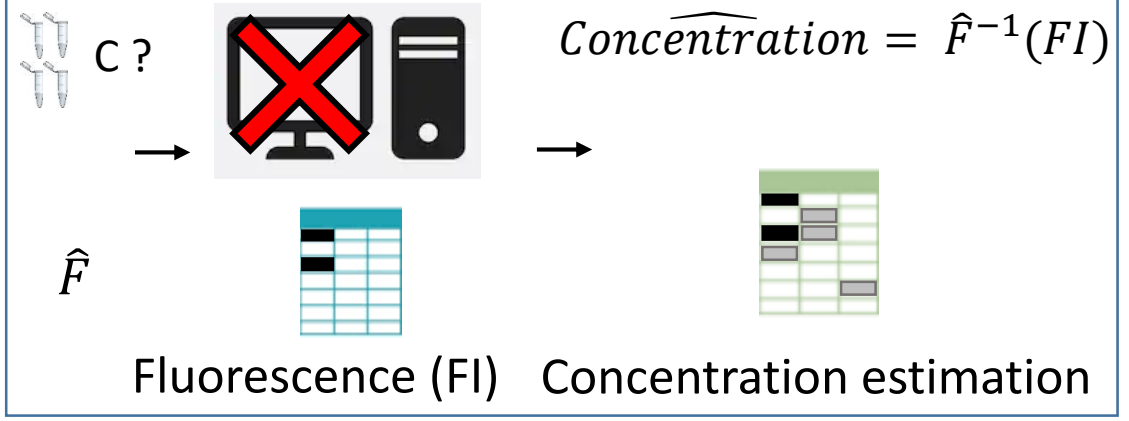
3 variables (X1, X2, X3)
500 observations
3 groups
30% of missing data on X1
Left-censored data

Assays pipeline

Curve fitting



Sample concentration estimation

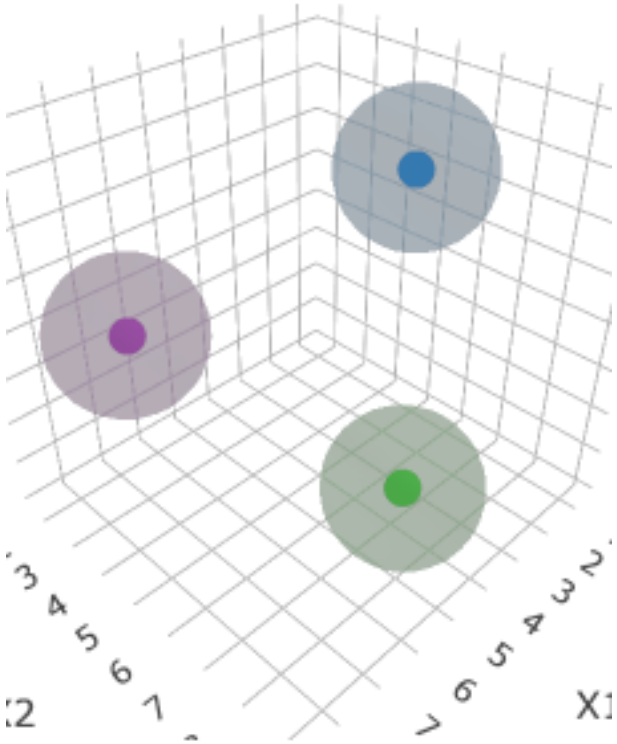


Data simulation : 3 scenarios

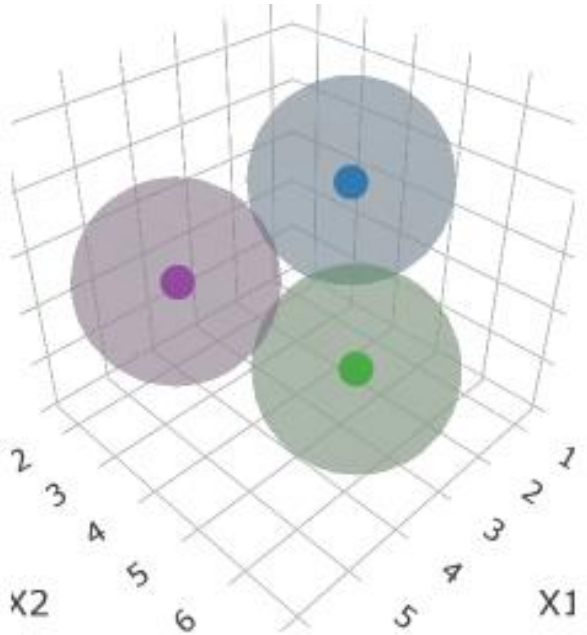
Missing data mechanism
Proportion of left-censored data

Cluster 1
Cluster 2
Cluster 3

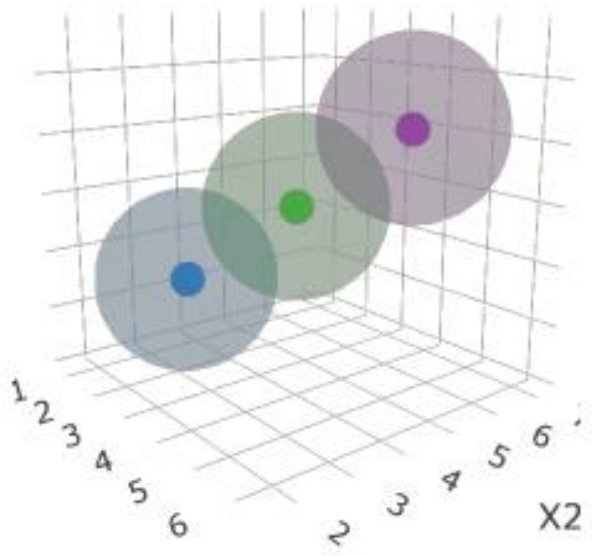
Scenario 1: Large separation



Scenario 2: Intermediate separation



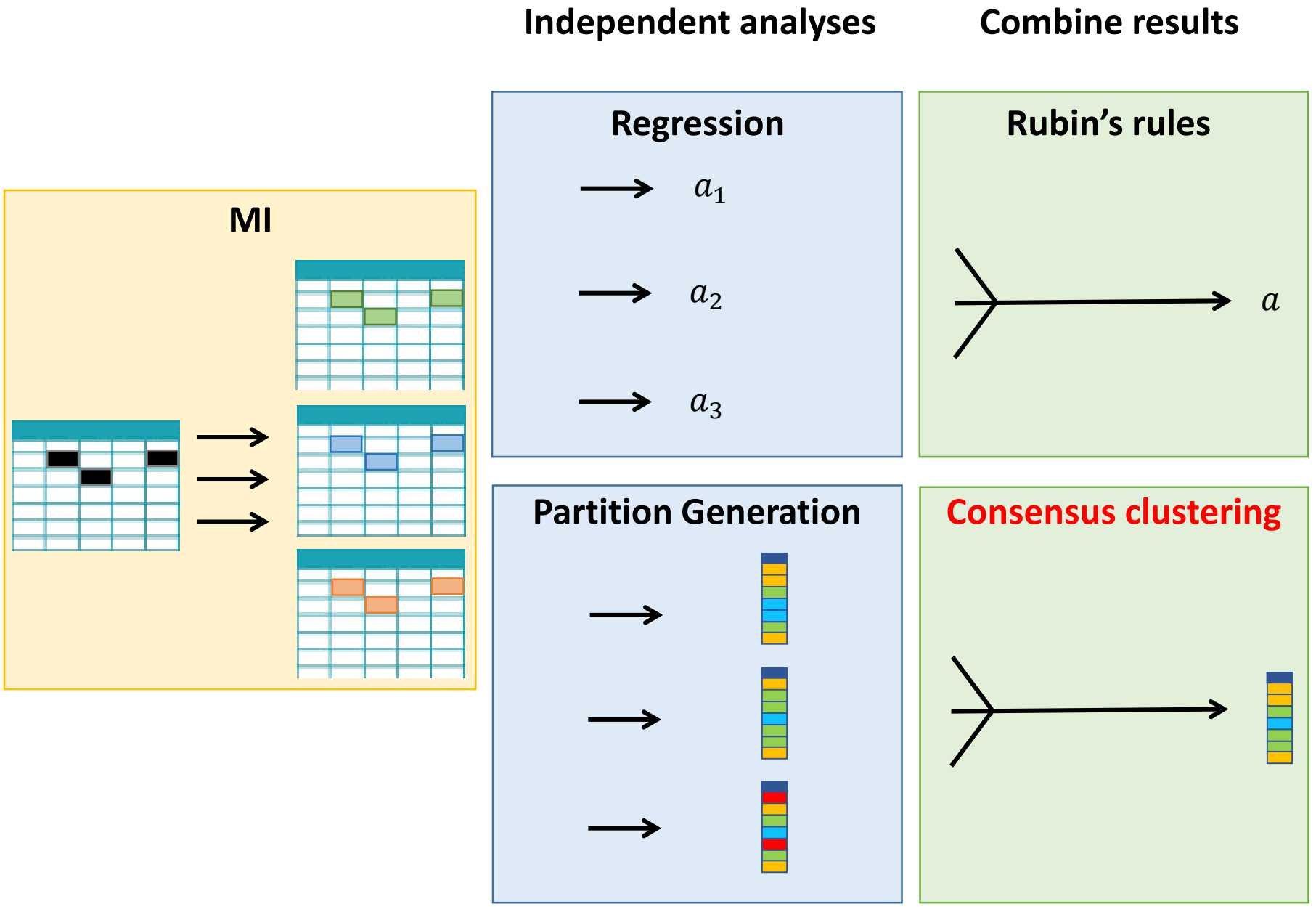
Scenario 3: Diagonal separation



Outline

- Simulation framework
- **Clustering procedures using MI**
- Performances on simulations

Clustering with multiple imputation

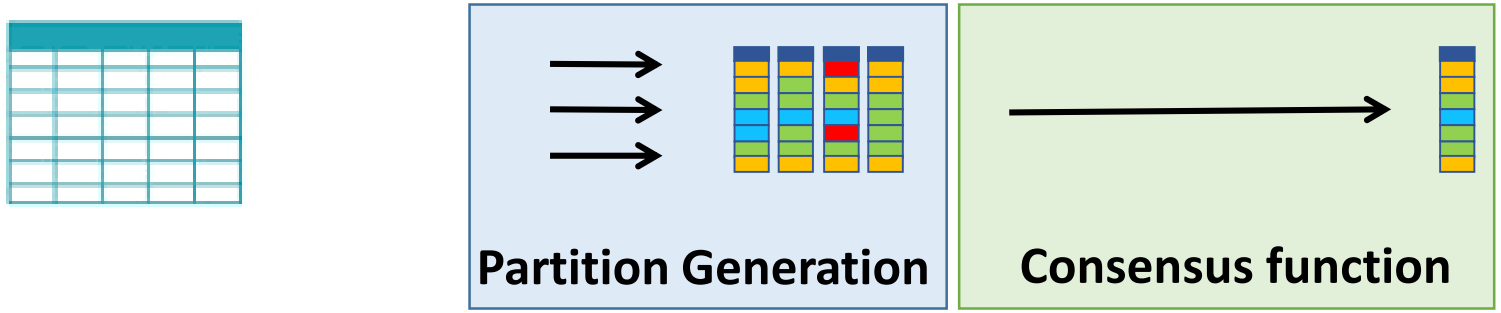


Consensus clustering

Principle:

Combining multiple clustering results to reveal consistency.

Consensus clustering

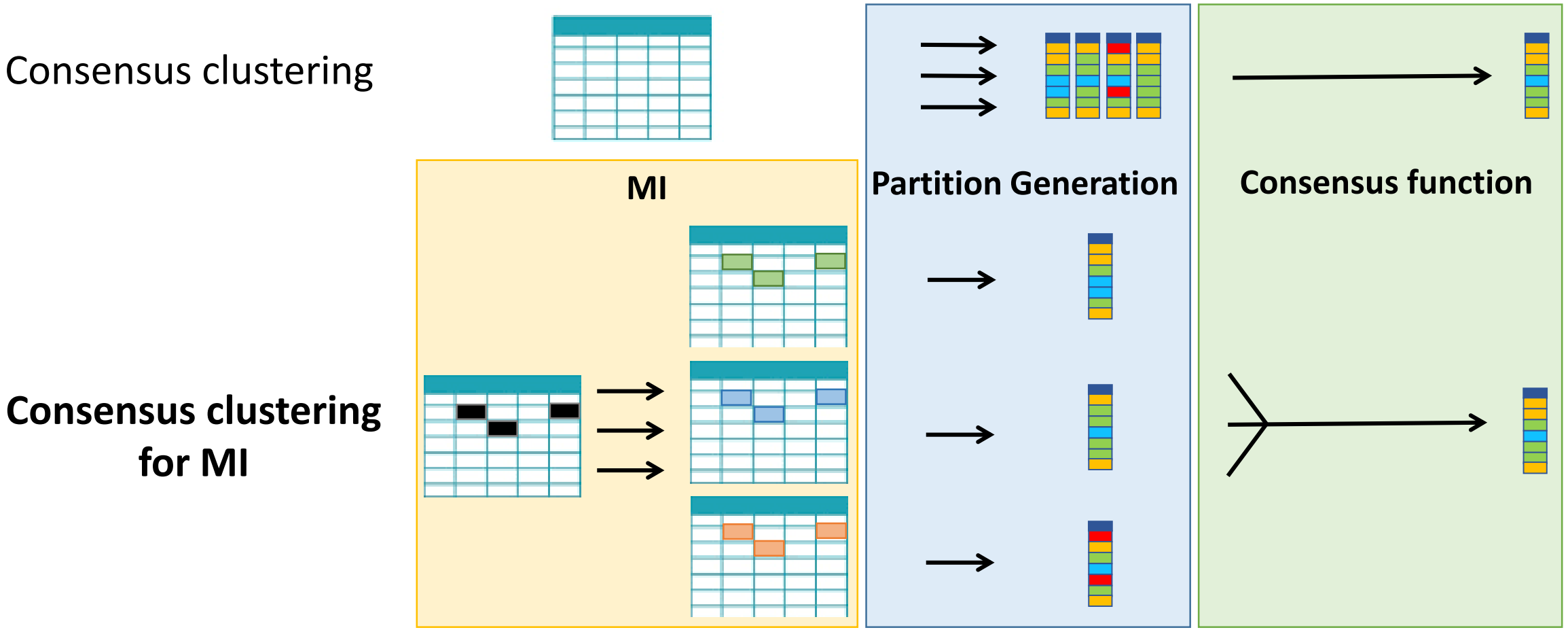


Vega-Pons S. and Ruiz-Shulclopper J., *International Journal of Pattern Recognition and Artificial intelligence*, 2011

Consensus clustering

Principle:

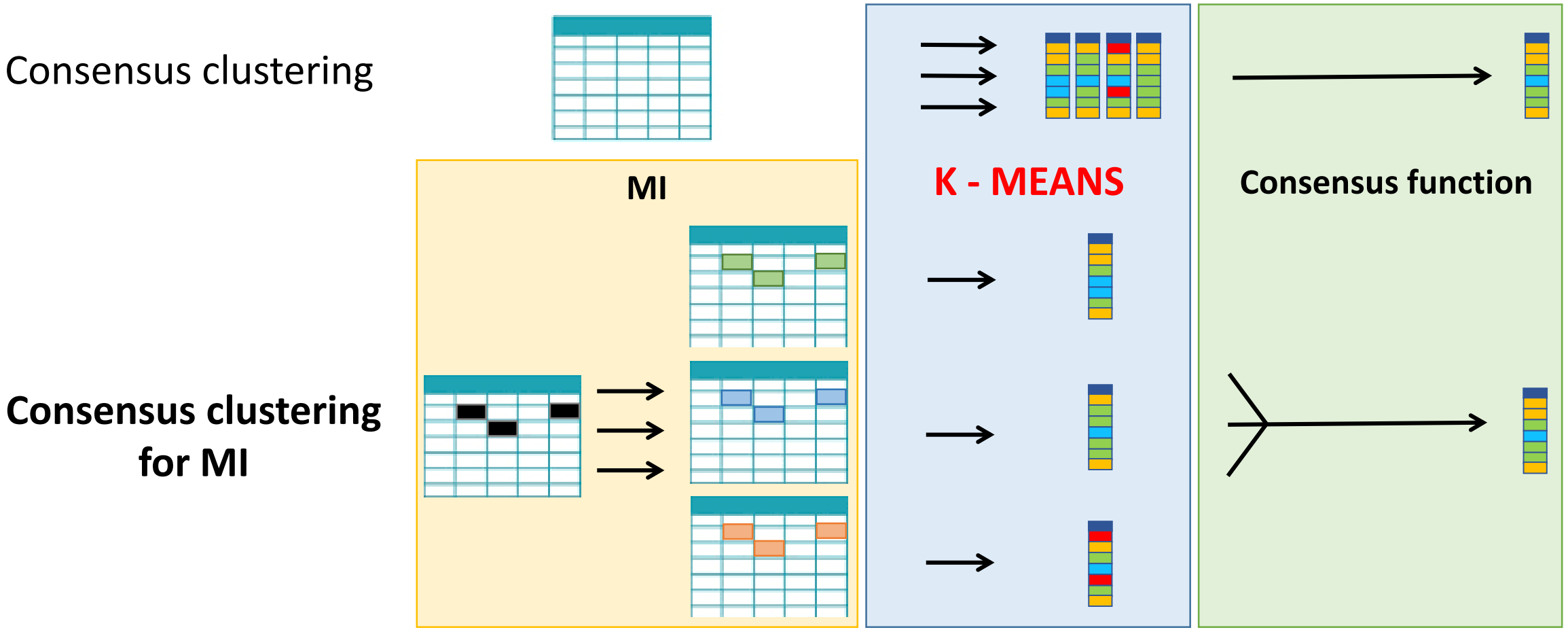
Combining multiple clustering results to reveal consistency.



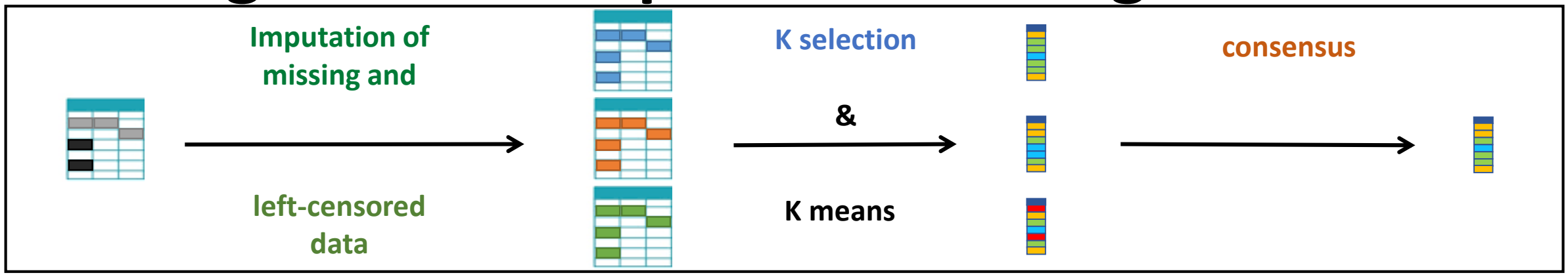
Consensus clustering

Principle:

Combining multiple clustering results to reveal consistency.



Clustering with incomplete data using MI



Methods for missing data
deletion
SSI (mice)
MI (mice)

Methods for left-censored data
SI: 1/2 LOD
SI: standard curve
SSI for left-censored data[1] (mice)
MI for left-censored data[1] (mice)

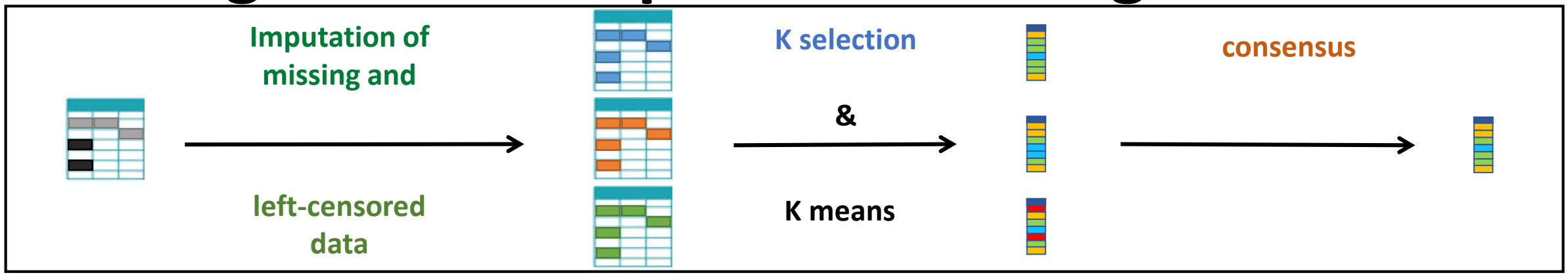
Criteria for K selection
CH[2]
CritCF[3]

Consensus methods
None
Multicons[4]
Combinatorial optimisation (*Bruckers*[5])
Object co-occurence (*Basagaña*[6])

MI = 5 imputations

1. Lapidus N., Chevret S., and Resche-Rigon M., *Statistics in medicine*, 2014
2. Calinski, R. B., and Harabasz, J., *Communications in Statistics*, 1974
3. Breaban M. and Luchian H., *Pattern Recognition*, 2011
4. Al-Najdi, A., Pasquier, N., and Precioso F., *ICAISC*, 2016
5. Bruckers L., Molenberghs G., and Dendale P., *Biometrical Journal*, 2017
6. Basagaña X., *et Al.*, *American Journal of Epidemiology*, 2012

Clustering with incomplete data using MI



Methods for missing data
deletion
SSI (mice)
MI (mice)

Methods for left-censored data
SI: 1/2 LOD
SI: standard curve
SSI for left-censored data[1] (mice)
MI for left-censored data[1] (mice)

Criteria for K selection
CH[2]
CritCF[3]

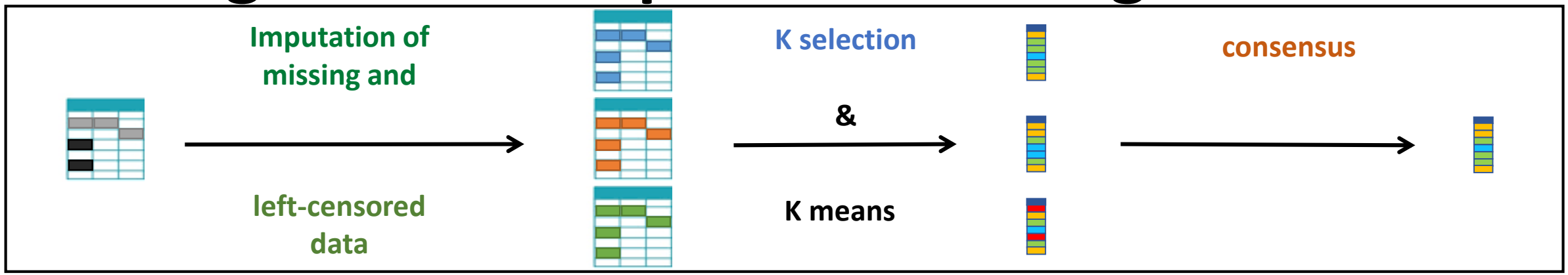
Consensus methods
None
Multicons[4]
Combinatorial optimisation (*Bruckers*[5])
Object co-occurence (*Basagaña*[6])

$$CH = \frac{BSS}{WSS} * \frac{n-k}{k-1}$$

$$CritCF = \left(\frac{2p}{p+1}\frac{1}{1+\frac{W}{B}}\right)^{\frac{log_2(k+1)+1}{log_2(p+1)+1}}$$

MI = 5 imputations

Clustering with incomplete data using MI



Methods for missing data
deletion
SSI (mice)
MI (mice)

Methods for left-censored data
SI: 1/2 LOD
SI: standard curve
SSI for left-censored data[1] (mice)
MI for left-censored data[1] (mice)

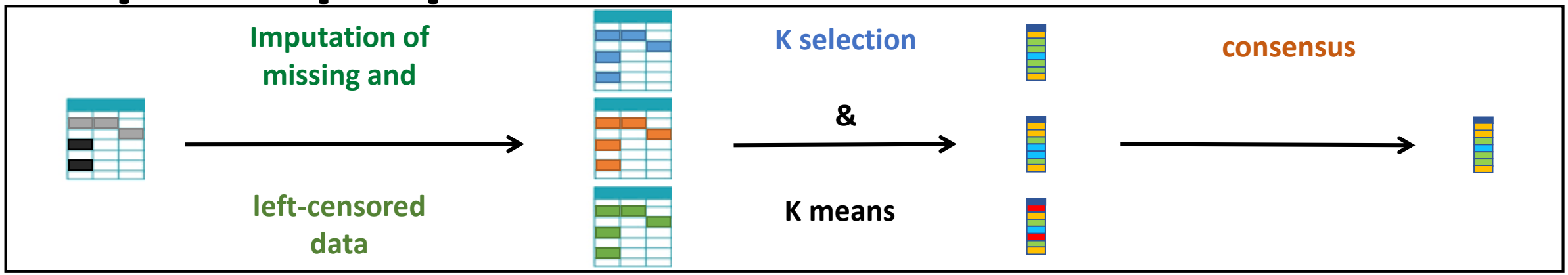
Criteria for K selection
CH[2]
CritCF[3]

Consensus methods
None
Multicons[4]
Combinatorial optimisation (*Bruckers*[5])
Object co-occurence (*Basagaña*[6])

MI = 5 imputations

1. Lapidus N., Chevret S., and Resche-Rigon M., *Statistics in medicine*, 2014
2. Calinski, R. B., and Harabasz, J., *Communications in Statistics*, 1974
3. Breaban M. and Luchian H., *Pattern Recognition*, 2011
4. Al-Najdi, A., Pasquier, N., and Precioso F., *ICAISC*, 2016
5. Bruckers L., Molenberghs G., and Dendale P., *Biometrical Journal*, 2017
6. Basagaña X., *et Al.*, *American Journal of Epidemiology*, 2012

Exemple of proposed method



Methods for missing data
deletion
SSI (mice)
MI (mice)

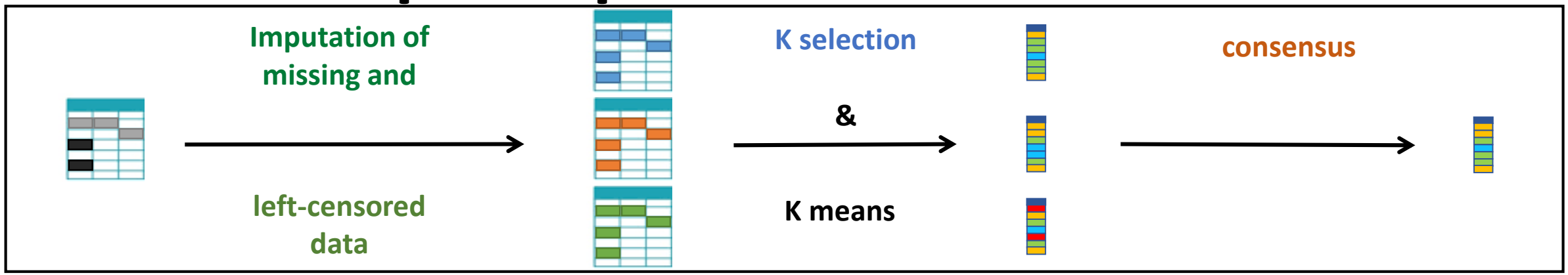
Methods for left-censored data
SI: 1/2 LOD
SI: standard curve
SSI for left-censored data (mice)
MI for left-censored data (mice)

Criteria for K selection
CH
CritCF

Consensus methods
None
Multicons
Combinatorial optimisation (*Bruckers*)
Object co-occurence (*Basagaña*)

MI Multicons CH

Stochastic simple imputation



Methods for missing data
deletion
SSI (mice)
MI (mice)

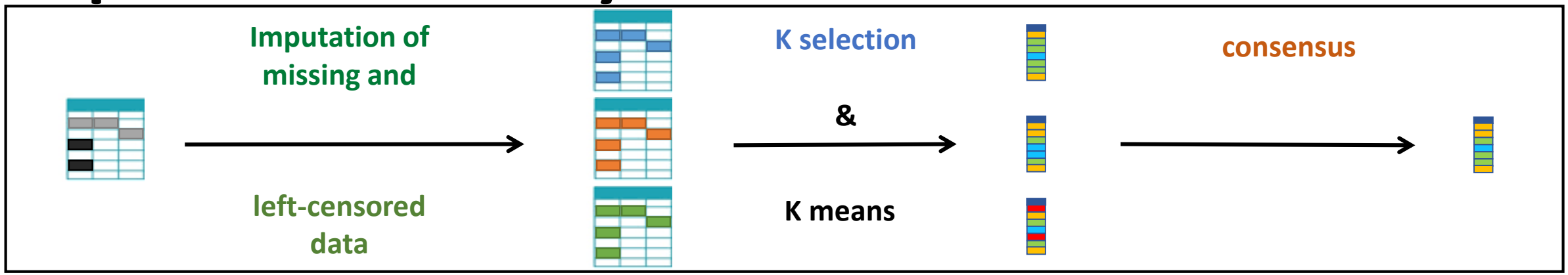
Methods for left-censored data
SI: 1/2 LOD
SI: standard curve
SSI for left-censored data (mice)
MI for left-censored data (mice)

Criteria for K selection
CH
CritCF

Consensus methods
None
Multicons
Combinatorial optimisation (*Bruckers*)
Object co-occurence (*Basagaña*)

SSI CH

Complete cases analysis



Methods for missing data
deletion
SSI (mice)
MI (mice)

Methods for left-censored data
SI: 1/2 LOD
SI: standard curve
SSI for left-censored data (mice)
MI for left-censored data (mice)

Criteria for K selection
CH
CritCF

Consensus methods
None
Multicons
Combinatorial optimisation (*Bruckers*)
Object co-occurence (*Basagaña*)

CCA CH

Outline

- Simulation framework
- Clustering procedures using MI
- **Performances on simulations**

How performance of clustering was evaluated

Cluster number

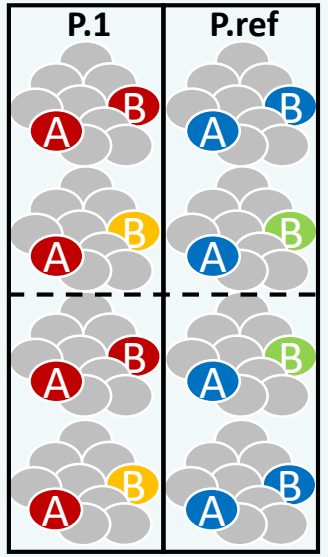
Adjusted Rand Index (ARI)

frequency of agreement over pairs,
adjusted for the chance of grouping elements

External validation criterion:
ARI

$$\mathrm{RI}(\mathrm{P}_1, \mathrm{P}_{\mathrm{ref}}) = \frac{\text{Number of concording pairs}}{\text{Number of pairs}}$$

$$\mathbf{ARI}(\mathbf{P_1}, \mathbf{P_{ref}}) = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00} + N_{01})(N_{01} + N_{11}) + (N_{00} + N_{10})(N_{10} + N_{11})}$$



Concording pair AB

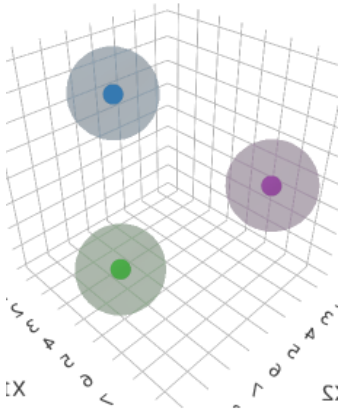
Non concording pair AB

1000 simulations

Performances on complete data (all sceanrios)

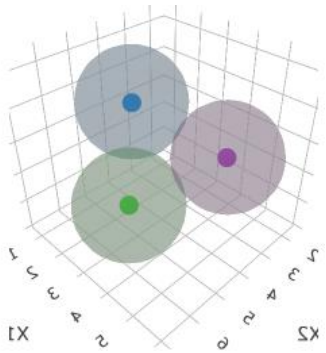
Correct number of clusters : 98.7%

Sc. 1



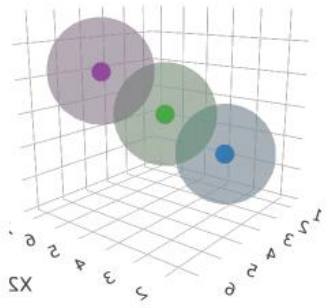
$\hat{k} = 4$: 3.6%

Sc. 2

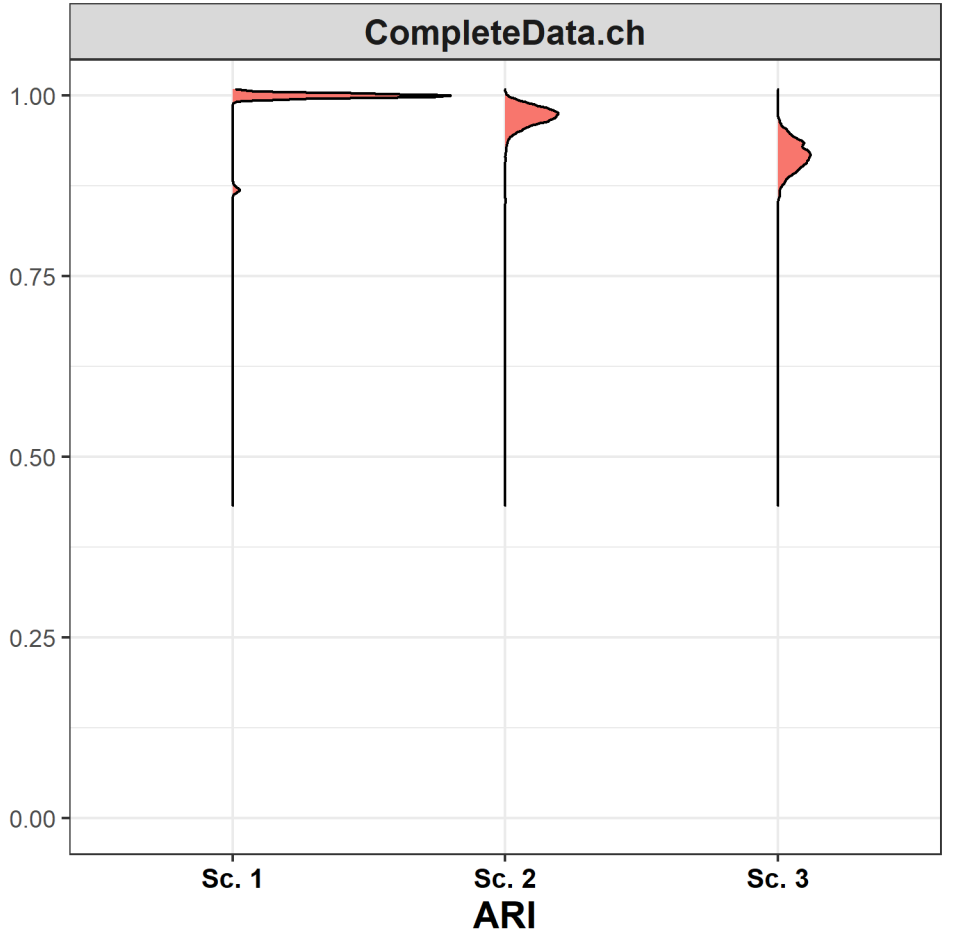


$\hat{k} = 4$: 0.1%

Sc. 3



$\hat{k} = 2$: 0.3%



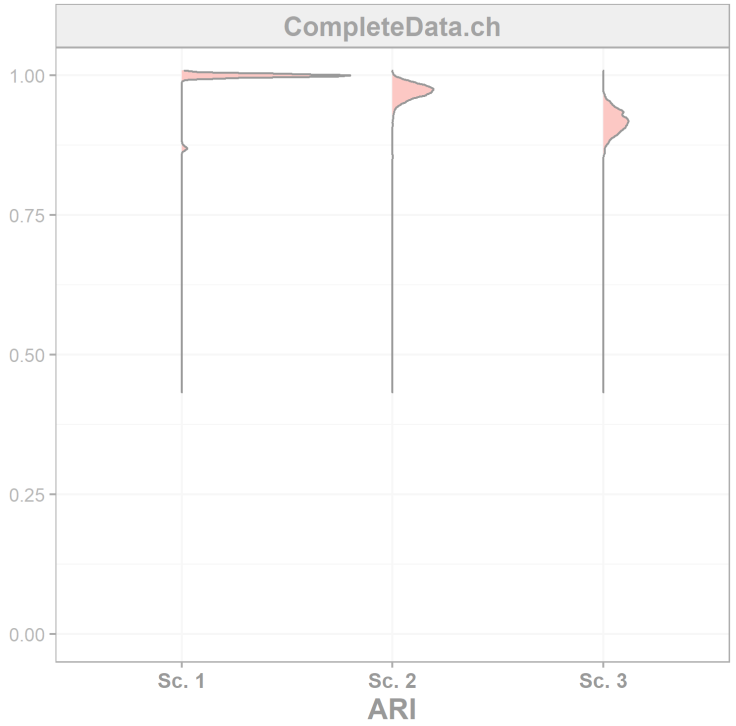
How performance of clustering was also evaluated

Cluster number

Δ Cluster number

Adjusted Rand Index (ARI)

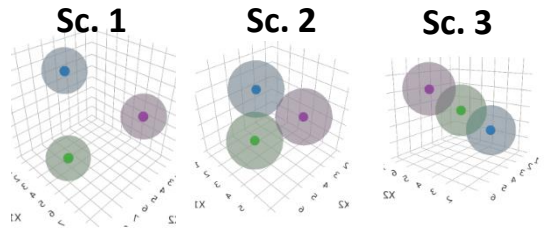
Δ ARI

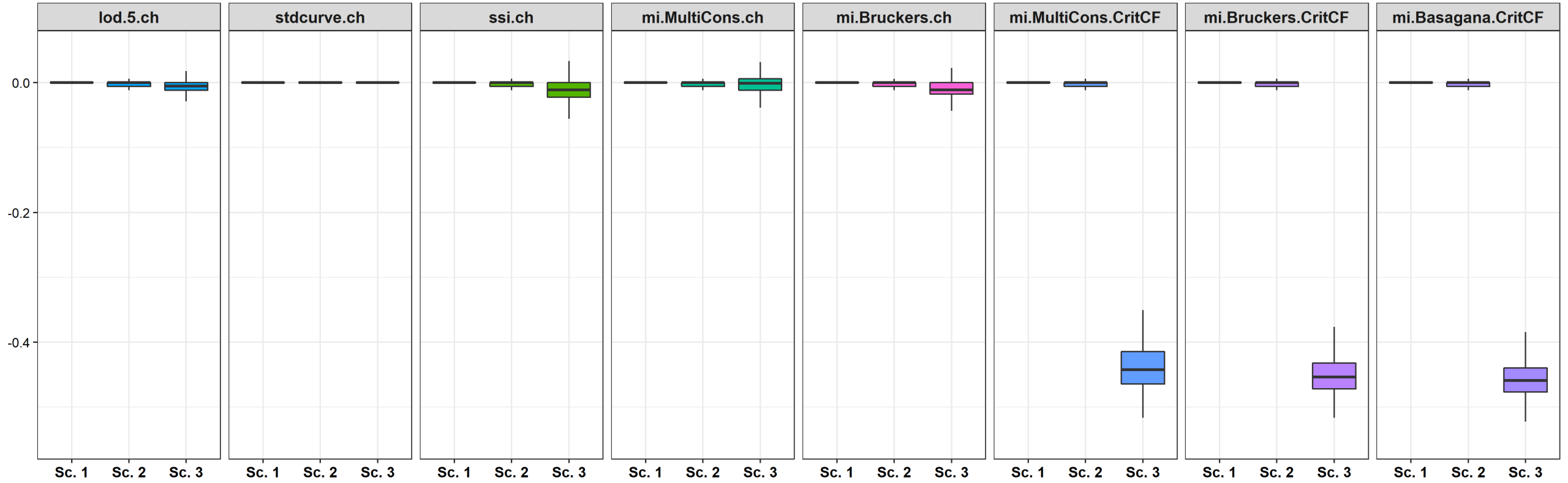


1000 simulations

Left-censored data & no missing data

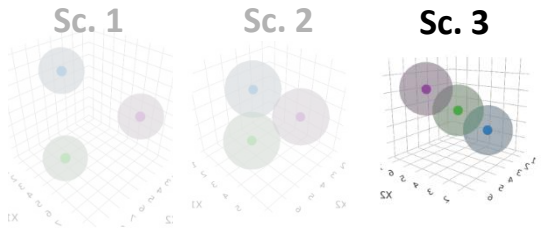
$$ARI - ARI_{Complete\ Data}$$



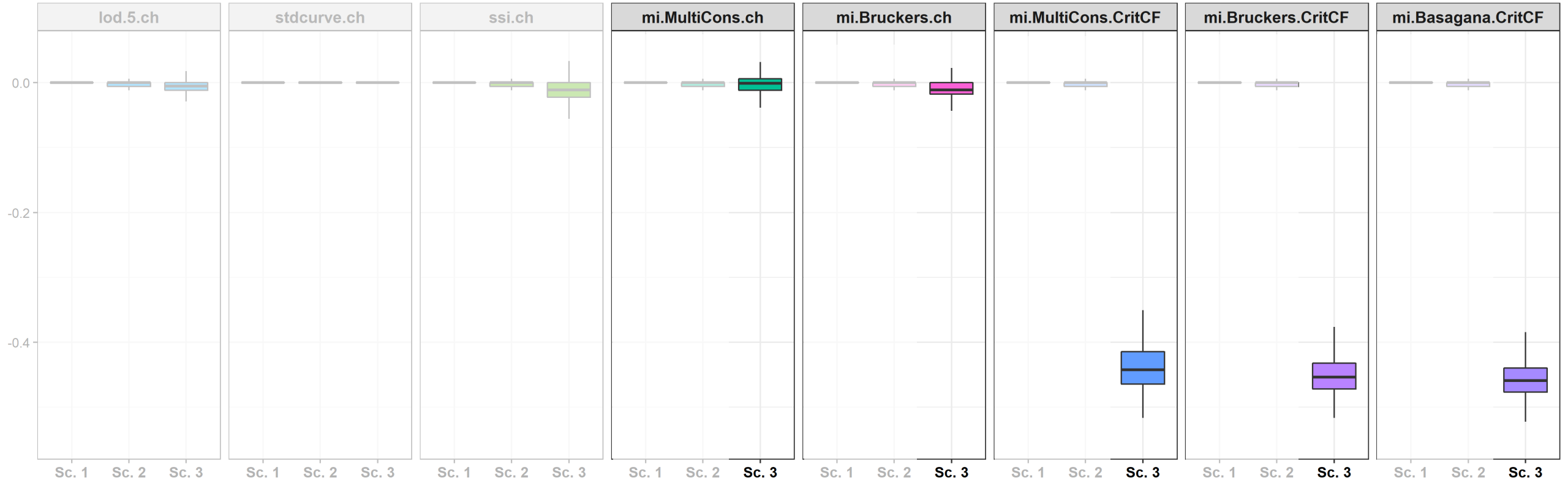


Left-censored data & no missing data

$\Delta_{Complete\ Data}$



Scenario 3: poor performances with CritCF



$\hat{k} = 2 : 2\%$
$\hat{k} \geq 4 : 4\%$

$\hat{k} = 2 : 2\%$

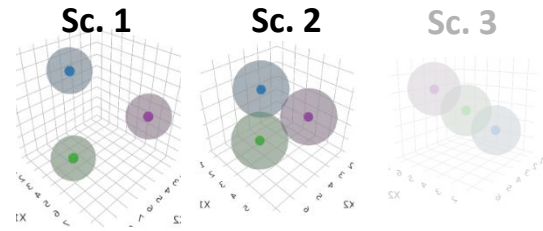
$\hat{\boldsymbol{k}} = \mathbf{2 : 79\%}$
$\hat{\boldsymbol{k}} \geq \mathbf{4 : \ 7\%}$

$\hat{\boldsymbol{k}} = \mathbf{2 : 94\%}$

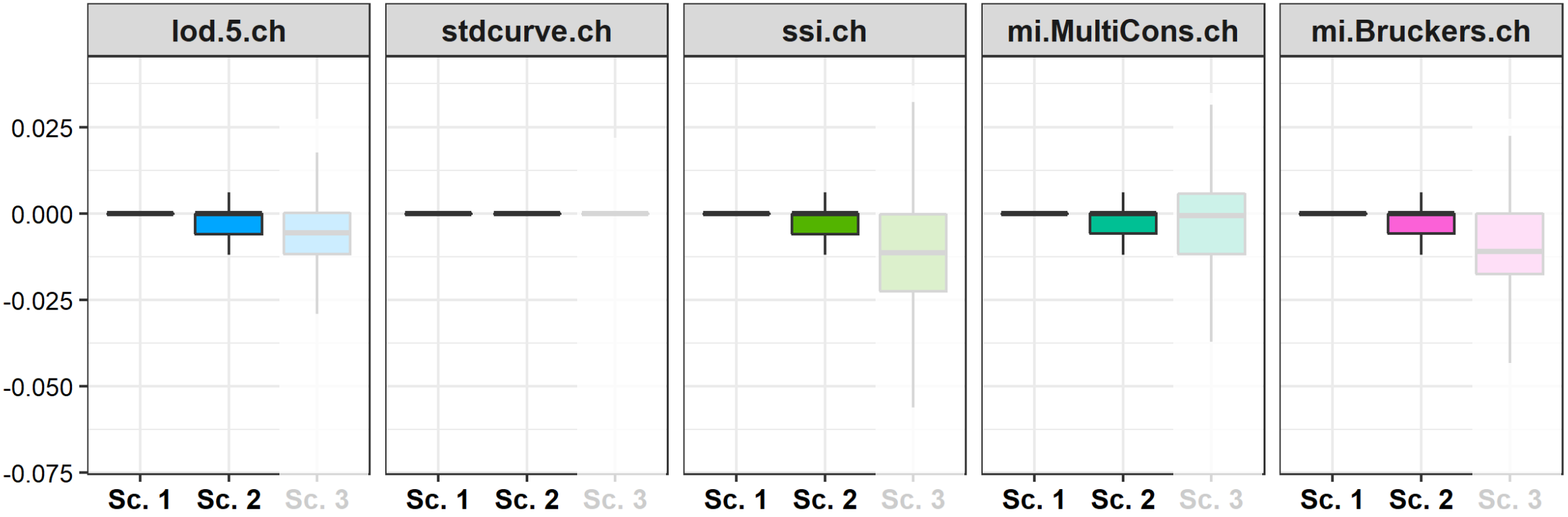
$\hat{\boldsymbol{k}} = \mathbf{2 : 99\%}$

Left-censored data & no missing data

$$ARI - ARI_{Complete\ Data}$$

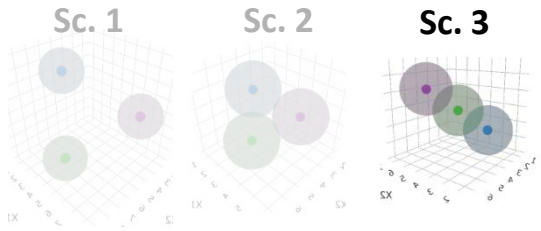


Scenario 1 & 2 : same performances of all methods

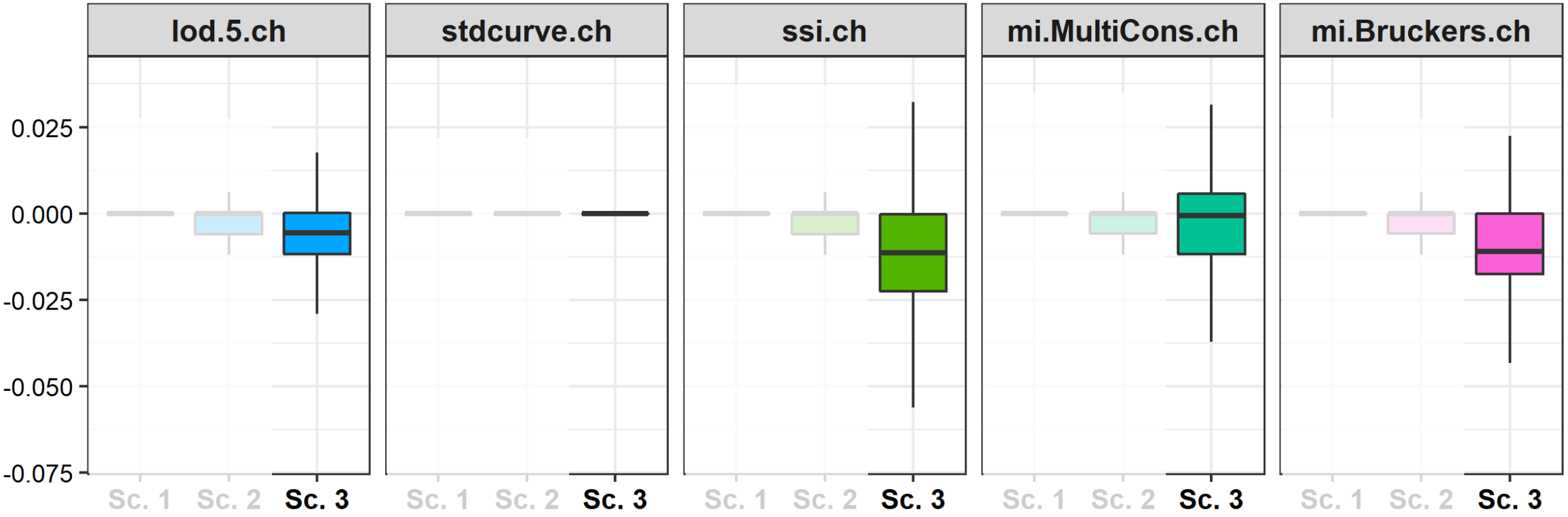


Left-censored data & no missing data

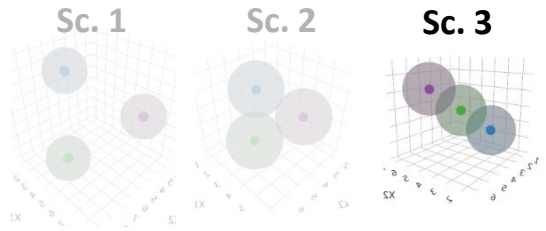
$$ARI - ARI_{Complete\ Data}$$



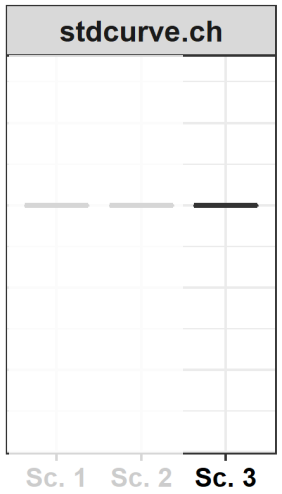
Scenario 3: best performances of standard curve imputation & MI with Multicons

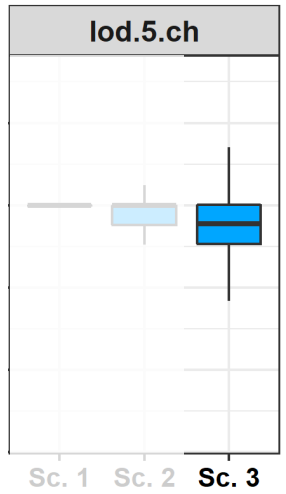


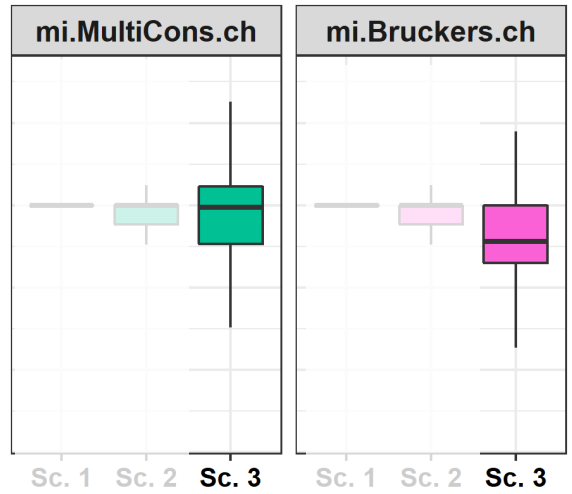
Left-censored data & no missing data

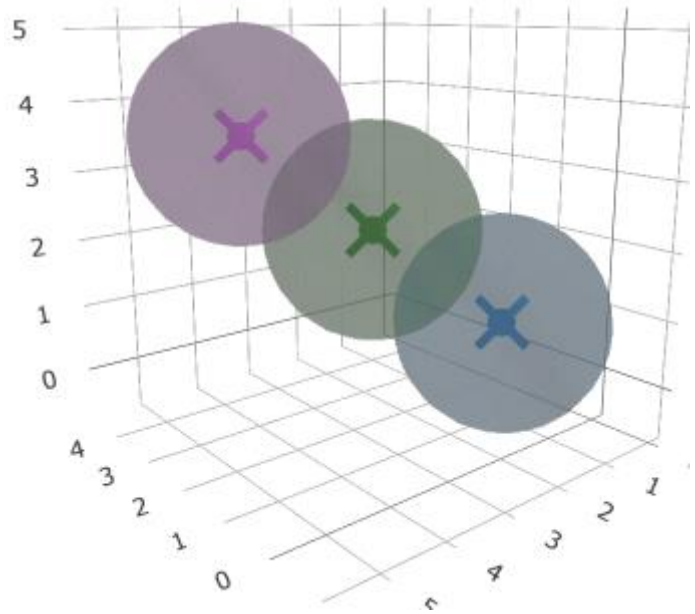


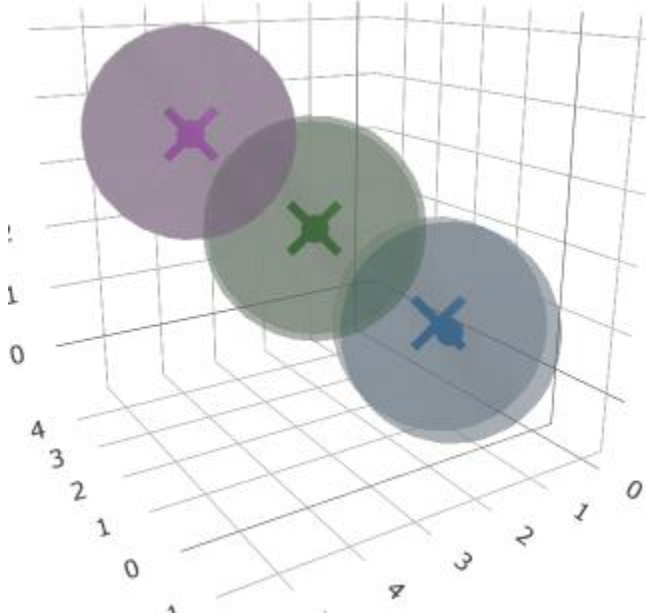
Scenario 3: best performances of standard curve imputation & MI with Multicons

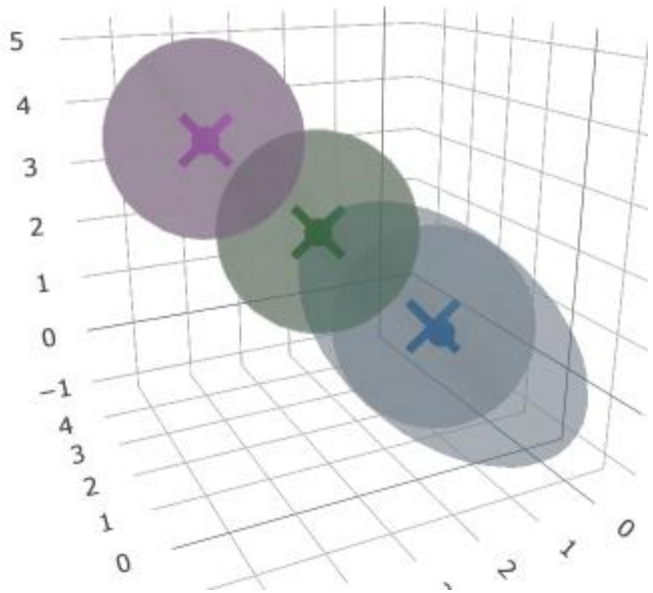








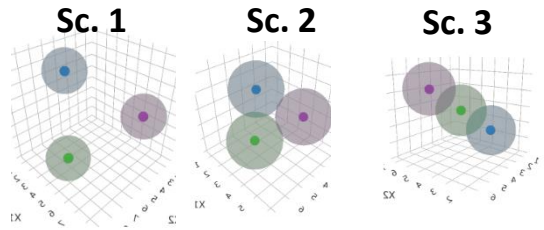




- Complete data
- After imputation

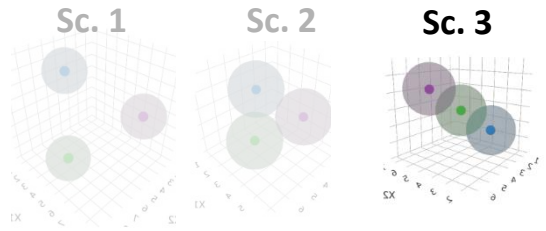
Missing data & no Left-censored data

$$ARI - ARI_{Complete\ Data}$$

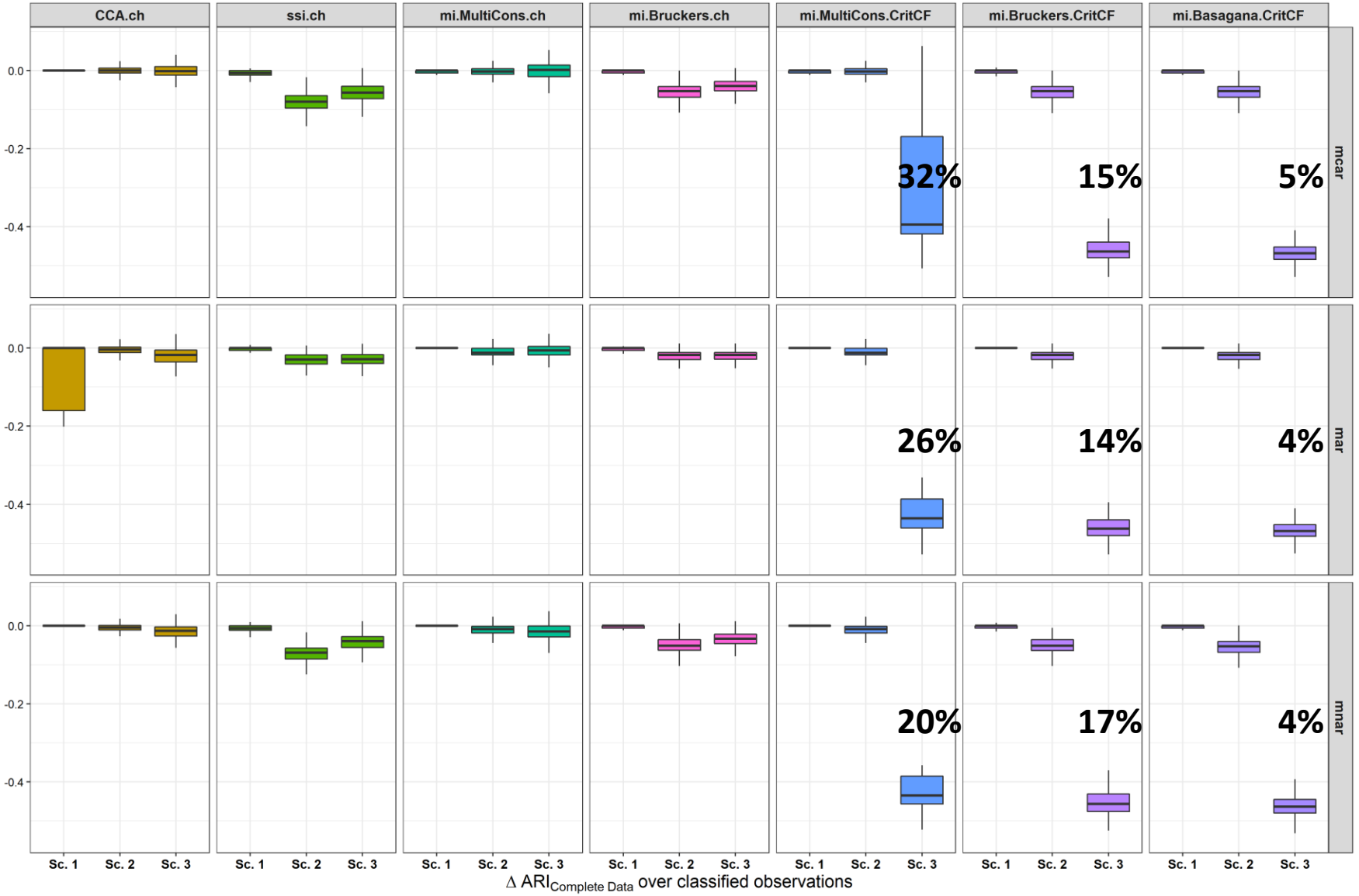


Missing data & no Left-censored data

$ARI - ARI_{Complete\ Data}$



Scenario 3: Poor performances with CritCF



MCAR

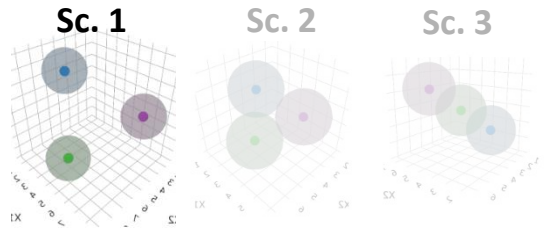
MAR

MNAR

$\hat{k} = 3$

Missing data & no Left-censored data

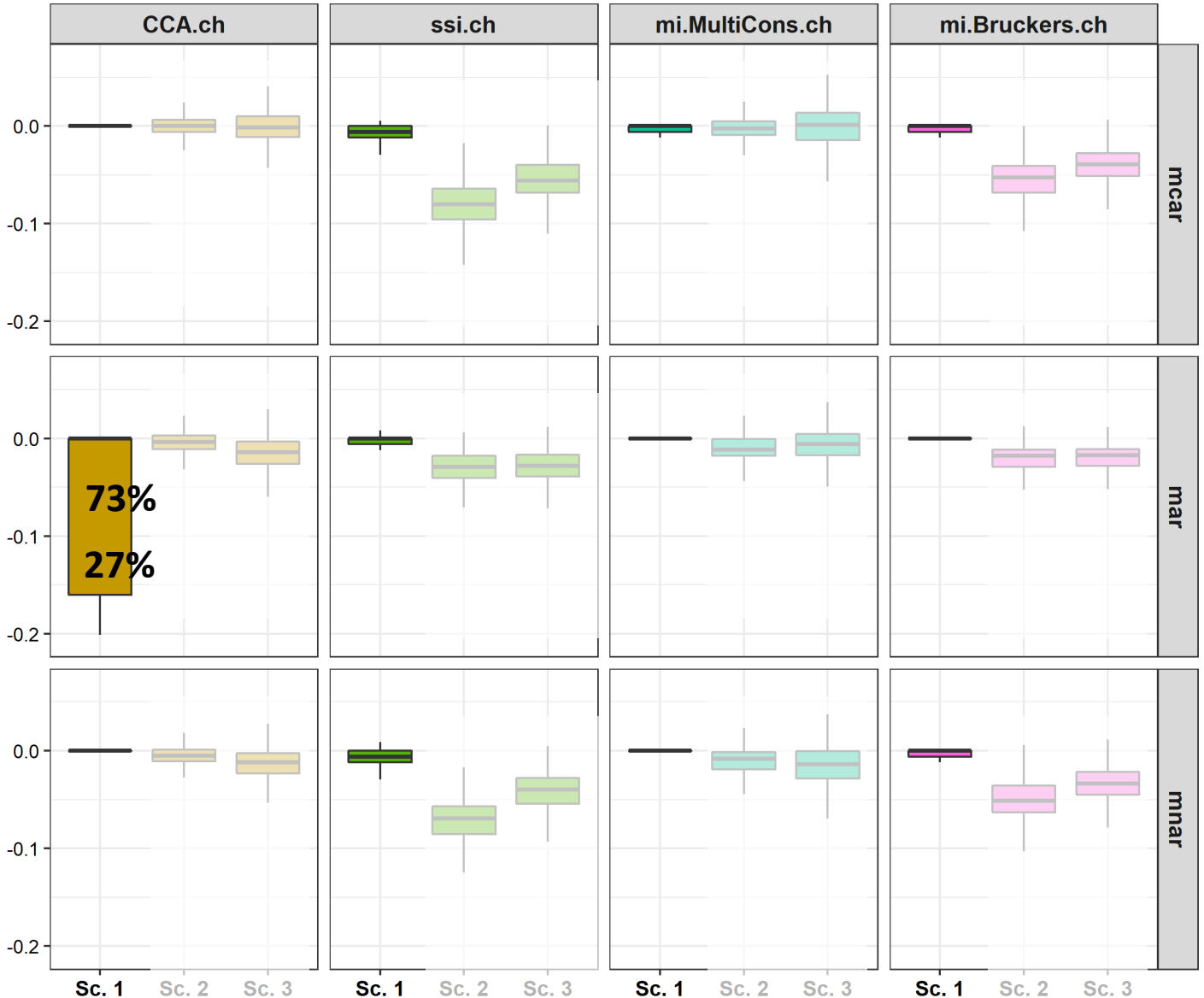
$$ARI - ARI_{Complete\ Data}$$



Scenario 1: under-performance of CCA with MAR mechanism

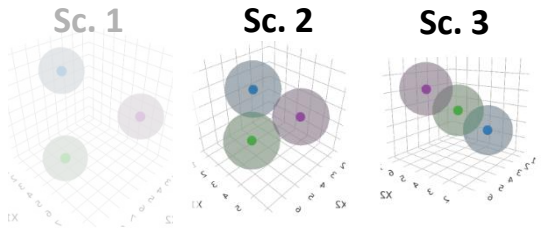
$\widehat{k} = 3$

$\widehat{k} = 4$

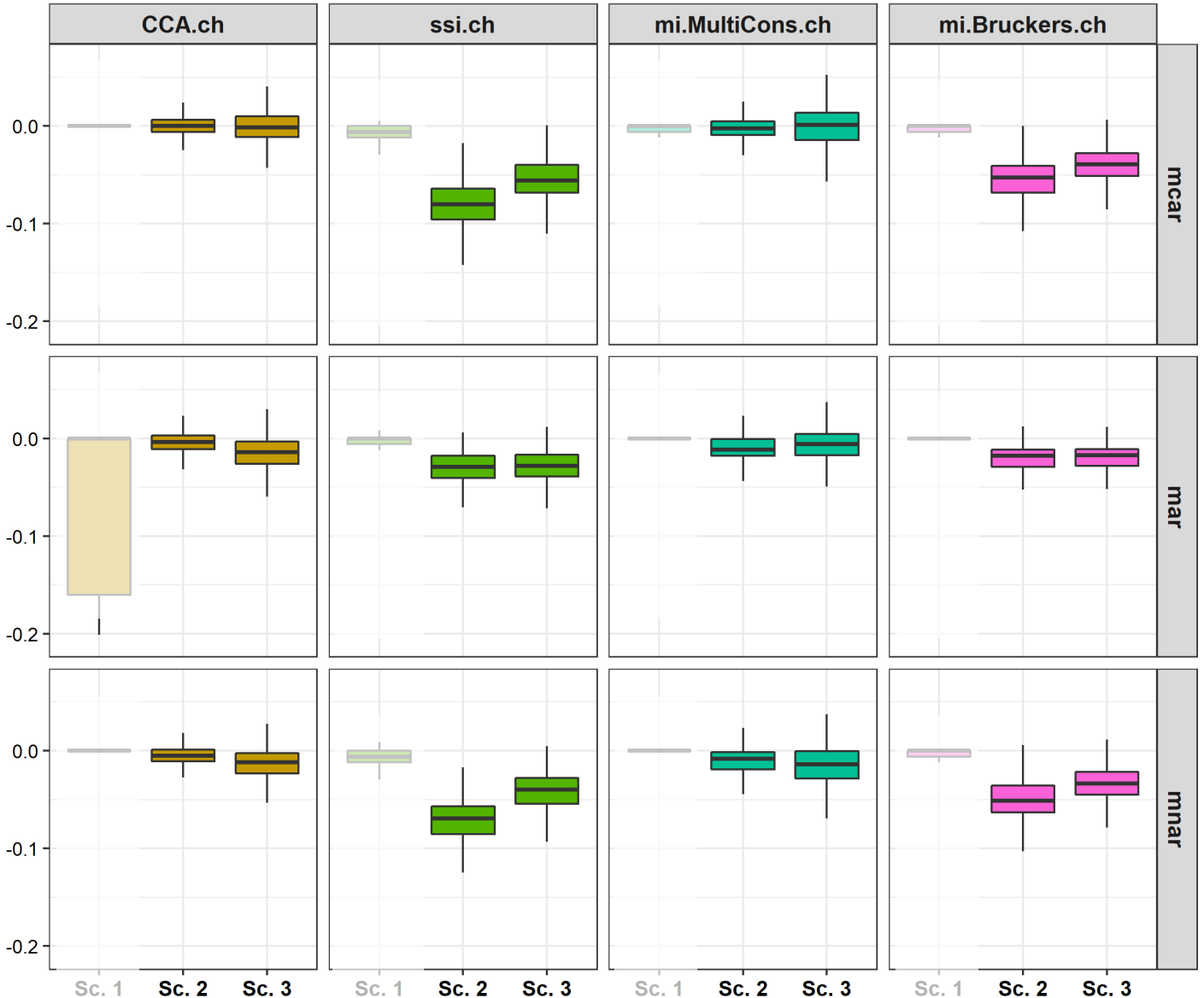


Missing data & no Left-censored data

$$ARI - ARI_{Complete\ Data}$$



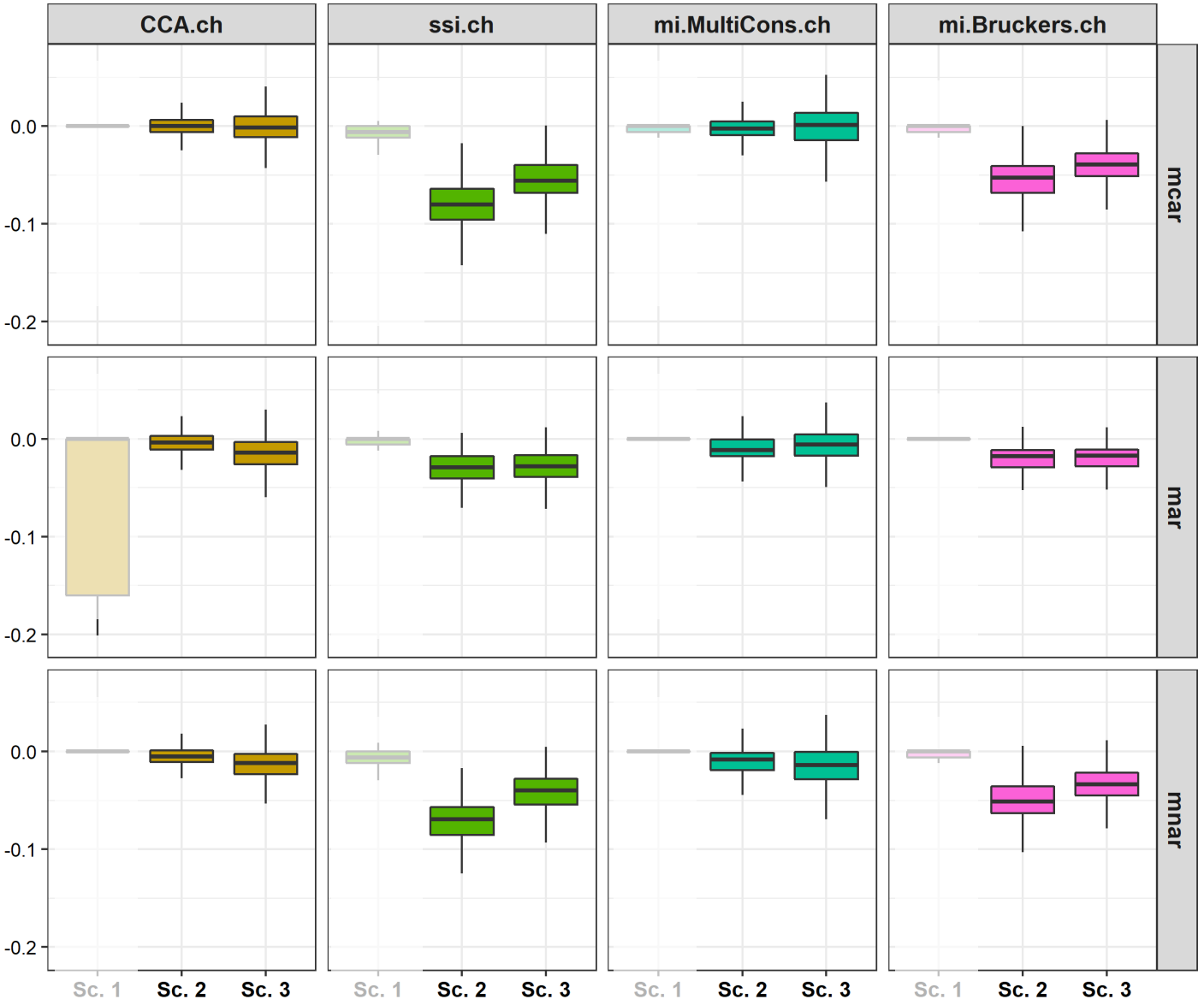
Scenario 2 & 3: Better performances under MAR than MCAR or MNAR

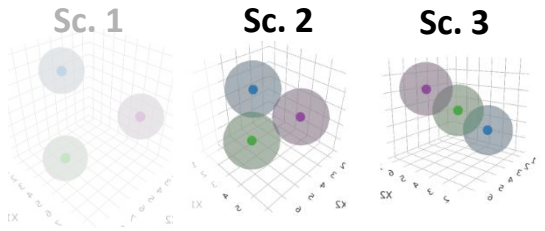


Missing data & no Left-censored data

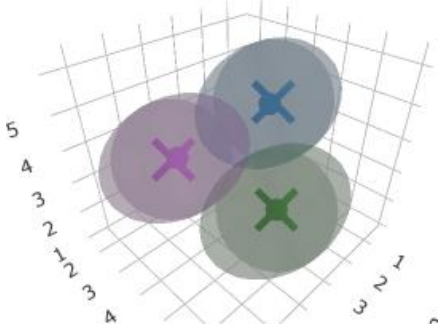
$$ARI - ARI_{Complete\ Data}$$

Scenario 2 & 3: Better performances under MAR than MCAR or MNAR

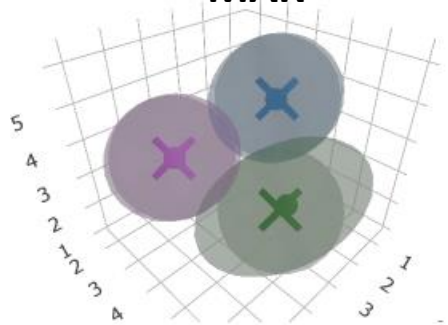




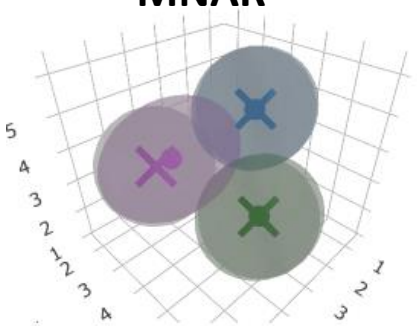
MCAR



MAR

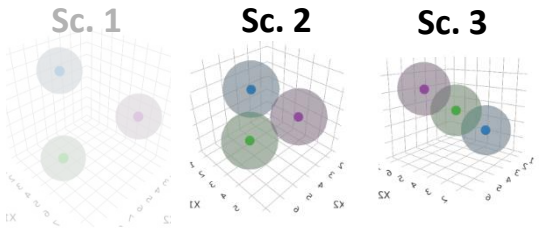


MNAR

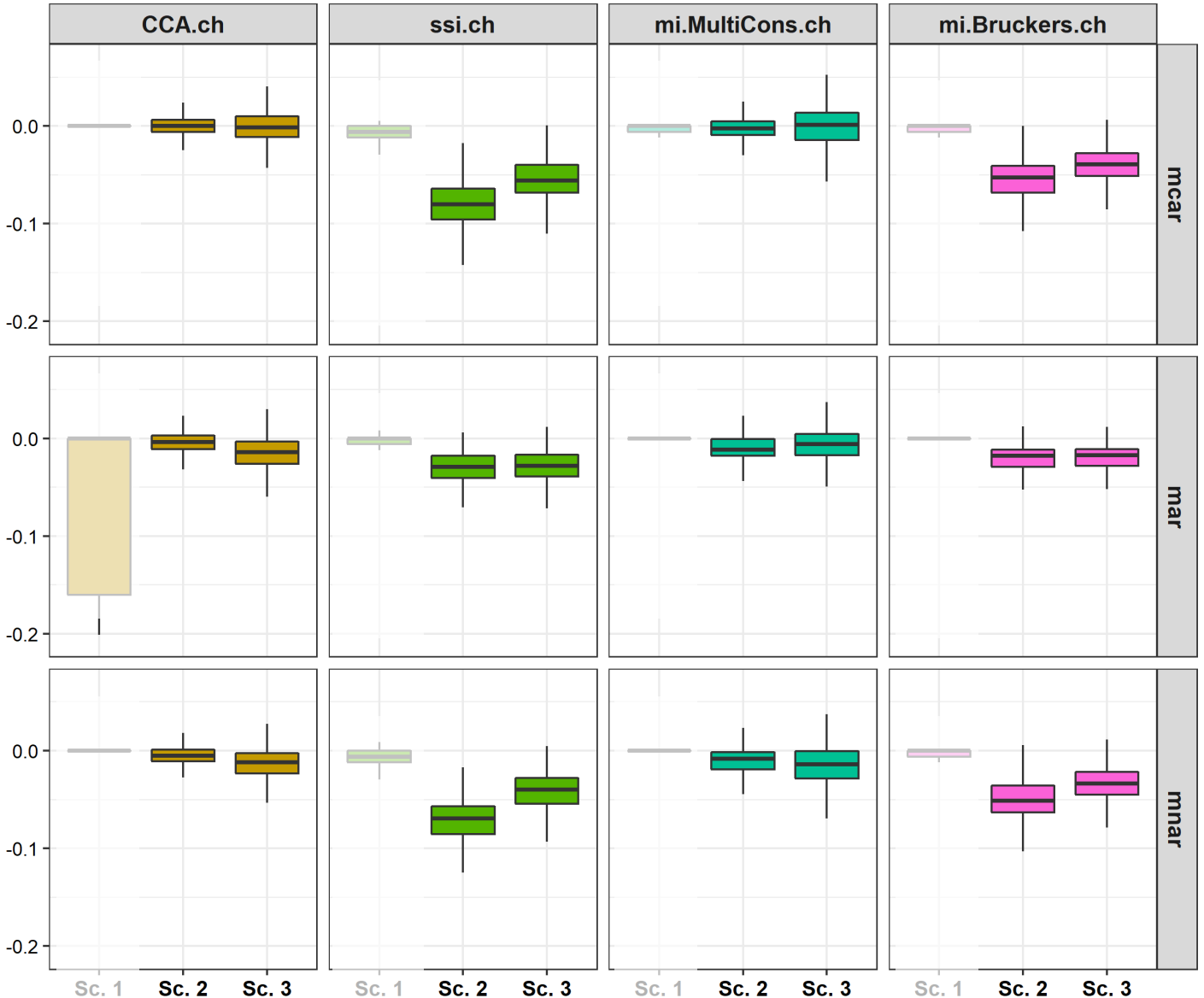


Missing data & no Left-censored data

$$ARI - ARI_{Complete\ Data}$$

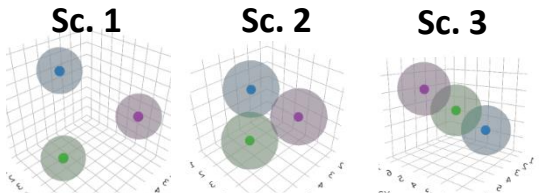


Scenario 2 & 3: Best performances with CCA & MI with Multicons



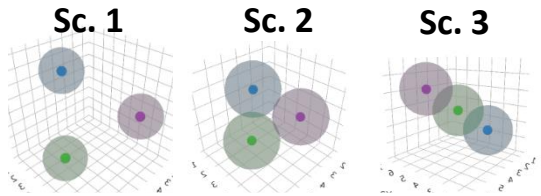
Missing data & Left-censored data

$$ARI - ARI_{Complete\ Data}$$



Missing data & Left-censored data

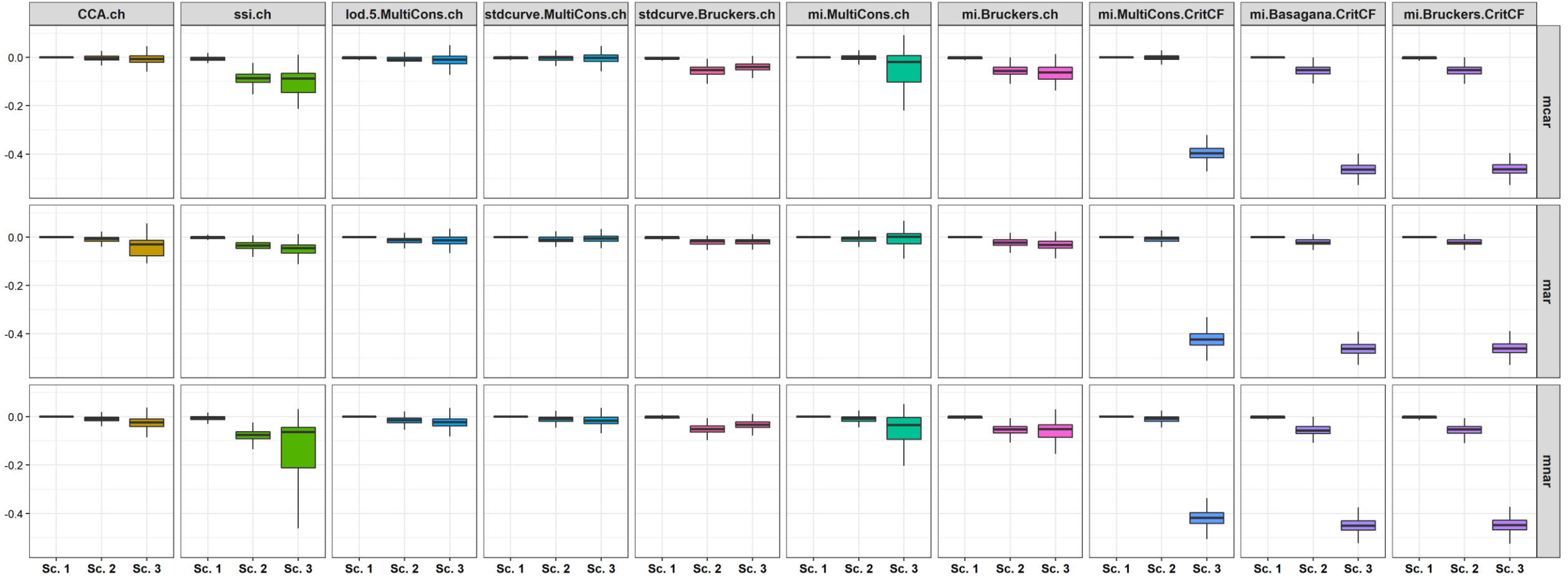
$$ARI - ARI_{Complete\ Data}$$



Sc. 1: no difference;

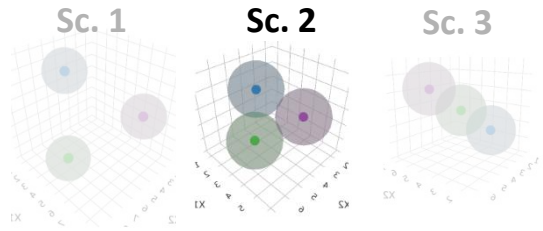
Sc. 2 & 3: MAR >> MCAR & MNAR, Multicons >> Bruckers & Basagana ;

Sc. 3: CH >> CritCF

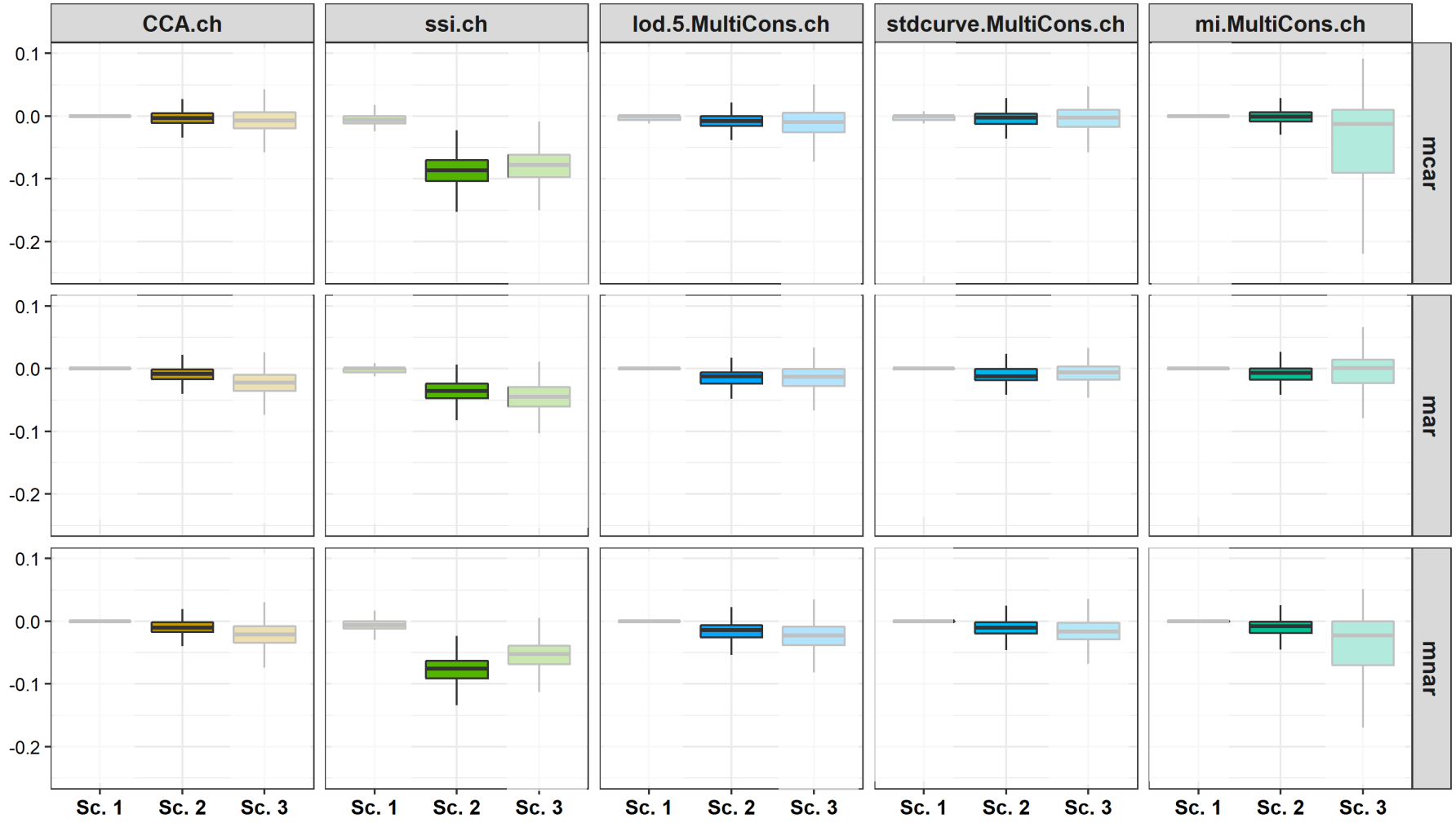


Missing data & Left-censored data

$$ARI - ARI_{Complete\ Data}$$



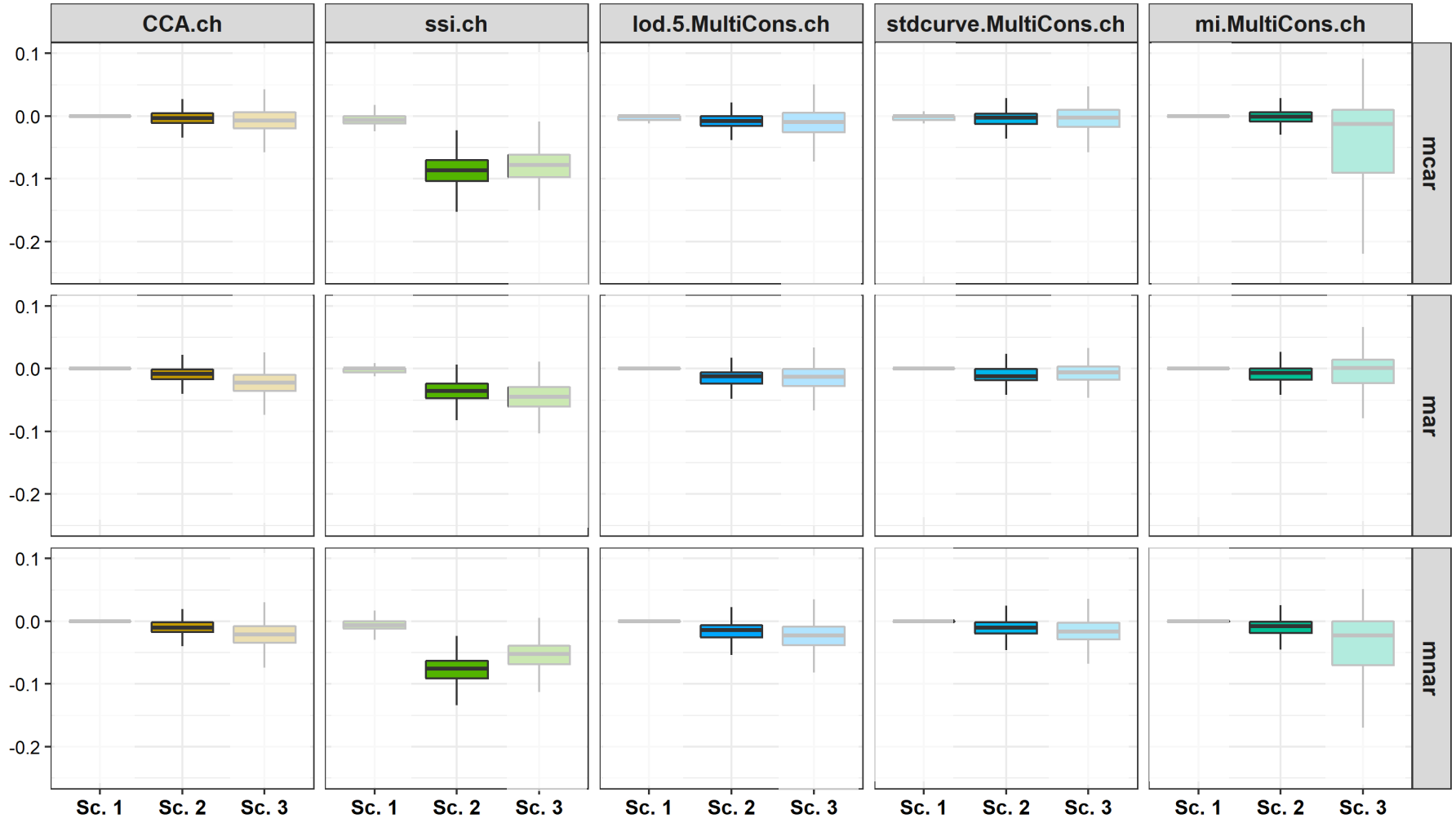
Scenario 2: Same performances of ½ LOD, Standard curve and MI

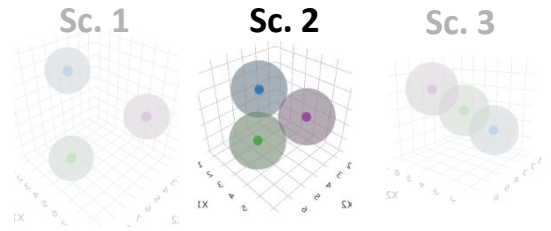


Missing data & Left-censored data

$$ARI - ARI_{Complete\ Data}$$

Scenario 2: Same performances of ½ LOD, Standard curve and MI

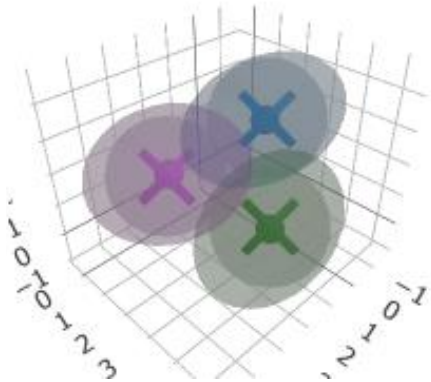




½ LOD SI

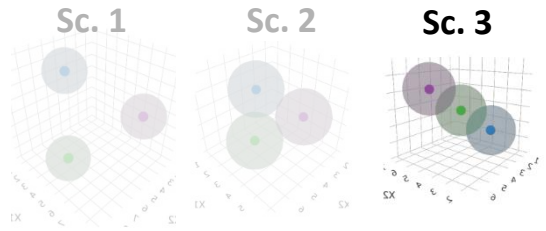
Standard curve SI

MI

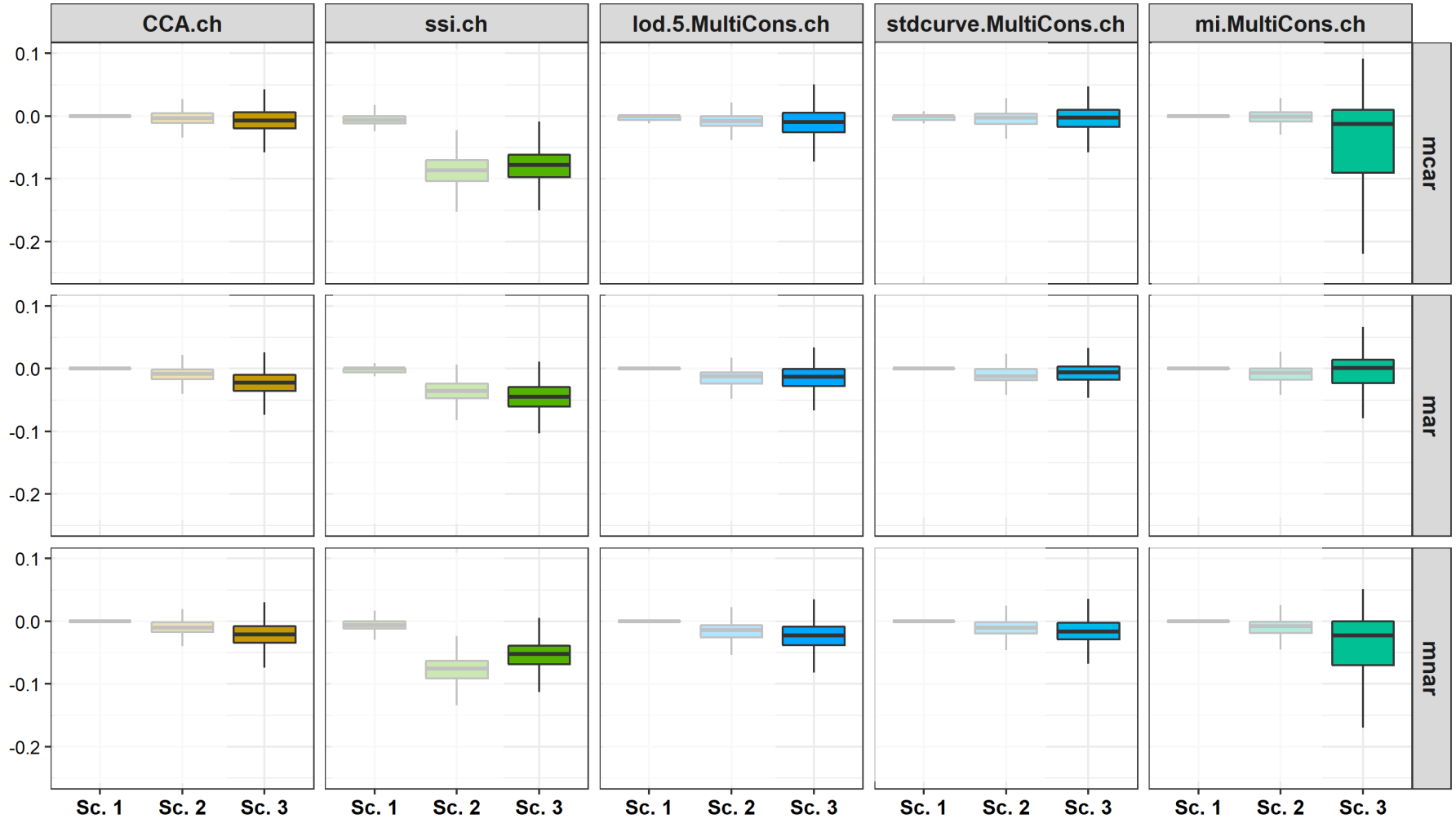


Missing data & Left-censored data

$$ARI - ARI_{Complete\ Data}$$



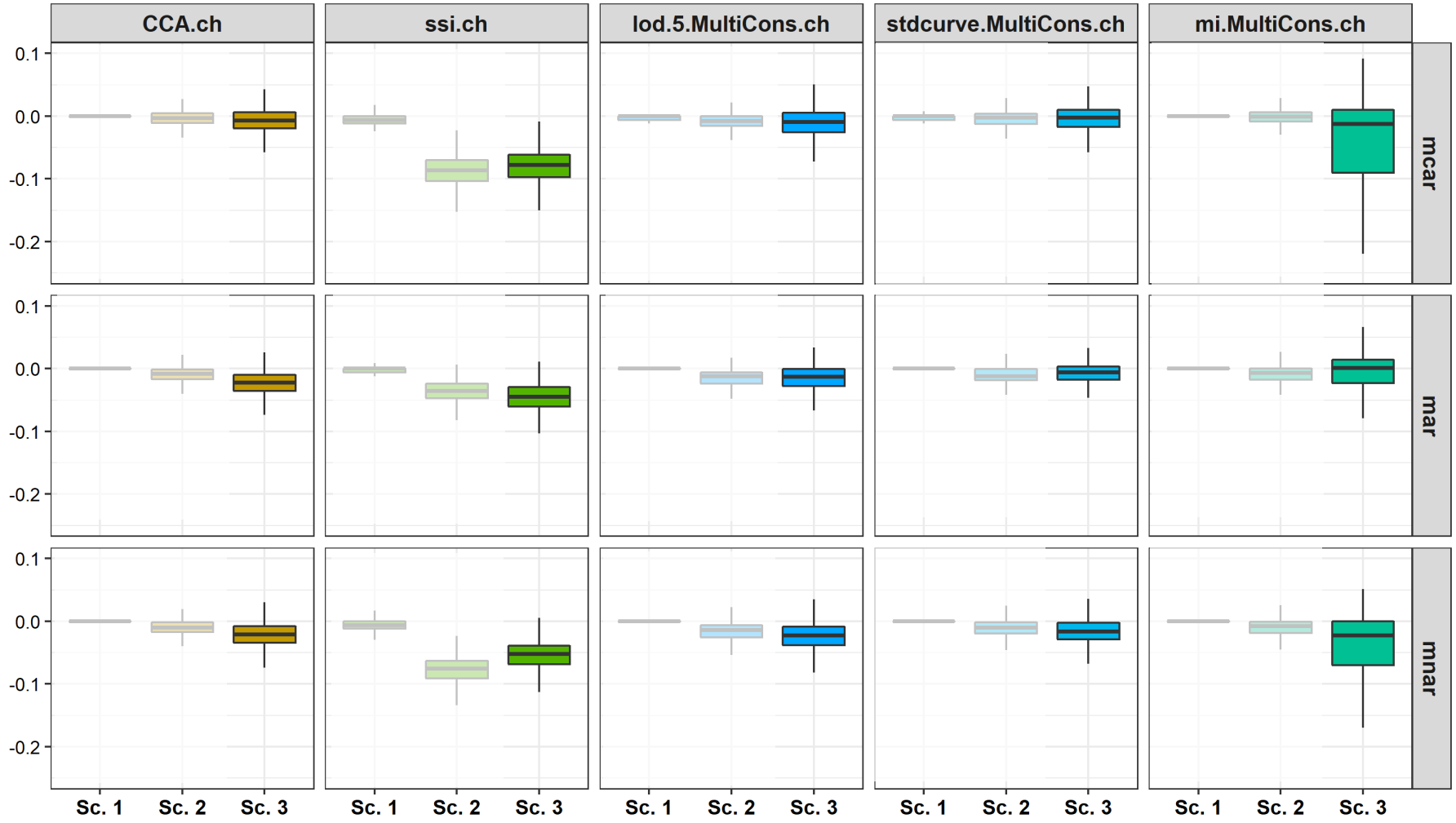
Scenario 3: More variability in MI performances compared to SI

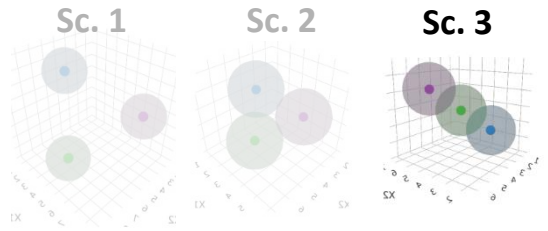


Missing data & Left-censored data

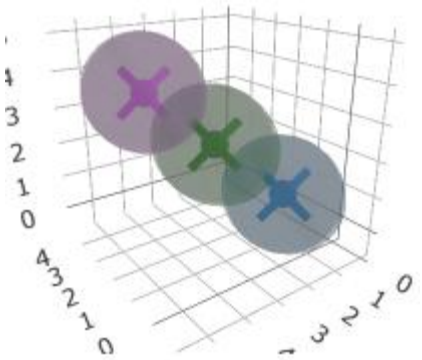
$$ARI - ARI_{Complete\ Data}$$

Scenario 3: More variability in MI performances compared to SI

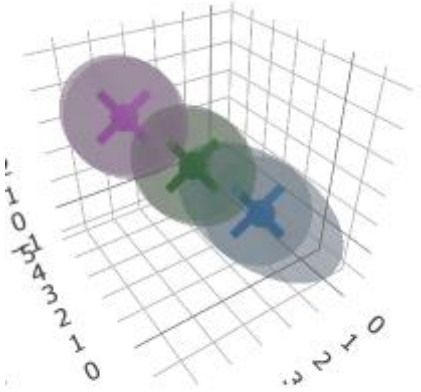




Standard curve SI



MI



Results

Left-censored data & no missing data

- **CH** >> CritfCF
- **MI with Multicons & standard curve** >> other methods

No Left-censored data & **missing data**

- **CH** >> CritfCF (∀ missing mechanism)
- **MI with Multicons & CCA** >> SSI, Bruckers consensus , Basagana consensus
- MAR >> MCAR & MNAR

Left-censored data & missing data

- **CH** >> CritfCF (∀ missing mechanism)
- **MI with Multicons & CCA** >> SSI, Bruckers consensus , Basagana consensus
- **½ LOD & standard curve** > MI for censored data (more variability)
- MAR >> MCAR & MNAR

Recomanded method: **Multicons**
with **imputation by standard curve for left censored data**
and **CH criterion for K selection**

Conclusion & Discussion

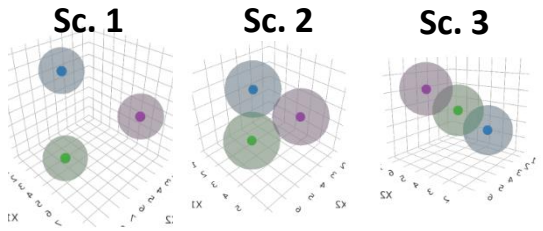
Results extandable to any dataset with missing data

Understand group structure distortion by incomplete data

MAR distortion seems to cause less difficulty than MNAR and MCAR

Performances compared to likelihood based methods

Results



Left-censored data & no missing data

- CH >> CritfCF
- MI with Multicons & standard curve >> other methods

No Left-censored data & **missing data**

- MI with Multicons & CCA >> SSI, Bruckers consensus , Basagana consensus
- CH >> CritfCF (∀ missing mechanism)
- MAR >> MCAR & MNAR

Left-censored data & missing data

Scenario 1:

- same performances as on complete data for all methods

Scenario 2:

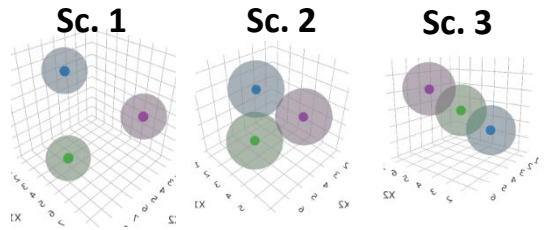
- under-performance of SSI, and Bruckers and Basagana consensus as compared to complete data (notably under MCAR and MNAR)
- No difference between ½ LOD, standard curve and MI for censored data
- Better performances under MAR as compared to MCAR and MNAR

Scenario 3:

- CH >> CritfCF (∀ missing mechanism)
- More variability in performances for MI as compared to ½ LOD and standard curve imputaiton
- Better performances under MAR as compared to MCAR and MNAR

Best methods: Multicons (with single imputaiton for censored data)

Results



Left-censored data & no missing data

Scenarios 1 and 2: - Same performances over all methods

Scenario 3: - CH >> CritfCF

- MI with Multicons & standard curve >> other methods

No Left-censored data & missing data

Scenario 1: - same performances as on complete data for all methods except CCA under-performs when missing mechanism is MAR

Scenario 2: - under-performance of SSI, and Bruckers and Basagana consensus as compared to complete data (notably under MCAR and MNAR)

- Better performances under MAR as compared to MCAR and MNAR

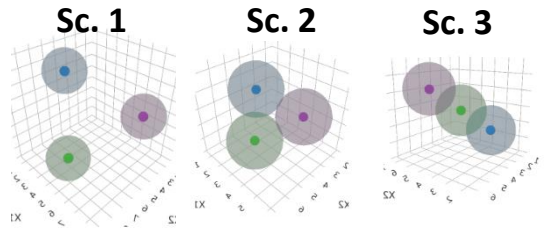
Scenario 3: - CH >> CritfCF (∀ missing mechanism)

- Better performances under MAR as compared to MCAR and MNAR

Best methods: MI with Multicons & CCA (∀ missing mechanism)

Left-censored data & no missing data

$\Delta_{Standard\ curve}$



Scenario 3: best performances of standard curve imputation & MI with Multicons