Practical Subgrouping in Medulloblastoma

Dr Reza Rafiee

Northern Institute for Cancer Research

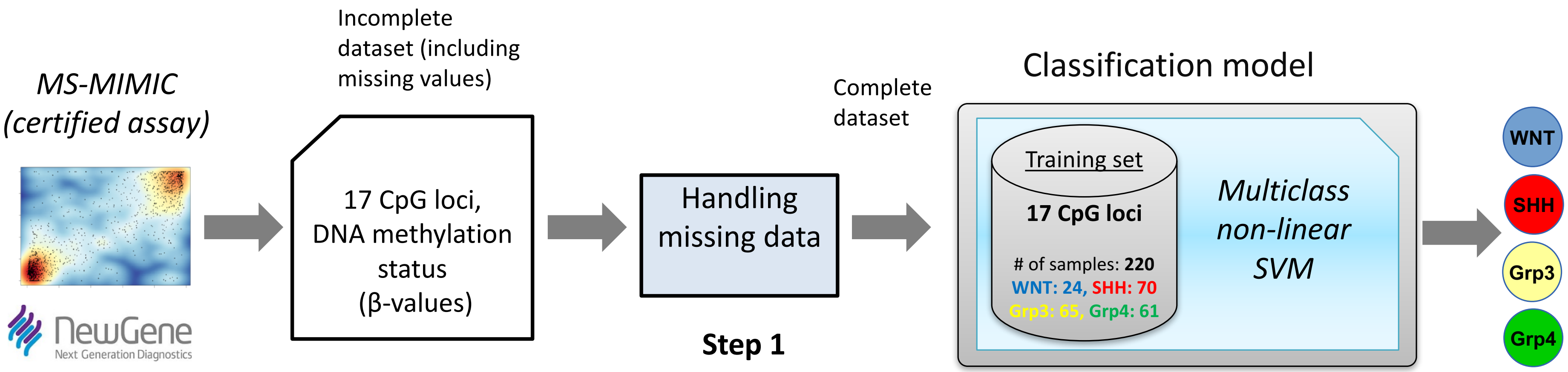
Newcastle University

10/04/2017

gholamreza.rafiee@ncl.ac.uk

Model and challenges

Aim: designing a reliable classification model to classify samples into one of the four known molecular subgroups.



MS-MIMIC (certified assay)

NewGene Next Generation Diagnostics

Incomplete dataset (including missing values)

17 CpG loci, DNA methylation status (β-values)

Handling missing data

Step 1

Complete dataset

Classification model

Training set

17 CpG loci

# of samples: **220**

WNT: 24, SHH: 70

Grp3: 65, Grp4: 61

Multiclass non-linear SVM

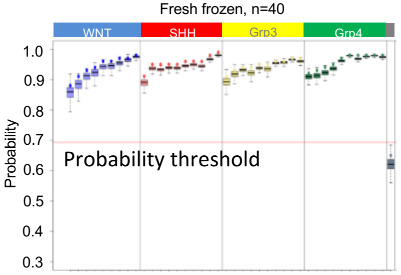
WNT

SHH

Grp3

Grp4

Step 2



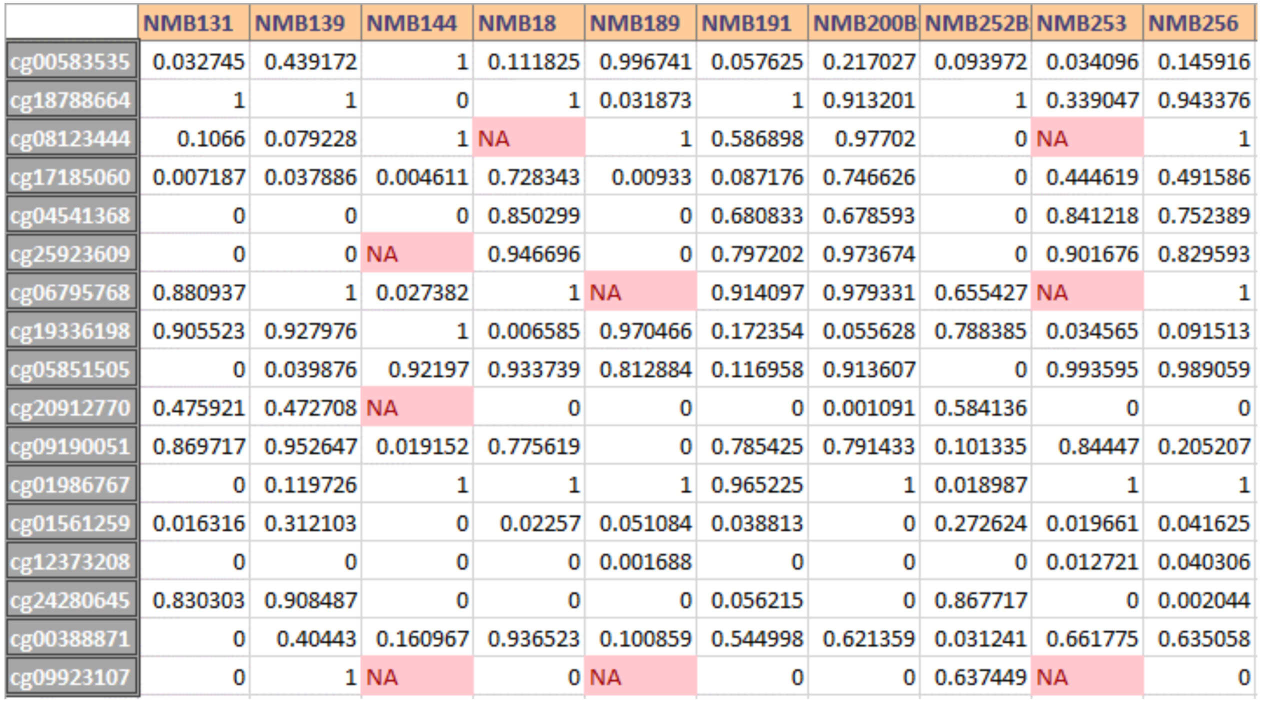
An example of incomplete dataset
(including missing data/β-values)

NA: Missing β-values

Samples

Features
(17 CpG loci)

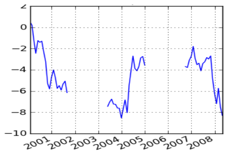
0 ≤ β-value ≤ 1

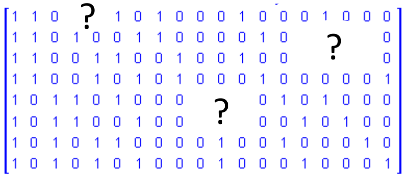


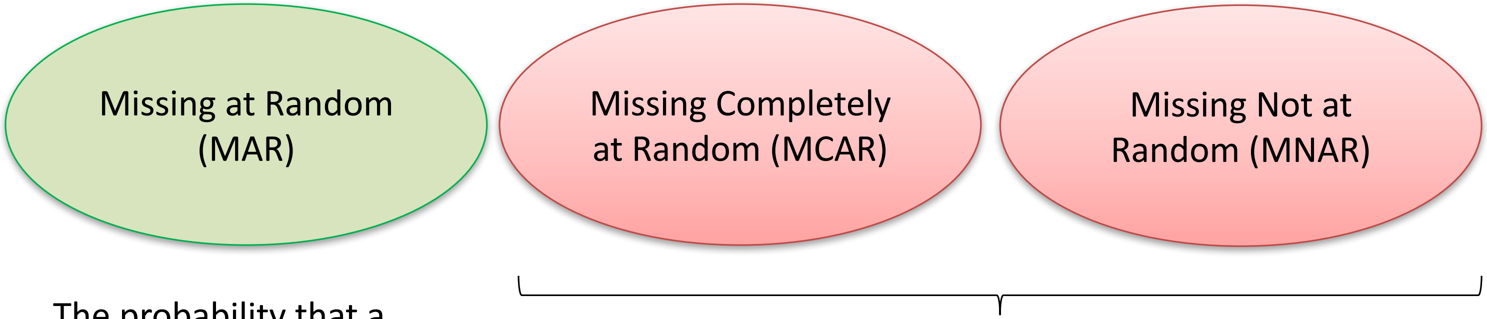
	NMB131	NMB139	NMB144	NMB18	NMB189	NMB191	NMB200B	NMB252B	NMB253	NMB256
cg00583535	0.032745	0.439172	1	0.111825	0.996741	0.057625	0.217027	0.093972	0.034096	0.145916
cg18788664	1	1	0	1	0.031873	1	0.913201	1	0.339047	0.943376
cg08123444	0.1066	0.079228	1	NA	1	0.586898	0.97702	0	NA	1
cg17185060	0.007187	0.037886	0.004611	0.728343	0.00933	0.087176	0.746626	0	0.444619	0.491586
cg04541368	0	0	0	0.850299	0	0.680833	0.678593	0	0.841218	0.752389
cg25923609	0	0	NA	0.946696	0	0.797202	0.973674	0	0.901676	0.829593
cg06795768	0.880937	1	0.027382	1	NA	0.914097	0.979331	0.655427	NA	1
cg19336198	0.905523	0.927976	1	0.006585	0.970466	0.172354	0.055628	0.788385	0.034565	0.091513
cg05851505	0	0.039876	0.92197	0.933739	0.812884	0.116958	0.913607	0	0.993595	0.989059
cg20912770	0.475921	0.472708	NA	0	0	0	0.001091	0.584136	0	0
cg09190051	0.869717	0.952647	0.019152	0.775619	0	0.785425	0.791433	0.101335	0.84447	0.205207
cg01986767	0	0.119726	1	1	1	0.965225	1	0.018987	1	1
cg01561259	0.016316	0.312103	0	0.02257	0.051084	0.038813	0	0.272624	0.019661	0.041625
cg12373208	0	0	0	0	0.001688	0	0	0	0.012721	0.040306
cg24280645	0.830303	0.908487	0	0	0	0.056215	0	0.867717	0	0.002044
cg00388871	0	0.40443	0.160967	0.936523	0.100859	0.544998	0.621359	0.031241	0.661775	0.635058
cg09923107	0	1	NA	0	NA	0	0	0.637449	NA	0

Categories of missingness

- Failure in:
  - Responding to a question (in surveys)
  - Equipment (sensors), recording mechanisms
  - Data entry
  - …






Missing at Random (MAR)
Missing Completely at Random (MCAR)
Missing Not at Random (MNAR)

The probability that a value is missing depends only on observed values.

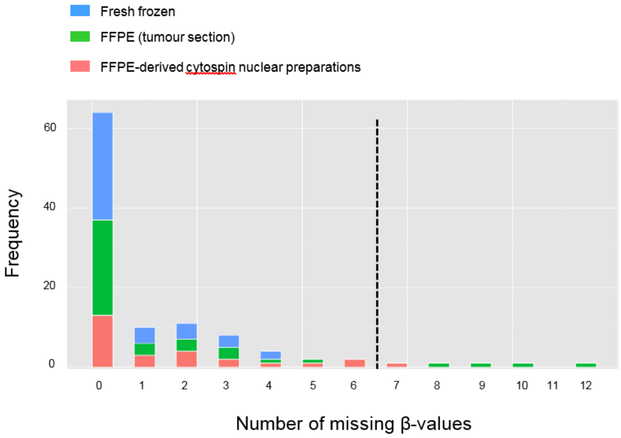
The missingness cannot be predicted from any other variables or sets of variables.

Missing data

Why missing: by using poor quality DNA (e.g., FFPE derived), some loci will fail to be assayed (**still is not clear the reason**).

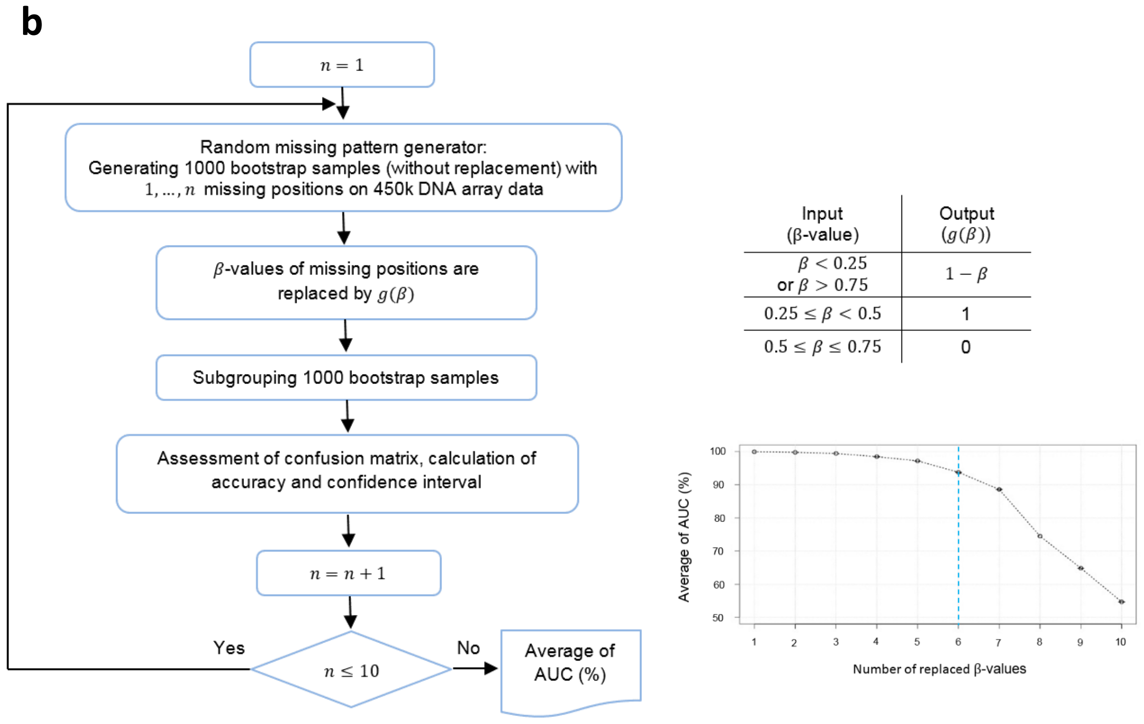
Two key questions: 1) what is the acceptable number of missing data (β-values)? 2) how to create a complete dataset from an incomplete one?

a



63/106 (59%) samples reported complete sets of β-values whereas **5/106 (5%)** samples had more than **7** missing β-values (QC measure for CpG locus-specific threshold; black line)

b


$n = 1$
Random missing pattern generator:
Generating 1000 bootstrap samples (without replacement) with $1, \dots, n$ missing positions on 450k DNA array data
β-values of missing positions are replaced by $g(\beta)$
Subgrouping 1000 bootstrap samples
Assessment of confusion matrix, calculation of accuracy and confidence interval
$n = n + 1$
Yes
$n \leq 10$
No
Average of AUC (%)
Average of AUC (%)
50
60
70
80
90
100
1
2
3
4
5
6
7
8
9
10
Number of replaced β-values

Input (β-value)	Output ($g(\beta)$)
$\beta < 0.25$ or $\beta > 0.75$	$1 - \beta$
$0.25 \leq \beta < 0.5$	1
$0.5 \leq \beta \leq 0.75$	0

Empirical determination of the maximal number of permissible missing β-values. a) The prediction accuracy of the SVM classifier model was evaluated *in silico* by replacing missing data with confounding methylation values, using the transformation shown in the table.

Using the 17-locus signature from 450k DNA methylation array data, random combinations of 1 to 10 β-values were replaced with confounding data and the performance of the classifier assessed. The average area under curve (AUC) from 1000 bootstraps was plotted. **An average AUC of > 94% is achieved up to 6 missing β-value data points. Assay performance declines with more than 6 missing β-value data points (QC threshold; blue dotted line).**

Package/library in R

- ‘Amelia’: Bootstrap + EM
- ‘mice’: Multivariate Imputation using Chained Equations
- ‘mi’: Multiple Imputation using an approximate Bayesian framework

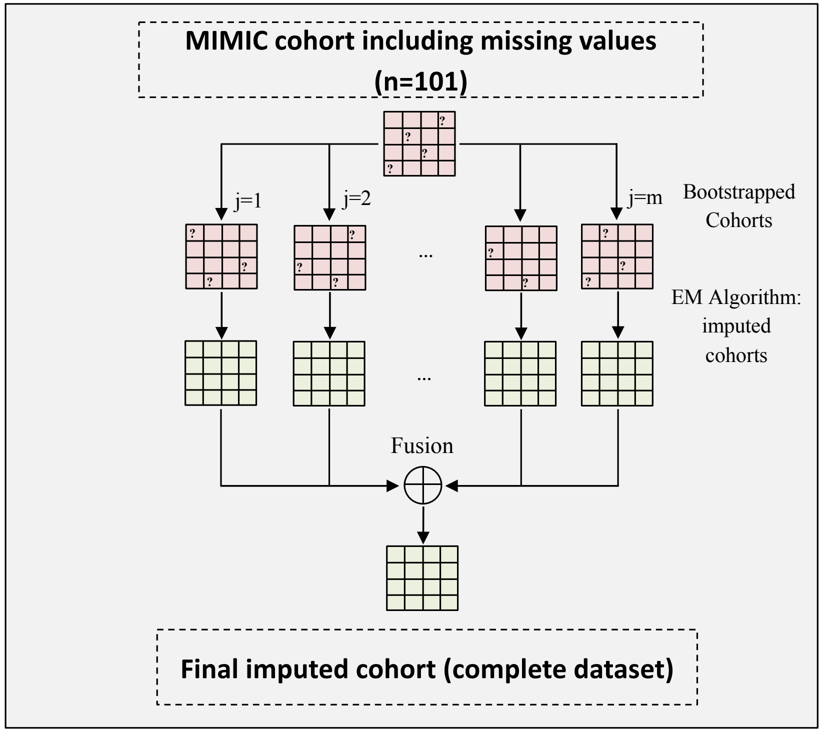
1) Diagnostics of the models
2) Provides graphics to visualize missing data patterns
3) Provides degree of sampling uncertainty
4) Applicable for categorical data as well

Multiple imputation modelling

using **Amelia** package in R

Assumptions to use this package: missing at random (MAR) and multivariate normality

MAR assumption: the pattern of missingness only depends on the observed data, not the unobserved data (missing)



MIMIC cohort including missing values (n=101)

Final imputed cohort (complete dataset)

'Impute' definition: assign (a value) to something by inference from the value of the products or processes to which it contributes.

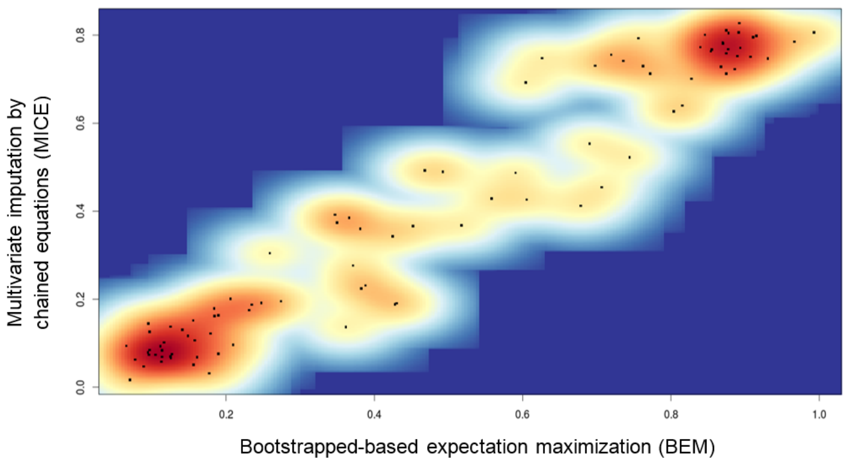
Bootstrapping: random sampling with replacement

Why we need bootstrapping: to simulate estimation uncertainty

Multiple imputation involves imputing *m* plausible values for each missing cell (reflecting the uncertainty about the missing value) in your data matrix and creating *m* "completed" data sets.

```
install.packages("Amelia", repos="http://r.iq.harvard.edu",
type = "source")
```

Imputation results by using “Amelia” and “mice” packages



Predicted subgroup is insensitive to multiple imputation modelling technique. Scatterplot of β-values generated by the bootstrapped-based expectation maximization (BEM) (*x* axis) and multivariate imputation by chained equations (MICE) (*y* axis) showing a strong correlation between the two methods (R^2=0.77).

Creating an optimal SVM classifier in R using e1071 package

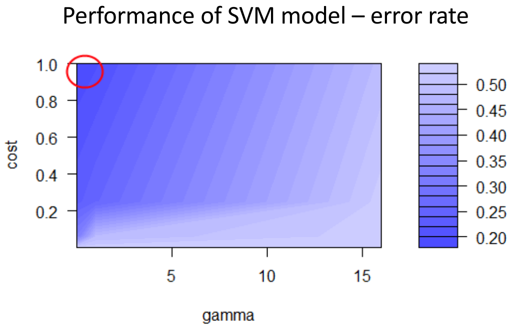
TUNING: a grid-based appraoch

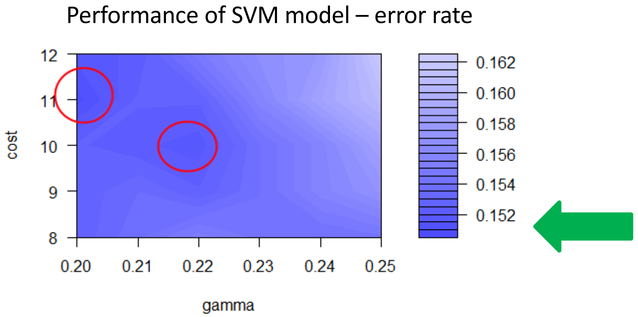
```
Tuning_model <- tune(svm, Trainingset450k17, label_vector,
scale = F, tolerance = 0.00001, type = "C-classification",
kernel = "radial", probability = T
ranges = list(cost= seq(0.0, 1.0, 0.2), gamma = seq(0, 15, 1)),
tunecontrol= tune.control(sampling = “cross”, cross=10), seed=1234)
```

The darkest shades of blue indicating the best (see the two plots).
Narrowing in on the darkest blue range and performing further tuning.

```
Plot(Tuning_model, xlime=range(0:15), ylime=range(0:1))

Plot(Tuning_model, xlime=range(0.2:0.25), ylime=range(8:12))
```





TRAINING:

```
Radial_model <- svm(Trainingset450k17, label_vector,
scale = F, tolerance = 0.00001, type = "C-classification",
kernel = "radial",
cost = optimum_cost, gamma = optimum_gamma,
probability = T, seed = 1234)
```

TESTING:

```
Radial_model <- predict(object= Radial_model, newdata = seq.test.BEM.97, probability=T)
```

Acknowledgment

Clifford, S.C.[1]

Schwalbe, E.C.[1,2]

Hicks, D.[1]

Bashton, M.[1], Enshaei, A.[1]

Gohlke, H.[3],Potluri, S.[1], Matthiesen, J.[1], Mather, M.[1], Taleongpong, P.[1], Chaston, R.[4], Scott, K.[4], Silmon, A.[4], Curtis, A.[4], Lindsey, J.C.[1], Crosier, S.[1], Smith, A.J.[1], Goschzik, T[5]., Doz, F[6]., Rutkowski, S[7]., Lannering, B.[8], Pietsch, T.[5], Bailey, S.[1], Williamson, D.[1],

[1]Northern Institute for Cancer Research, Newcastle University, Newcastle upon Tyne, U.K.
[2]Northumbria University, Newcastle upon Tyne, U.K.
[3]Agena, Hamburg, Germany
[4]NewGene, Newcastle upon Tyne, U.K.
[5]Department of Neuropathology, University of Bonn Medical Center, Bonn, Germany
[6]Institut Curie and University Paris Descartes, Paris, France
[7]University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[8]Department of Pediatrics, University of Gothenburg and The Queen Silvia Children's Hospital, Gothenburg, Sweden