A family of mixture models for beta valued DNA methylation data

Koyel Majumdar[1], Isobel C Gormley[1], Thomas B Murphy[1], Antoinette S Perry[2,3], Ronald W Watson[2] and Romina Silva[2,3]

[1]School of Mathematics and Statistics, [2]School of Medicine, UCD Conway Institute, [3]School of Biology and Environmental Science, University College Dublin

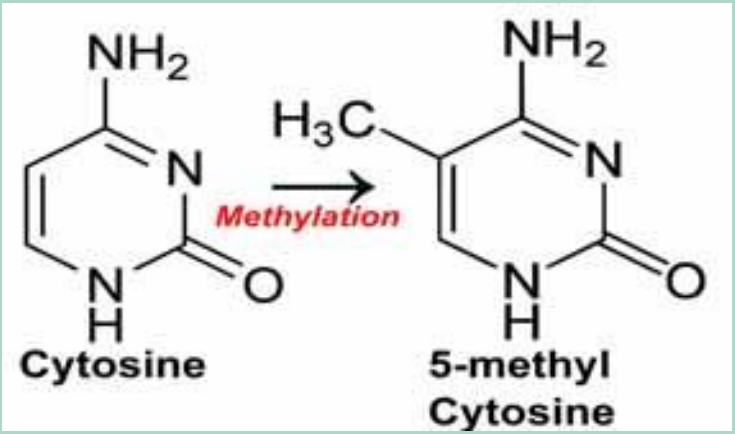
Science Foundation Ireland sfi For what's next

SFI Centre for Research Training in Foundations of Data Science

UCD DUBLIN

Necessity for statistical models for analysing DNA methylation data

- An epigenetic change where a methyl group is added or removed from the 5’ carbon of the cytosine ring.
- The cytosine-guanine dinucleotide (CpG) sites which remain unmethylated in normal cells can get methylated in abnormal cells such as cancer cells.
- Identifying the differentially methylated regions (DMR) between the benign and malignant tissue samples can help in the diagnosis of the disease and its treatment.
- The Illumina MethylationEPIC BeadChip microarray can be used for methylation profiling of 866,830 CpG sites in the human genome.
- Methylation states: Hypomethylation, Hemimethylation and Hypermethylation.
- Aim is to analyze such DNA methylation data collected by Silva et al.[1] from benign and tumor prostate tissues of 4 patients.
- Develop beta mixture models (BMM) to uncover groups of CpG sites with similar methylation profiles in order to identify DMRs in an efficient manner and obviate the need for arbitrary thresholds to identify the methylation states.



https://www.well.ox.ac.uk/ogc/epigenetic-analysis-on-illumina-epic-arrays/fig-1-cytosine-methylation/

Beta mixture models

- Data: ***C*** CpG sites from ***N*** patients' ***R*** tissue samples belongs to ***K*** groups.
- The log-likelihood function for the generalized mixture model is,

$$L(\theta|X) = \sum_{c=1}^{C} \log[\sum_{k=1}^{K} \tau_k \prod_{n=1}^{N} \prod_{r=1}^{R} Beta(x_{cnr}; \alpha_{knr}, \delta_{knr})]$$

where,
θ is the parameters to be estimated for the model, and
X is the dataset of beta values for the CpG sites

- K-means clustering is used to obtain an initial clustering and the method of moments is used to obtain initial values for the EM algorithm.
- Digamma function Approximation: The M-step solution in EM algorithm is not available in closed form; hence the parameter estimation using numerical approximation is inefficient for such high-dimensional data.
- The lower bound approximation for the digamma function is used to obtain a closed form solution [2] which is valid for all y > 1/2:

$$\psi(y) > \log(y - \frac{1}{2})$$

- The Bayesian Information Criterion (BIC) [3] is used to select the optimal model and the adjusted Rand index [4] is used for measurement of agreement between the different models.

Family of mixture models

C.. Model

Assumptions:
- ***K =3 , R = 1***
- The BMM parameters (α,δ) are constant, and each patient belongs to the same mixture model.

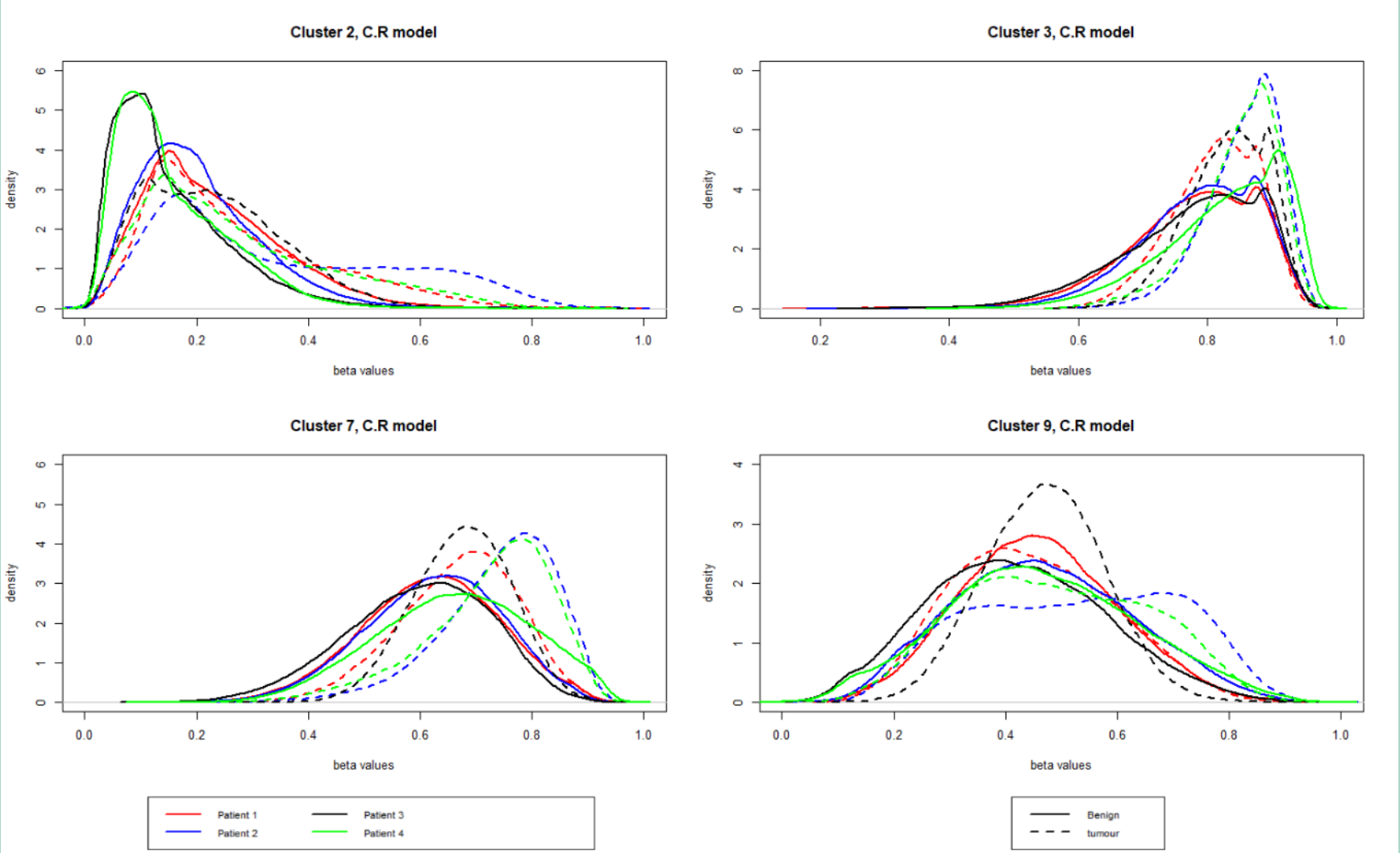
CN. Model

Assumptions:
- ***K = 3, R = 1***
- Each patient belongs to a different BMM which results in varying BMM parameters for each patient

C.R Model

Assumptions:
- ***K = 3, R = 2***.
- The samples belong to ***R*** number of BMM and the K^R groups obtained for changes in methylation state between the given samples.

Results: C.R model



- The DNA samples from benign and tumor cells of prostate tissue are obtained from 4 patients.
- The BMM approach accurately identifies more DMRs than conventional methods and is more computationally efficient than other proposed beta mixture models.
- Additional 110 DMRs related to prostate cancer gene were obtained.
- Non-parametric test results suggested the beta value of the RARB, APC and SFRP2 genes were greater in the tumor samples than in the benign samples for all patients.

References

[1] Silva, R., Moran, B., Russell, N.M., Fahey, C., Vlajnic, T., Manecksha, R.P., Finn, S.P., Brennan, D.J., Gallagher, W.M., Perry, A.S.: Evaluating liquid biopsies for methylomic profiling of prostate cancer. Epigenetics 15(6-7), 715-727 (2020), doi: 10.1080/15592294.2020.1712876.

[2] Diamond, H.G., Straub, A.: Bounds for the logarithm of the Euler gamma function and its derivatives. Journal of Mathematical Analysis and Applications 433(2), 1072-1083 (2016), doi:10.1016/j.jmaa.2015.08.034.

[3] Schwarz, G.: Estimating the dimension of a model. Annals of Statistics 6, 461-464 (1978)

[4] Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification 2, 193-218 (1985), doi:10.1007/BF01908075.