QUEEN'S UNIVERSITY BELFAST
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Data Analysis & Visualisation

CSC3062

BEng (CS & SE), MEng (CS & SE), BIT & CIT

Dr Reza Rafiee

Semester 1 – 2019/2020

Multiple Imputation Modelling and Diagnostics

What is multiple imputation?

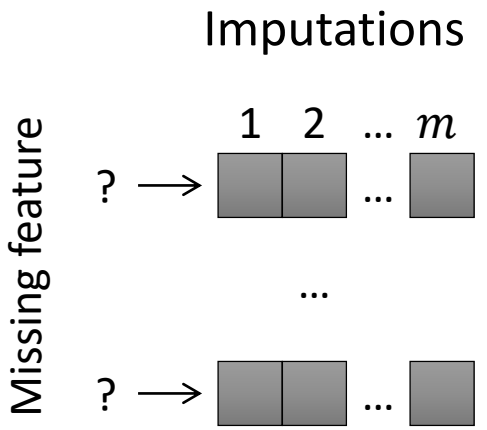
- This statistical technique (algorithm) takes the incomplete dataset (i.e., including missing data) and returns $\boldsymbol{m}$ imputed datasets with no missing values.

$\boldsymbol{m}$ is a user-selected parameter

Multiple imputation

- Each missing feature is imputed (filled in) with a set of $m > 1$ plausible values which reflect the uncertainty about the missing feature.

$$\begin{array}{c} \\ \text{Feature 1} \\ \dots \\ \text{Feature 17} \end{array} \begin{array}{c} \begin{array}{ccc} 1 & \dots & 103 \end{array} \\ \begin{bmatrix} 0.9 & \cdots & ? \\ \vdots & \ddots & \vdots \\ ? & \cdots & 0.1 \end{bmatrix} \end{array}$$

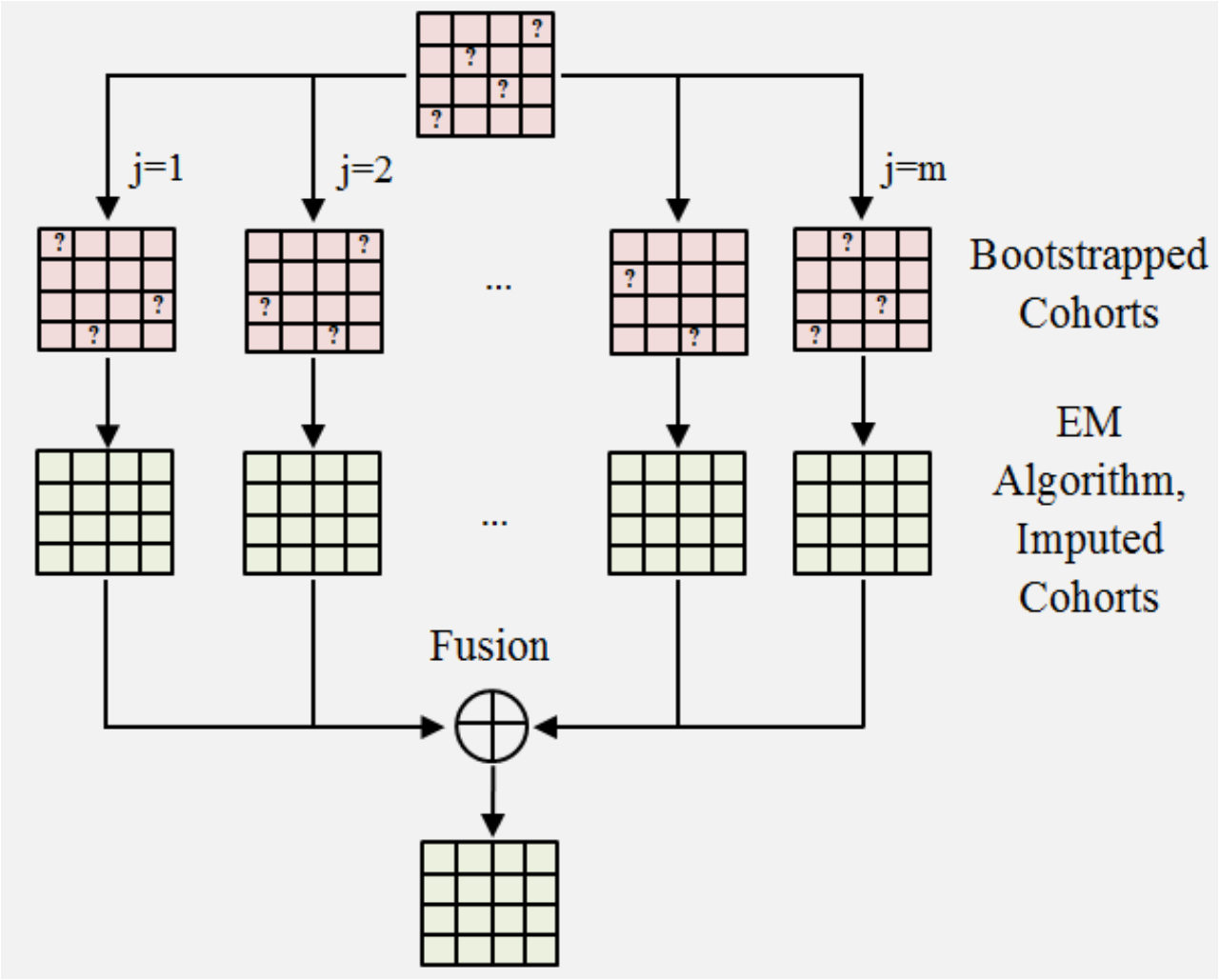


Multiple imputation modelling techniques

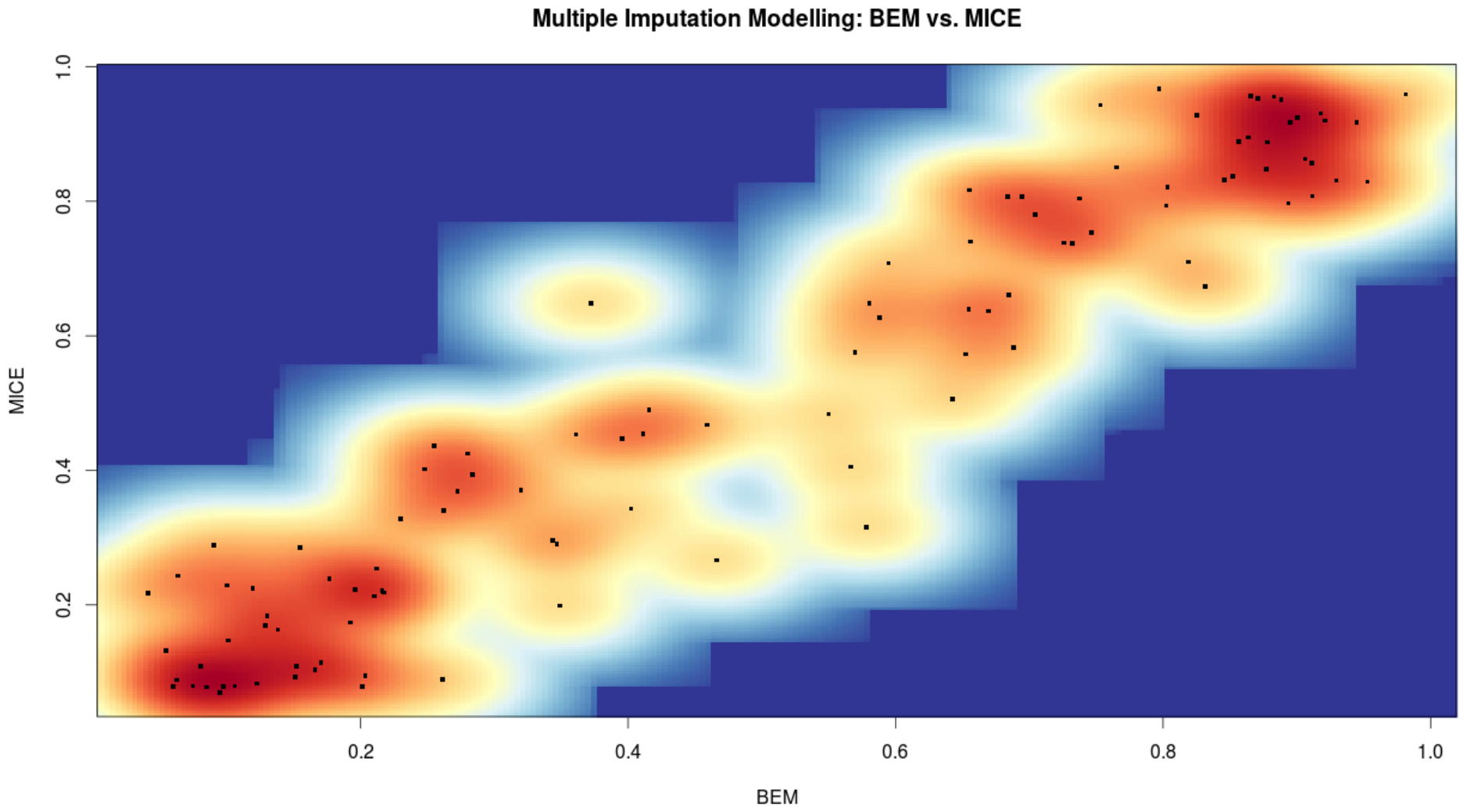
- Multivariate Imputation by Chained Equations (MICE)
- Bootstrapped Expectation-Maximisation (BEM)
- Multiple Imputation using an approximate Bayesian framework (MI)

Multiple imputation modelling techniques

- Bootstrapped Expectation-Maximisation (BEM)



Visualising BEM vs. MICE using scatter plot



Impact of a method (BEM or MICE)

Using different multiple imputation methods may affect the final results (e.g., classification results)

BEM

		Reference subgroup			
		WNT	SHH	Grp 3	Grp 4
Predicted Subgroup	WNT	22	0	0	0
	SHH	0	23	0	0
	Grp 3	0	0	23	0
	Grp 4	0	0	0	28
	NC+	2	4	1	0
	Total	24	27	24	28

MICE

		Reference subgroup			
		WNT	SHH	Grp 3	Grp 4
Predicted Subgroup	WNT	22	0	0	0
	SHH	0	22	0	0
	Grp 3	0	1	23	0
	Grp 4	0	0	0	28
	NC	2	5	0	0
	Total	24	28	23	28

Summarising the performance of a classification algorithm using a "confusion matrix". A matrix (table) shows the discrepancy between predicted and reference subgroup.

+NC: Non-classifiable

Efficiency of Multiple Imputation

- Efficiency of an estimate based on m imputation is approximately:

$$(1 + \frac{\gamma}{m})^{-1}$$

Where γ is the fraction of missing information for the quality being estimated.

1) Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys . J. Wiley & Sons, New York.
2) Schafer, Joseph L. and Maren K. Olsen. 1998. \Multiple imputation for multivariate missing-data problems: A data analyst's perspective." Multivariate Behavioral Research 33(4):545-571.

Efficiency of m imputations for 17 probes

Feature #	missing fraction	Efficiency of m imputation per feature	Average of efficiency (12 feature)
1	0.165048544	0.991815118	**0.995782732**
2	0.048543689	0.997578693	m=20
3	0.038834951	0.998062016	
4	0	-	
5	0.077669903	0.996131528	
6	0.009708738	0.999514799	
7	0.155339806	0.992292871	
8	0.009708738	0.999514799	
9	0.077669903	0.996131528	
10	0.038834951	0.998062016	
11	0	-	
12	0.242718447	0.988009592	
13	0	-	
14	0.019417476	0.999030068	
15	0	-	
16	0	-	
17	0.13592233	0.993249759	

Installation and Updates from R

To install the Amelia package on any platform, simply type the following at the R command prompt,

```
> install.packages("Amelia")
> update.packages()
```