betaclust: a family of mixture models for beta valued DNA methylation data

Koyel Majumdar[1], Romina Silva[2,3], Antoinette Sabrina Perry[2,3], Ronald William Watson[3], Thomas Brendan Murphy[1] and Isobel Claire Gormley[1*]

Abstract

Background: The DNA methylation process has been extensively studied for its role in cancer. Promoter cytosine-guanine dinucleotide (CpG) island hypermethylation has been shown to silence tumour suppressor genes. The methylation state of a CpG site is hypermethylated if both the alleles are methylated, hypomethylated if neither of the alleles are methylated and hemimethylated otherwise. Identifying the differentially methylated CpG (DMC) sites between benign and tumour samples can help understand the disease.

The Illumina MethylationEPIC BeadChip microarray quantifies the methylation level at a CpG site as a beta value which lies within [0,1). There is a lack of suitable methods for modelling the beta values in their innate form. For this reason, the beta values are usually transformed into M-values for analysis. The DMCs are identified using M-values or beta values via multiple t-tests but this can be computationally expensive. Also, arbitrary thresholds are often selected and used to identify the methylation state of a CpG site.

We propose a family of novel beta mixture models (BMMs) which use a model-based clustering approach to cluster the CpG sites in their innate beta form to (i) objectively identify methylation state thresholds and (ii) identify the DMCs between different samples. The family of BMMs employs different parameter constraints that are applicable to different study settings. Parameter estimation proceeds via an Expectation-Maximisation algorithm, with a novel approximation during the M-step providing tractability and computational feasibility.

Results: Performance of the BMMs is assessed through a thorough simulation study, and the BMMs are used to analyse a prostate cancer dataset and an esophageal squamous cell carcinoma dataset. The BMM approach objectively identifies methylation state thresholds and identifies more DMCs between the benign and tumour samples in the prostate and esophageal cancer data than conventional methods, in a computationally efficient manner. The empirical cumulative distribution function of the DMCs related to genes implicated in carcinogenesis indicates hypermethylation of CpG sites in the tumour samples in both cancer settings.

Conclusion: An R package `betaclust` is provided to facilitate the widespread use of the developed BMMs to provide objective thresholds to determine methylation state and to computationally efficiently identify DMCs by clustering DNA methylation data in its innate form.

Keywords: DNA methylation; model-based clustering; beta mixture model; digamma function; EM algorithm.

Background

Epigenetics is the study of heritable changes in gene activity that do not involve any explicit change to the DNA sequence [1]. DNA methylation is one of the epigenetic processes where a methyl group is added to or removed from the 5′ carbon of the cytosine ring [2]. This process assists in regulating gene expression and is essential for the development of an organism, but irregular changes in DNA methylation patterns can lead to damaging health effects [3]. The DNA methylation process has been extensively studied in the context of cancer, and its treatment [4, 5]. It has been observed that cytosine-guanine dinucleotide (CpG) islands that remain unmethylated in normal cells can become methylated in abnormal cells such as cancer cells [6]. A CpG site is hypomethylated if neither of the alleles are methylated, hemimethylated if either of the alleles are methylated or hypermethylated if both the alleles are methylated. A differentially methylated region (DMR) is a genomic region that has different methylation states between different samples, which may have been taken from tissues of an individual over time, different tissues from the same individuals or other individuals [7]. In several cancer studies, it has been observed that tumour suppressor genes are silenced by hypermethylation of promoter regions of those genes [8–10]. A better understanding of the disease can be achieved by identifying the regions that are differentially methylated between the benign and the tumour samples.

The Illumina MethylationEPIC BeadChip microarray [11] is used to interrogate over 850,000 CpG sites and retrieve methylation profiling of the CpG sites in the human genome. The Illumina microarray produces a sample of methylated (M) and unmethylated (U) light signal intensities, and the level of methylation is measured as a ratio of these intensities. The *beta* value is used to quantify the methylation level and is calculated as $beta = \max(M)/(\max(M) + \max(U) + \chi)$, where χ is a constant offset added to regularise the values for very low M and U values [12]. The methylation level at a CpG site is quantified by this *beta* value and is constrained to lie between 0 and 1 as it measures the proportion of methylated and unmethylated signal intensities. The *beta* values are continuous and a value close to 1 suggests that the site is hypermethylated, while values close to 0 represent hypomethylation. The two probe intensities are assumed to be gamma-distributed as they can take only positive values, and their ratio results in beta distributed variables. Thus, the *beta* values can be modelled using a beta distribution.

The *beta* values in general have higher variance in the center of the [0,1) interval than towards the two endpoints. This results in the data being heteroscedastic, which imposes serious challenges for analysing them as they violate the assumptions for the ubiquity of Gaussian models. Hence, the *beta* values are usually converted to M-values using a logit transformation, $M = \log(beta/(1 - beta))$ as these values are statistically more convenient. For the M-values, a positive value represents hypermethylation, whereas a negative value represents hypomethylation of a CpG site. The transformed data can be modelled using Gaussian models as the data are no longer bounded and lie within $(-\infty, \infty)$ [12]. However, transformation make the results less biologically interpretable. Hence there is a need to statistically model the *beta* values in their innate form.

In many studies, thresholds of *beta* values are subjectively selected to identify the three methylation states [13]. A *beta* value < 0.2 is often used to suggest hypomethylation of a CpG site whereas a value > 0.8

*Correspondence: claire.gormley@ucd.ie

[1]School of Mathematics and Statistics, University College Dublin, Belfield, Dublin 4, Ireland

Full list of author information is available at the end of the article

is used to suggest hypermethylation. This is because the *beta* values within the interval [0.2,0.8] are approximately linearly related to the M-values [12]. An objective approach to determining thresholds between methylation states is required.

For bounded data, a transformation-based approach to density estimation using the Gaussian mixture model has been proposed [14]. The bounded data are first transformed using a range-power transformation, and then the density is estimated using the Gaussian mixture model. While this approach makes the use of Gaussian mixture models feasible for bounded data, the results are less biologically interpretable than if the untransformed data had been modelled directly.

Several studies have used mixture models to model a subset of CpG sites and cluster samples into latent groups of biologically related samples [15–20]. A small number of CpG sites related to the genes of interest are analysed in these studies, and it is difficult to computationally scale these models to analyse data from the complete microarray. The `Methylmix` [21] R [22] package uses a univariate beta mixture model to uncover patient subgroups with similar DNA methylation levels for a specific CpG site. Methylation levels then are compared to mean DNA methylation levels of normal tissue samples and a Wilcoxon rank sum test is used to establish hyper and hypomethylated genes relative to normal. Other studies identify DMCs using M-values or *beta* values via multiple t-tests or Wilcoxon rank-sum tests [23, 24]. A nonparametric test, based on Cuzick test [25] calculating p-values for each CpG site individually has been proposed for modelling *beta* values to identify DMCs between multiple treatments [26]. It is efficient in analyzing smaller arrays but is computationally intensive for arrays containing more than 28,000 CpG sites.

We propose a novel family of beta mixture models (BMMs) which use a model-based clustering approach to cluster the CpG sites in their innate beta form to (i) objectively identify methylation state thresholds and (ii) identify the DMCs between different samples. The DMRs of interest can then be retrieved from the identified DMCs. The BMMs are capable of clustering the entire microarray of CpG sites in a computationally efficient manner. Performance is assessed through a thorough simulation study, and the BMMs are used to analyse a prostate cancer dataset and an esophageal squamous cell carcinoma dataset. The capability of the BMMs to model the *beta* values, to objectively identify thresholds and efficiently cluster all the CpG sites to identify DMCs is demonstrated.

Data

Simulated data

To assess the performance of the proposed BMM method, a thorough simulation study is conducted. Twenty simulated datasets consisting of methylation values for $C = 20,000$ CpG sites are generated. While the real cancer datasets analysed here have a larger value of C, due to computational constraints $C = 20,000$ is used in the simulation studies. Each simulated dataset consists of *beta* values from two samples (sample A and sample B) from each of $N = 4$ patients. Hypomethylated CpG sites were generated from a $Beta(2, 20)$ distribution, with hemi- and hyper- values generated from $Beta(4, 3)$ and $Beta(20, 2)$ distributions respectively. Reflecting what is typically observed in DNA methlyation data, the proportion of CpG sites that were simulated as hypomethylated, hemimethylated and hypermethylated was 0.35, 0.35 and 0.3 respectively. This resulted in 29% of the CpG sites being differentially methylated i.e. hypomethylated in one sample and hypermethylated in the other, or vice versa.

Cancer data

Triggering of cancer and disease advancement is related to epigenetic changes such as hypermethylation of target genes resulting in gene inactivation. Hypermethylation of certain tumour suppressor genes has been observed during the early stages of prostate cancer disease [27]. In comparison with benign esophageal mucosa, the esophageal squamous cell carcinoma (ESCC) genome was observed to contain focal areas of hypermethylation and widespread areas of hypomethylation in tumour samples [28]. The identification of methylation changes in target genes can help in early cancer diagnosis. Moreover, changes in methylation patterns during treatment can assist in understanding the effectiveness of different therapeutic approaches.

Here, DNA methlyation data from a prostate cancer study and from an ESCC study are appositely modeled in their innate beta form to (i) objectively identify methylation state thresholds and (ii) uncover DMCs between the two samples in each cancer setting using a model-based clustering approach.

Prostate cancer data

Prostate cancer (PCa) is the fifth major cause of cancer-related mortality globally [29]. DNA methylation samples were collected for a study of methylation profiling [30]. The study cohort consisted of four patients with metastatic prostate cancer disease. Tissue samples from matched biopsy cores (tumour and histologically matched normal – herein benign) were collected from each patient, and DNA was extracted from these samples. Methylation profiling of these DNA samples was done using the Infinium MethylationEpic Beadchip [31]. The methylation array data consisted of 694,923 CpG sites and *beta* values for each CpG site for the two DNA samples were collected from each of the four patients. The raw methylation array data was quality controlled and pre-processed using the RnBeads [32] R package [30]. Further, 103 CpG sites with missing *beta* values were removed from the dataset. The resulting prostate cancer dataset therefore contained *beta* values for $C = 694{,}820$ CpG sites from each of $R = 2$ DNA samples collected from each of $N = 4$ patients.

Esophageal squamous cell carcinoma data

Esophageal squamous cell carcinoma (ESCC) is a subtype of esophageal cancer characterized by aberrant DNA methylation. A study was conducted to investigate abnormal genes in ESCC, and DNA samples were collected from 15 patients' benign and tumour tissues [33]. Paired samples from 4 randomly selected patients were considered here. After the removal of 6,446 CpG sites with missing *beta* values, the ESCC dataset contained *beta* values for $C = 474,869$ CpG sites from each of $R = 2$ DNA samples collected from each of $N = 4$ patients.

Method

The beta distribution has support on the interval $[0, 1]$ and is parameterized by two positive shape parameters, α and δ. If $\alpha > 1$ and $\delta > 1$ then the distribution is unimodal and if the parameters are < 1 then the distribution is bimodal. A uniform distribution is obtained if $\alpha = 1$ and $\delta = 1$.

Given the properties of the *beta* values, the beta distribution is used here to appositely model the methylation level x_{cnr} of the c^{th} CpG site ($c = 1, \ldots, C$), from the n^{th} patient ($n = 1, \ldots, N$), from their r^{th} DNA sample ($r = 1, \ldots, R$) i.e., for $0 \leq x_{cnr} \leq 1$:

$$f(x_{cnr}|\alpha, \delta) \sim \text{Beta}(x_{cnr}|\alpha, \delta) = \frac{{x_{cnr}}^{\alpha-1}(1 - x_{cnr})^{\delta-1}}{\text{B}(\alpha, \delta)},$$

where $\text{B}(\alpha, \delta) = (\Gamma(\alpha)\Gamma(\delta))/\Gamma(\alpha + \delta)$ is the beta function, defined in terms of the gamma function $\Gamma(\cdot)$.

The DNA methylation data are collected in the multivariate dataset $\mathbf{X}$ of dimension $C \times NR$ where each

of the NR columns contains the methylation levels of the C CpG sites in one of the R samples from each of the N patients.

A beta mixture model

A mixture model assumes the observed data have been generated from a heterogeneous population. Each CpG site is assumed to have been generated by one of the K groups or clusters in the heterogeneous population using a probabilistic model. The parameter $\boldsymbol{\theta}$ is used here to denote all the shape parameters in a beta mixture model, i.e., $\boldsymbol{\theta} = (\boldsymbol{\alpha}_1, \boldsymbol{\delta}_1, ..., \boldsymbol{\alpha}_K, \boldsymbol{\delta}_K)$, where $\boldsymbol{\alpha_k}$ and $\boldsymbol{\delta_k}$ are the shape parameters of cluster k. The mixing proportions $\boldsymbol{\tau} = (\tau_1, ..., \tau_K)$ lie between 0 and 1, $\sum_{k=1}^{K} \tau_k = 1$, and denote the probability of belonging to cluster k $\forall k = 1, \ldots, K$. Independence is assumed across patients and samples, given a CpG site's cluster membership, leading to the probability density function for such a beta mixture model (BMM):

$$\begin{aligned} f(\mathbf{X}|\boldsymbol{\tau}, \boldsymbol{\theta}) &= \prod_{c=1}^{C} \sum_{k=1}^{K} \tau_k f(\mathbf{X}|\boldsymbol{\alpha}_k, \boldsymbol{\delta}_k) \\ &= \prod_{c=1}^{C} \sum_{k=1}^{K} \tau_k \prod_{n=1}^{N} \prod_{r=1}^{R} \text{Beta}(x_{cnr}|\alpha_{knr}, \delta_{knr}) \end{aligned} \tag{1}$$

The parameters $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$ of the BMM can be estimated by maximising the associated log-likelihood function:

$$\begin{aligned} \ell(\boldsymbol{\tau}, \boldsymbol{\theta}|\mathbf{X}) = \sum_{c=1}^{C} \log \Bigg[&\sum_{k=1}^{K} \tau_k \\ &\prod_{n=1}^{N} \prod_{r=1}^{R} \text{Beta}(x_{cnr}; \alpha_{knr}, \delta_{knr}) \Bigg] \end{aligned} \tag{2}$$

The direct computation of maximum likelihood estimates (MLEs) of $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$ from (2) is complex, and an incomplete data approach is therefore used here. The latent binary vector $\mathbf{z}_c = (z_{c1}, \ldots, z_{cK})$ is introduced for each CpG site c, where z_{ck} is 1 if CpG site c belongs to the k^{th} group and 0 otherwise. The $C \times K$ matrix $\mathbf{Z}$ is combined with the associated *beta* values for each CpG site to form the complete data $(\mathbf{X}, \mathbf{Z})$. Therefore, the complete data log-likelihood function is

$$\begin{aligned} \ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}, \mathbf{Z}|\mathbf{X}) = &\sum_{c=1}^{C} \sum_{k=1}^{K} z_{ck} \{\log \tau_k + \\ &\sum_{n=1}^{N} \sum_{r=1}^{R} \log[\text{Beta}(x_{cnr}; \alpha_{knr}, \delta_{knr})]\}. \end{aligned} \tag{3}$$

The complete data log-likelihood function (3) can be used to find the MLEs $\hat{\boldsymbol{\tau}}$ and $\hat{\boldsymbol{\theta}}$ using the Expectation-Maximisation (EM) algorithm [34]. Further, a probabilistic clustering solution is available from the expected value of z_{ck}, the posterior probability of CpG site c belonging to cluster k, provided on convergence of the EM algorithm.

A family of BMMs

Each CpG site is assumed to have one of the $M = 3$ methylation states: hypomethylation, hemimethylation or hypermethylation. The most generalised BMM is defined in (1) which models the CpG sites as belonging to K latent groups. By introducing a variety of constraints on the parameters of this generalised BMM, a family of three beta mixture models is developed. Each of the 3 models serves a specific purpose e.g., to cluster the CpG sites into the 3 methylation states allowing objective inference of methylation state thresholds or facilitating the identification of DMCs between different samples, as detailed below.

The K·· model The K·· model facilitates clustering of the C CpG sites into one of $K = M$ methylation states, based on a single sample ($R = 1$) from each of N patients. Under the K·· model the shape parameters of each cluster are constrained to be equal for each patient. Thus, for the K·· model the complete data

log-likelihood function is

$$\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}, \mathbf{Z}|\mathbf{X}) = \sum_{c=1}^{C}\sum_{k=1}^{K} z_{ck}\{\log \tau_k + \sum_{n=1}^{N}\sum_{r=1}^{1} \log[\text{Beta}(x_{cnr}; \alpha_{k\cdot\cdot}, \delta_{k\cdot\cdot})]\}.$$

The K·· model is used here to identify the methylation state of each CpG site in a single sample, allowing the objective inference of thresholds between methylation states.

The KN· model The KN· model is used to cluster each of the C CpG sites into one of $K = M$ methylation states, based on data from a single sample ($R = 1$) from each of N patients. While the KN· model has a similar purpose to the K·· model, it differs in that it is less parsimonious as it allows cluster and patient-specific shape parameters. The complete data log-likelihood function for the KN· model is

$$\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}, \mathbf{Z}|\mathbf{X}) = \sum_{c=1}^{C}\sum_{k=1}^{K} z_{ck}\{\log \tau_k + \sum_{n=1}^{N}\sum_{r=1}^{1} \log[\text{Beta}(x_{cnr}; \alpha_{kn\cdot}, \delta_{kn\cdot})]\}.$$

Similar to the the K·· model, the KN· model also allows inference on the methylation state of each CpG site in a sample and the objective inference of thresholds between the 3 methylation states.

The K·R model The K·R model facilitates the identification of differentially methylated CpG sites between R DNA samples collected from each of N patients. The K·R model assumes $K = M^R$ clusters, each identifying one combination of the M methylation states in the R samples. The shape parameters are allowed to vary for each sample type but are constrained to be equal for each patient. The complete data log-likelihood function for the K·R model is

$$\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}, \mathbf{Z}|\mathbf{X}) = \sum_{c=1}^{C}\sum_{k=1}^{K} z_{ck}\{\log \tau_k + \sum_{n=1}^{N}\sum_{r=1}^{R} \log[\text{Beta}(x_{cnr}; \alpha_{k\cdot r}, \delta_{k\cdot r})]\}.$$

This model enables the identification of DMCs between R DNA samples as illustrated in the simulation studies and applications that follow.

Parameter estimation

For each of the 3 models in this family of BMMs, the parameters can be estimated and the cluster membership for each CpG site inferred using the EM algorithm. Here, we illustrate this procedure for the most general BMM given in (1). Parameter estimation derivations for the K··, KN· and K·R models are detailed in Appendices 1–3.

The EM algorithm consists of two steps which are iterated until convergence. In the expectation step (E-step) the expected value of the complete data log-likelihood function is obtained, conditional on the observed data and the current parameter estimates. The maximisation step (M-step) maximises the expected complete data log-likelihood function with respect to the parameters. To obtain the parameter estimates $\hat{\boldsymbol{\tau}}$ and $\hat{\boldsymbol{\theta}}$, the E and M-steps are iterated until convergence to at least a local optimum i.e. until

$$\ell(\boldsymbol{\tau}^{(t+1)}, \boldsymbol{\theta}^{(t+1)}|\mathbf{X}) - \ell(\boldsymbol{\tau}^{(t)}, \boldsymbol{\theta}^{(t)}|\mathbf{X}) < \epsilon$$

where ϵ is an arbitrarily chosen small value.

An initial clustering of the C CpG sites is obtained using k-means clustering and the method of moments is then used to calculate the initial values of $\boldsymbol{\tau}, \boldsymbol{\alpha}_k$ and $\boldsymbol{\delta}_k$ provided to the EM algorithm. The E and M steps then proceed as follows:

- E-step: the expected value of z_{ck} is calculated, i.e. the posterior probability of CpG site c belonging

to the k^{th} cluster, conditional on the current parameter estimates. At iteration $t+1$

$$\begin{aligned}\hat{z}_{ck} =& \mathbf{E}[z_{ck}|\mathbf{X}, \boldsymbol{\tau}^{(t)}, \boldsymbol{\theta}^{(t)}] \\ =& \frac{\tau_k^{(t)} \prod_{n=1}^{N} \prod_{r=1}^{R} \text{Beta}(x_{cnr}; \alpha_{knr}^{(t)}, \delta_{knr}^{(t)})}{\sum_{k'=1}^{K} [\tau_{k'}^{(t)} \prod_{n=1}^{N} \prod_{r=1}^{R} \text{Beta}(x_{cnr}; \alpha_{k'nr}^{(t)}, \delta_{k'nr}^{(t)})]}.\end{aligned}$$

- M-step: estimates of the parameters $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$ are calculated by maximising the expected complete data log-likelihood function, given the $\hat{\mathbf{Z}}$ values from the E-step.

For the M-step, the expected complete data log-likelihood function is maximised by differentiating it w.r.t the parameters. Solutions for the mixing proportions are available in closed form as

$$\hat{\tau}_k = \sum_{c=1}^{C} \hat{z}_{ck}/C, \ \forall \ k = 1, \ldots, .K.$$

For the shape parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$, the expected complete data log-likelihood function to be optimized is

$$\begin{aligned}\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}|\mathbf{X}, \hat{\mathbf{Z}}) =& \sum_{c=1}^{C} \sum_{k=1}^{K} \hat{z}_{ck} \{\log \tau_k + \\ & \sum_{n=1}^{N} \sum_{r=1}^{R} [(\alpha_{knr} - 1) \log x_{cnr} + \\ & (\delta_{knr} - 1) \log(1 - x_{cnr}) - \\ & \log \text{B}(\alpha_{knr}, \delta_{knr})]\}.\end{aligned} \tag{4}$$

Differentiating (4) w.r.t α_{knr} yields

$$\begin{aligned}\frac{\partial \ell_C}{\partial \alpha_{knr}} =& \sum_{c=1}^{C} \hat{z}_{ck} \{\log x_{cnr} - [\psi(\alpha_{knr}) - \\ & \psi(\alpha_{knr} + \delta_{knr})]\}\end{aligned} \tag{5}$$

where ψ is the logarithmic derivative of the gamma function known as the digamma function, $\psi(\alpha_{knr}) = \partial \log \Gamma(\alpha_{knr})/\partial \alpha_{knr}$.

Similarly, the derivative of $\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}|\mathbf{X}, \hat{\mathbf{Z}})$ w.r.t δ_{knr} is,

$$\begin{aligned}\frac{\partial \ell_C}{\partial \delta_{knr}} =& \sum_{c=1}^{C} \hat{z}_{ck} \{\log(1 - x_{cnr}) - \\ & [\psi(\delta_{knr}) - \psi(\alpha_{knr} + \delta_{knr})]\}.\end{aligned} \tag{6}$$

Closed form solutions for $\hat{\alpha}_{knr}$ and $\hat{\delta}_{knr}$ are not available due to the presence of the digamma function. To obtain the maximised parameter estimates, numerical optimisation algorithms such as BFGS [35] and BHHH [36] could be used. However, for the large datasets considered here, use of these algorithms proved to be computationally infeasible.

A digamma approximation

Here an approximation to the digamma function is used to allow for closed form solutions for the shape parameters. The lower bound for the digamma function for all $y > 1/2$ [37] is

$$\psi(y) > \log(y - 1/2). \tag{7}$$

Given the context, in our family of BMMs we assume that the beta distributions are unimodal and bounded, meaning the shape parameters are greater than 1. Thus, the lower bound approximation holds and was empirically observed to be a very close approximation of the digamma function. Thus the lower bound is used in (5) and (6) resulting in closed-form solutions at the M-step of the EM algorithm i.e.,

$$\begin{aligned}\frac{\partial \ell_C}{\partial \alpha_{knr}} \approx& \sum_{c=1}^{C} \hat{z}_{ck} \sum_{n=1}^{N} \sum_{r=1}^{R} \Bigg[\log x_{cnr} - \\ & \log \frac{\alpha_{knr} - 1/2}{\alpha_{knr} + \delta_{knr} - 1/2}\Bigg]\end{aligned} \tag{8}$$

and

$$\begin{aligned}\frac{\partial \ell_C}{\partial \delta_{knr}} \approx& \sum_{c=1}^{C} \hat{z}_{ck} \sum_{n=1}^{N} \sum_{r=1}^{R} \Bigg[\log(1 - x_{cnr}) - \\ & \log \frac{\delta_{knr} - 1/2}{\alpha_{knr} + \delta_{knr} - 1/2}\Bigg].\end{aligned} \tag{9}$$

Equating equations (8) and (9) to zero, we get the approximate estimates of α_{knr} and δ_{knr} as,

$$\hat{\alpha}_{knr} = 0.5 + \frac{0.5\exp(-y_2)}{\{[\exp(-y_2)-1][\exp(-y_1)-1]\}-1}$$

and

$$\hat{\delta}_{knr} = \frac{0.5\exp(-y_2)[\exp(-y_1)-1]}{\{[\exp(-y_2)-1][\exp(-y_1)-1]\}-1},$$

where $y_1 = (\sum_{c=1}^{C} \hat{z}_{ck} \log x_{cnr})/(\sum_{c=1}^{C} \hat{z}_{ck})$ and

$y_2 = (\sum_{c=1}^{C} \hat{z}_{ck} \log(1-x_{cnr}))/(\sum_{c=1}^{C} \hat{z}_{ck})$.

Calculating methylation state thresholds

To objectively derive the thresholds between methylation states, without loss of generality, we denote by clusters 1 and 2 the clusters representing hypomethylated and hypermethylated CpG sites respectively. The ratio of fitted density estimates ω_j for cluster $j = 1, 2$ is calculated as

$$\omega_j = \frac{\tau_j f(\mathbf{X}|\boldsymbol{\alpha}_j, \boldsymbol{\delta}_j)}{\sum_{k \neq j} \tau_k f(\mathbf{X}|\boldsymbol{\alpha}_k, \boldsymbol{\delta}_k)}. \quad (10)$$

The threshold separating e.g., the hypomethylated and hemimethylated clusters is calculated as the minimum beta value at which $\omega_1 \geq 1$. Similarly, the threshold dividing the hemimethylated and hypermethylated clusters is calculated to be the maximum beta value at which the $\omega_2 \geq 1$.

In the K·· model, as the shape parameters are constrained to be equal for each patient, a single set of thresholds is calculated for all patients. As the shape parameters vary for each patient in the KN· model, a set of thresholds is calculated for each patient.

Optimal model assessment

Comparison of the fit of the developed models is possible through use of e.g., the Akaike information criterion (AIC) [38], the Bayesian information criterion (BIC) [39], or the integrated complete log-likelihood criterion (ICL) [40, 41]. The model which minimises the AIC, BIC and/or ICL is selected as the optimal model. The AIC and BIC values are defined as

$$AIC = 2Q - 2\log(\hat{L})$$
$$BIC = Q\log(C) - 2\log(\hat{L}),$$

where $\hat{L}$ is the maximised value of the likelihood function and Q is the number of parameters in the model.

The ICL penalizes the BIC by including an entropy term favouring well separated clusters. The ICL is defined as

$$ICL = BIC + 2\sum_{c=1}^{C}\sum_{k=1}^{K} g_{ck}\log(\hat{z}_{ck}),$$

where $g_{ck} = 1$ if the c^{th} CpG site belongs to the k^{th} cluster and 0 otherwise.

In what follows, the adjusted Rand index (ARI) [42] is used to obtain a measure of agreement between different clustering solutions where a value of 1 suggests the two clustering solutions are in full agreement.

Results

Simulated data results

Estimating methylation state thresholds

The first objective of the simulation study was to cluster the CpG sites in a sample into 3 clusters representing the 3 methylation states and then to calculate the threshold points between these states. To achieve this, the K·· and KN· models were fitted to the data from sample A in each of the 20 simulated datasets with the true K·· generating model selected by BIC to be the optimal model in each case. In Figure 1 the density estimates under the clustering solution of the K·· model for a single simulated dataset are displayed. The hypomethylated CpG sites are clustered in cluster 1, while the hypermethylated and hemimethylated CpG sites are in clusters 2 and 3 respectively. The proportion of CpG sites belonging to each cluster is also displayed in Figure 1 which are notably very close to

the true mixing proportions. As parameters are constrained to be equal for each patient in the K·· model a single set of thresholds is calculated for all 4 patients. The threshold point of 0.242 indicates that any CpG site with a lower beta value is likely to be hypomethylated. Similarly any CpG site with a beta value greater than the second threshold point of 0.808 is likely to be hypermethylated. These objectively inferred threshold points are very close to the true threshold points of 0.244 and 0.808.

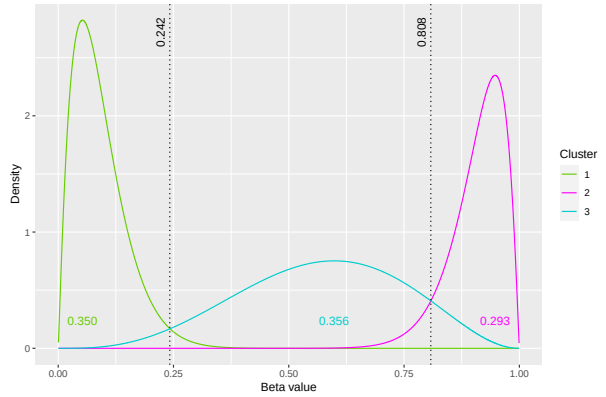

Figure 1: Fitted density estimates under the clustering solution of the K·· model on a simulated dataset from sample A. The thresholds between methylation states are illustrated by the black dotted lines. The proportion of CpG sites belonging to each cluster is also displayed.

Table 1: Contingency table under the K·· model for CpG sites from patients in sample A.

		Cluster		
		1	2	3
Methylation state	Hypermethylation		5866	12
	Hemimethylation		8	7109
	Hypomethylation	7004		1

The true methylation state and estimated cluster membership are presented in Table 1 for a single simulated dataset, with an ARI of 99.7%. An ARI of 1 was obtained when comparing the clustering solutions of the two fitted models. The mean ARI across the 20 simulated datasets for the K·· model was 0.996 (standard deviation 0.00057) and for the KN· model was 0.996 (standard deviation 0.00055), demonstrating accurate and stable clustering solutions. The mean ARI when comparing the K·· and KN· clustering solutions was 0.999 (standard deviation 0.0001). A summary of the parameter estimates and kernel density plots under the K·· model are available in Appendices 4–5.

Identifying DMCs

To identify differentially methylated CpG sites between multiple DNA samples in the simulated data, the K·R model is fitted. For each CpG site, as there are $R = 2$ sample types in each simulated dataset, $K = M^R = 9$ different combinations of methylation states are possible across samples A and B. The CpG sites that are hypomethylated in one sample and hypermethylated in the other sample and vice versa, are of prime interest as they indicate potential epigenetic changes in the cancer genome. Table 2 shows the contingency table of the clustering solution for a single simulated dataset against the true cluster memberships, with an ARI of 0.929. The fitted density estimates of the clustering solution for this single simulated dataset are shown in Figure 2. The density estimates show that cluster 1 captures DMCs which are hypermethylated in sample A and hypomethylated in sample B while cluster 2 contains DMCs which are hypomethylated in sample A and hypermethylated in sample B. The proportions of CpG sites clustered into clusters 1 and 2 are equal to the true mixing proportions. The mean ARI across all 20 simulated datasets was 0.908 (standard deviation 0.05), indicating accurate and stable clustering. A summary of the parameter estimates under the fitted K·R model is available in Appendix 4 and kernel density estimates are in Appendix 6.

Table 2: Contingency table under the K·R model to detect DMCs in a single simulated dataset.

	Cluster								
Methylation state change	1	2	3	4	5	6	7	8	9
Hypermethylation - Hemimethylation							1993	14	
Hypermethylation - Hypermethylation						30	975		
Hypermethylation - Hypomethylation	2861		5						
Hemimethylation - Hypomethylation	3		1335		661				
Hypomethylation - Hemimethylation		7		979					
Hemimethylation - Hemimethylation						5	2	2967	
Hypomethylation - Hypomethylation									3060
Hypomethylation - Hypermethylation		2958		1					
Hemimethylation - Hypermethylation						2140		4	

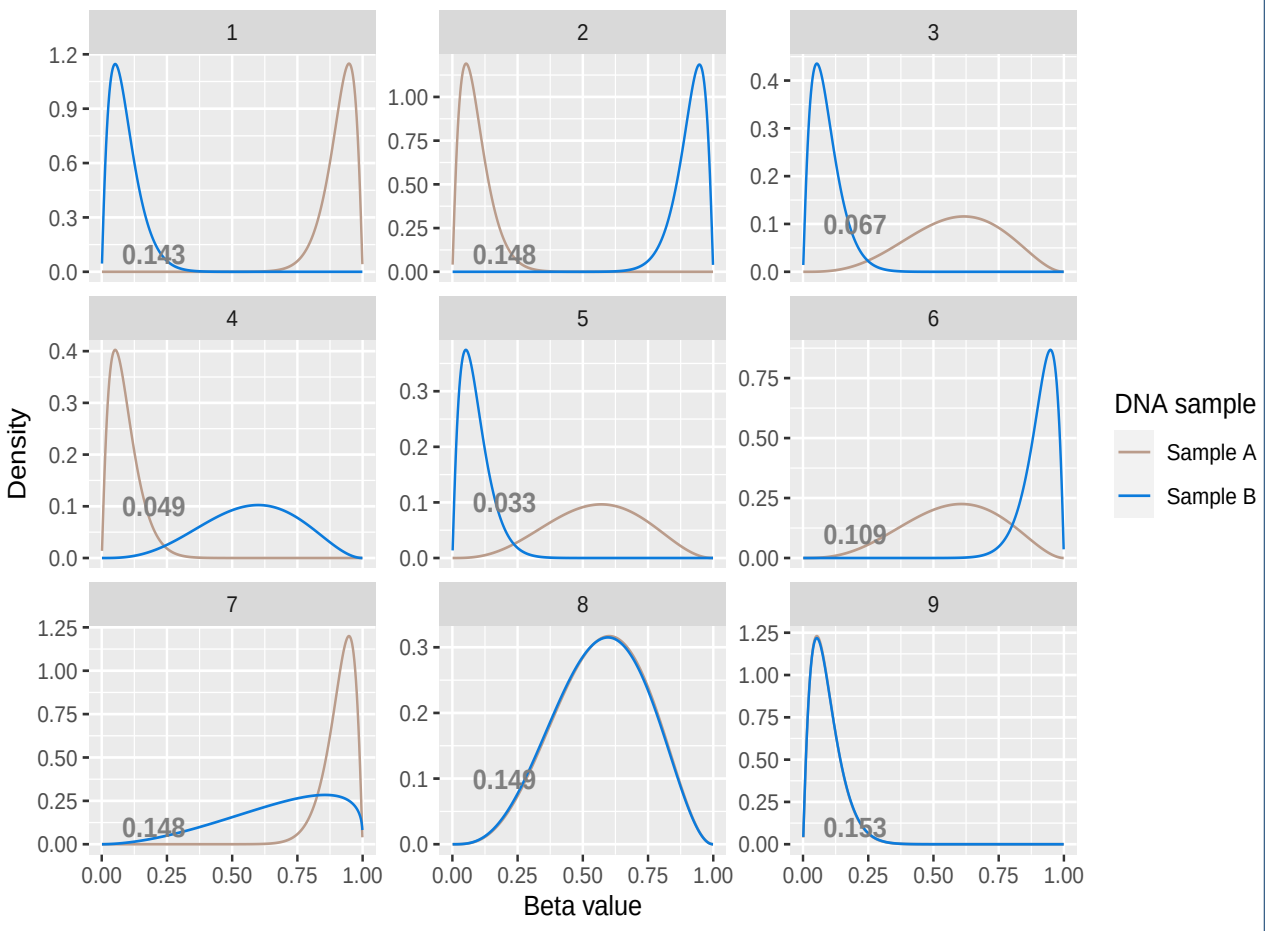

Figure 2: Fitted density estimates under the K·R model on a simulated dataset. The proportion of CpG sites belonging to each of the 9 clusters is displayed in the relevant panel.

Prostate cancer data results

Estimating methylation state thresholds

The PCa dataset has *beta* values for each of $C = 694{,}820$ CpG sites from $R = 2$ DNA samples collected from $N = 4$ patients. The cluster the CpG sites into the 3 known methylation states and objectively infer

the methylation state thresholds in the benign samples and in the tumour samples, the K·· and KN· models are fitted. The BIC suggests that the KN· model is optimal for both the benign and the tumour samples. This is intuitive, particularly for the tumour samples where the degree of disease varies for each patient, as the KN· model allows for patient specific shape parameters.

The fitted density estimates under the KN· model for the benign sample collected from patient 1 are displayed in Figure 3. The inferred methylation state thresholds are 0.258 and 0.747 for the benign sample, and 0.19 and 0.751 for the tumour sample collected from patient 1. The hypermethylation state thresholds in the benign and tumour samples are very close; in contrast, the hypomethylation state thresholds are quite different. While these objective thresholds are close to the subjective values suggested in the literature of 0.2 and 0.8, the difference results in more hypo- and hypermethylated CpG sites being identified by the BMM as DMCs. The ARIs between the KN· and K·· solutions of 0.94 for the benign sample and 0.96 for the tumour sample indicate good clustering agreement between the two models. The kernel density estimates under the KN· model are available in Appendix 8 and a summary of the parameter estimates under the KN· model is available in Appendix 4.

As the parameters are allowed to vary for each patient under the KN· model, different parameter estimates result for each patient, giving different methylation state thresholds. The fitted density estimates for the samples of patients 2, 3 and 4 are available in Appendix 7. Patient 1 was known to have a higher degree of disease than the other patients and their matched normal DNA methylation profile was more tumour-like than benign. For patient 1 the methylation threshold for identifying hypermethylated CpG sites (0.747) was lower than that for the other patients (0.774, 0.766 and 0.814 for patients 2, 3 and 4 respectively) suggesting more hypermethylated CpG sites in patient 1's benign sample than in the other patients' samples. A similar pattern was observed in the methylation thresholds inferred from the patients' tumour samples (see Appendix 7) in that the threshold was lower for patient 1.

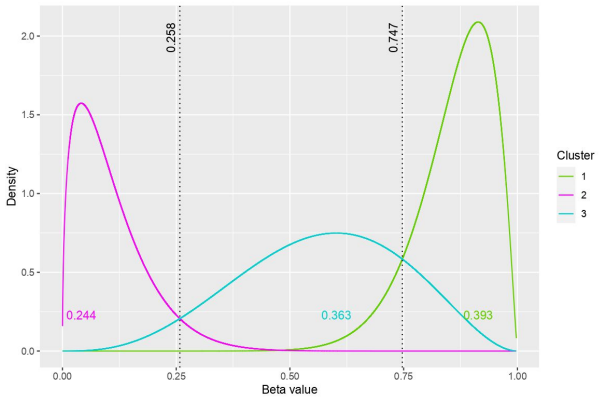

Figure 3: Fitted density estimates under the KN· model for DNA methylation data from the benign sample from patient 1. The methylation state thresholds are indicated by the black dotted lines. The proportion of CpG sites belonging to each cluster is also displayed.

Identifying DMCs in the PCa data

To identify the differentially methylated CpG sites in the PCa data, the K·R model is fitted to the $C \times NR$ dimensional dataset. As there are $M^R = 9$ different combinations of methylation states possible across the benign and tumour samples, the $K = 9$ model is fitted. In Figure 4, the density estimates of the clustering solution are illustrated and Table 3 summarises the parameter estimates under the K·R model. Based on these, clusters 1–4 are deemed to identify the CpG sites that are differentially methylated as their methylation state changes between the benign and tumour samples. The CpG sites that are hypermethylated in

the benign sample and hemimethylated in the tumor sample belong by cluster 1. Clusters 2–4 show a methylation shift from hypo or hemimethylation in the benign prostate tissues to hypermethylation in the tumour tissues. The CpG sites clustered in cluster 5 are in a hemimethylated state in both samples, those in clusters 6 and 7 are hypermethylated in both samples, whereas those in clusters 8 and 9 are hypomethylated. Cluster 4 captures the largest proportion of CpG sites with τ_4 =0.163, whereas cluster 1 identifies the lowest proportion of CpG sites with τ_1 =0.058. The K·R model identifies 44.6% of the CpG sites (i.e., those belonging to clusters 1–4) as differentially methylated between the benign and tumour samples. The kernel density estimates under the K·R model are illustrated in Appendix 9.

The maximum possible clustering uncertainty when clustering the CpG sites into K clusters is $1 - 1/K = 8/9$. Figure 5 illustrates the clustering uncertainties for all the CpG sites under the fitted the K·R model; all CpG sites have clustering uncertainties well below the maximum possible value demonstrating that the CpG sites are clustered with high certainty.

Using the K·R model, 309,889 CpG sites were identified as differentially methylated whereas the conventional approach [30] identified 32,452 DMCs. Of those identified as DMCs by the K·R model, 140 related to genes implicated in prostate cancer carcinogenesis. The DMCs were mapped to DMRs by defining a DMR to occur when ≥ 3 adjacent CpG sites were identified as being differentially methylated. The K·R model identified 60,135 DMRs while the conventional approach [30] identified 1690 DMRs. Ten differentially methylated CpG sites were mapped to the GSTP1 genes, 35 DMCs were mapped to APC genes, 31 were mapped to RASSF1 genes, 22 were mapped to SFRP2 genes and 48 were mapped to RARB genes. For example, hypermethylation of RARB promoter genes is a significant biomarker in diagnosing prostate cancer [43]. Figure 6 displays the methylation levels of the DMCs belonging to the RARB genes for the benign and tumour samples; the median *beta* value is higher in the tumour sample than in the benign sample for all patients. Through non-parametric tests, the *beta* values were shown to be significantly higher in the tumour samples than in the benign samples for the CpG sites related to these genes ($p < 0.05$). Further, Figure 7 shows the empirical cumulative distribution function (ECDF) for DMCs related to the RARB genes for benign and tumour samples. The ECDF illustrates that the DMCs have increased *beta* values in the tumour samples compared to the benign samples.

To further verify the approach, the family of BMMs was also fitted to the publicly available esophageal squamous cell carcinoma dataset. Similar results to those obtained for the prostate cancer dataset were obtained and the results are discussed in detail in Appendix 10.

Software

An R package called `betaclust` has been developed to facilitate the widespread use of the developed family of BMMs. Through the package, each BMM can be fitted separately, or all the BMM models can be fitted simultaneously via a wrapper function. The AIC, BIC and ICL are provided to facilitate selection of the optimal model. Summaries of the fitted models' parameter estimates are also provided. A variety of static plotting options are available for visualization, similar to those shown here. In addition, interactive density plots can be constructed. The package is available at github.

Discussion

DNA methylation is being studied widely for disease diagnosis and treatment [4, 5]. Technology advancements have led to the development of microarrays that can process e.g., 850,000 CpG sites from a DNA sam-

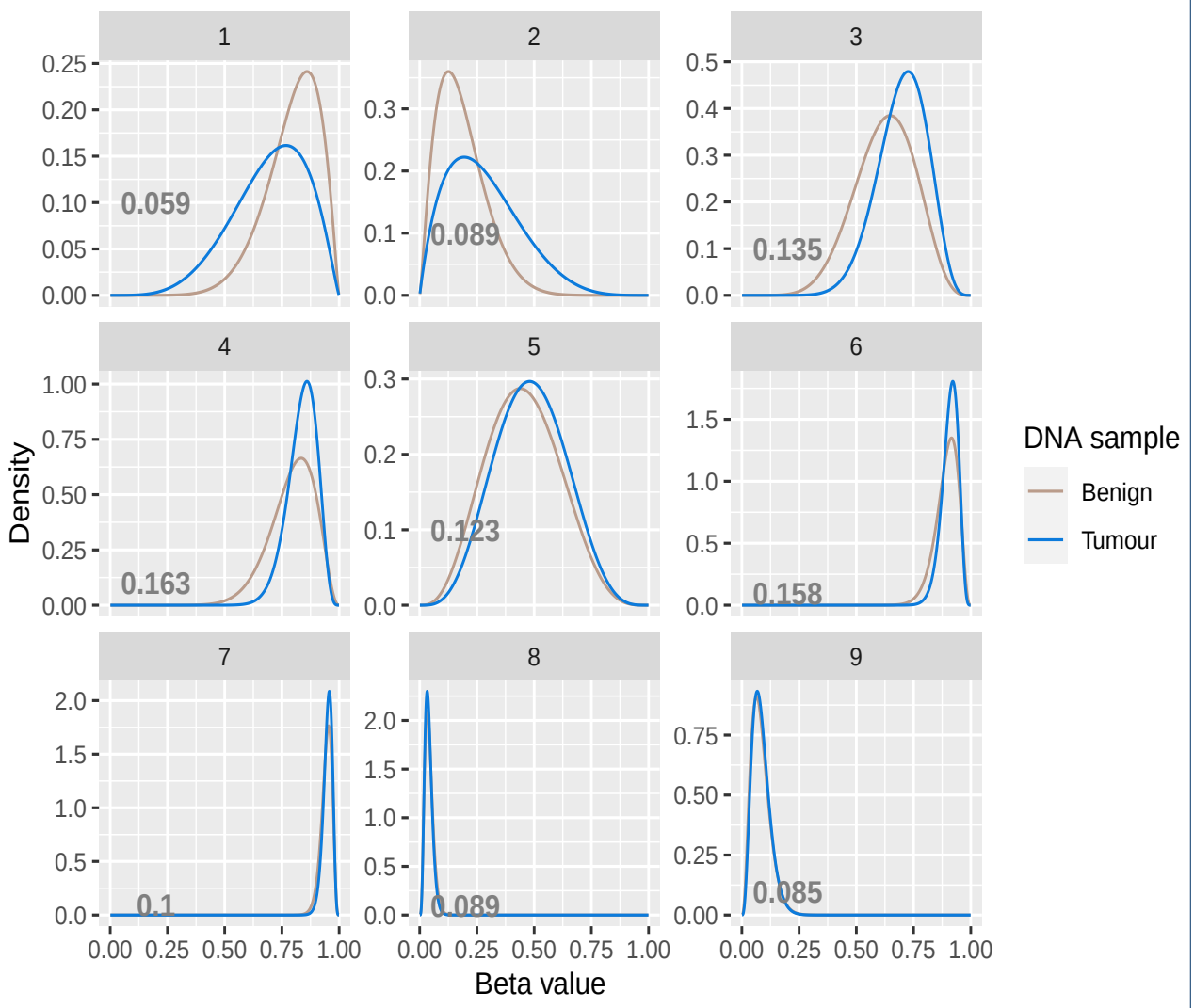

Figure 4: Fitted density estimates under the clustering solution of the K·R model with $K = 9$ for the DNA methylation data from benign and tumour prostate samples. The proportion of CpG sites belonging to each of the 9 clusters is displayed in the relevant panel.

ple but it has been difficult to explore such large arrays with currently available computational power [11]. Analysis has also been limited by a lack of appropriate statistical methods for the bounded nature of the multivariate DNA methlyation data. The methylation states of CpG sites are often of interest and are typically identified using thresholds which are defined in literature based on intuition [12, 13] rather than using a statistical, objective approach. Multiple t-tests or Wilcoxon rank sum tests are often used to compare methylation levels between samples to identify DMCs [23, 24]. The *beta* values are usually transformed to Gaussian distributed values using a logit transformation [12]. The family of BMMs developed here advocates against transforming the data and proposes modelling the data in its innate form.

Table 3: Beta distributions' parameter estimates for the PCa dataset under the K·R model.

(a) Benign samples

Cluster	$\hat{\alpha}$	$\hat{\delta}$	Mean	Std. deviation
1	8.815	2.277	0.795	0.116
2	2.324	10.223	0.185	0.106
3	8.005	4.810	0.625	0.130
4	13.006	3.387	0.793	0.097
5	4.071	4.924	0.453	0.157
6	33.720	4.006	0.894	0.050
7	84.926	5.023	0.944	0.024
8	4.842	112.897	0.041	0.018
9	3.749	41.317	0.083	0.041

(b) Tumour samples

Cluster	$\hat{\alpha}$	$\hat{\delta}$	Mean	Std. deviation
1	5.076	2.231	0.695	0.160
2	1.975	5.040	0.282	0.159
3	12.058	5.170	0.700	0.107
4	27.000	5.249	0.837	0.064
5	4.686	4.990	0.484	0.153
6	56.506	5.734	0.908	0.036
7	111.727	5.978	0.949	0.020
8	5.455	133.043	0.039	0.016
9	4.194	45.197	0.085	0.039

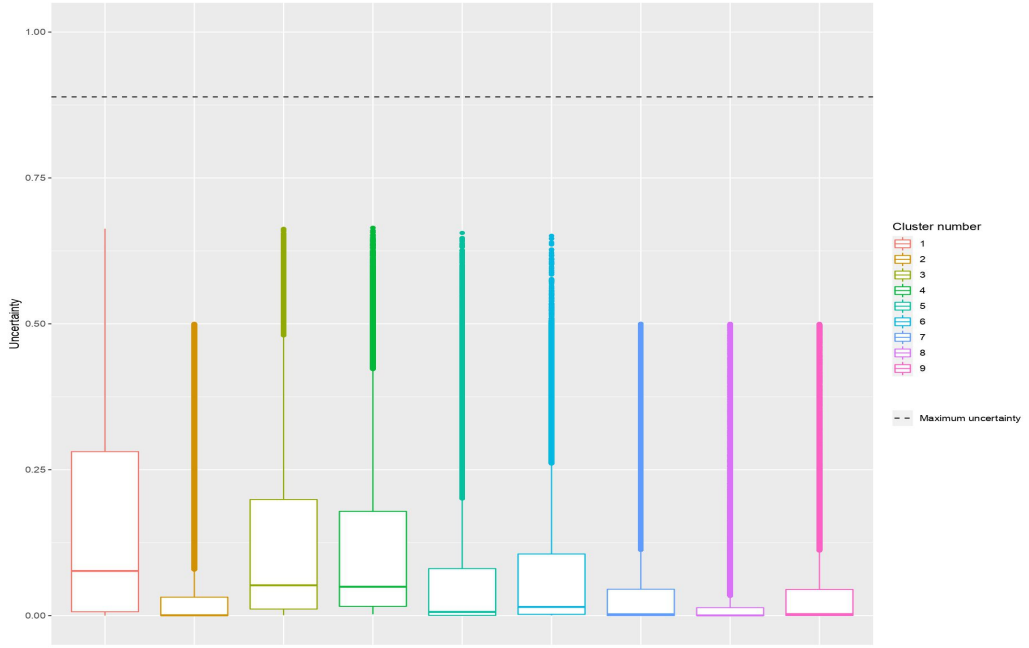

Figure 5: Clustering uncertainties for CpG sites in the PCa data.

We developed a novel family of beta mixture models for the high-dimensional DNA methlyation data to cluster the CpG sites in their innate beta form to (i) objectively identify methylation state thresholds and (ii) identify the DMCs between different samples. The BMMs were evaluated on and applied to simulated datasets, a motivating prostate cancer dataset and a an esophageal squamous cell carcinoma dataset. The BMMs use a model-based clustering approach and are computationally efficient, utilising parallel programming and a digamma function approximation. Run times are reduced from e.g, $\approx$ 65 hours using numerical optimisation at the M-step of the EM algorithm, to 15 minutes when the digamma approximation is used to analyse the PCa data. The objective inference of methylation thresholds using the K·· and KN· models demonstrated that the thresholds of 0.2 and 0.8 defined in literature are not appropriate for every

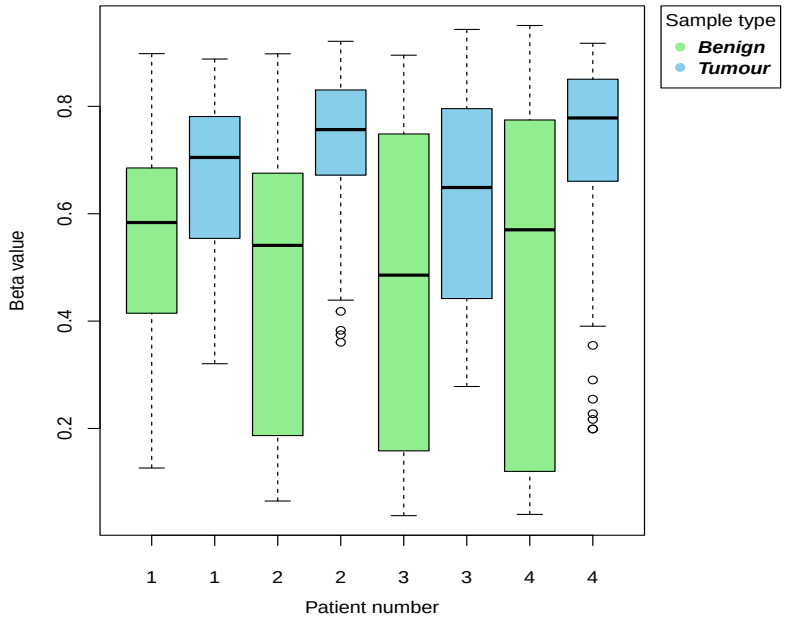

Figure 6: Methlyation levels of the differentially methylated CpG sites related to the RARB genes in the benign and tumour samples.

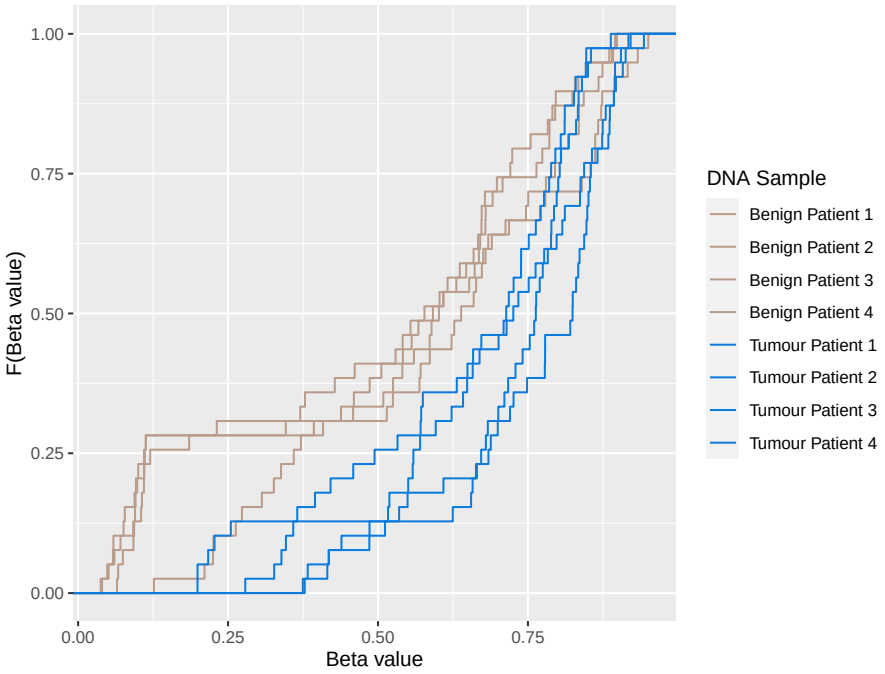

Figure 7: ECDFs for the DMCs related to the RARB genes for all patient samples.

scenario. The thresholds inferred from each patient's data showed variability, reflecting the different stages of disease in the patients. The developed K·R model clusters CpG sites from multiple DNA samples to determine the CpG sites with differential methylation. In the motivating PCa dataset orders of magnitude more CpG sites were identified as being differentially methylated compared to conventional approaches, opening new avenues of research.

A key assumption of the BMMs proposed here is that the methylation states of adjacent CpG sites are conditionally independent given their cluster membership. However, methylation levels of adjacent CpG sites often have high correlation e.g., if a CpG site is hypermethylated, neighbouring CpG sites tend to be hypermethylated [44]. The BMMs could be modified to consider the spatial correlation between adjacent CpG sites. Additionally, the BMMs first locate DMCs, from which DMRs are then determined. The family of BMMs could to be broadened so that the intermediate step of identifying DMCs is removed and the spatial structure in the data is taken into account in order to directly find DMRs of interest.

Finally, the methylation state of a human genome changes over time depending on clinical conditions. Longitudinal methylation data are often collected to study the effect of environmental changes or treatments on disease progression. Such data are vast and current appraoches struggle to handle these extensive data in their innate form. In order to analyze methylation changes over time in multiple patients, similar to [45], the BMMs could be further enhanced to model dependency over time.

Conclusions

Motivated by a prostate cancer application, a family of BMMs was developed for DNA methylation data to infer methylation state thresholds and to identify DMCs by modelling the data in its innate form. More DMCs were objectively identified than by current subjective approaches leading to a deeper insight to the methylation data and opening the opportunity to gain a finer understanding of the epigenetic changes. Further, the BMMs are computationally efficient, achieved by an approximation of the digamma function. The developed family of BMMs can be used to identify differentially methylated CpG sites, thus embedding this epigenetic analysis in a model-based framework.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author's contributions

KM, TBM and ICG conceived, designed, analysed and implemented the BMMs, the `betaclust` package and wrote the manuscript. RS and ASP conceived, acquired and processed the prostate DNA methylation data. RWW and ASP proposed verification of the findings on additional data. All authors provided critical feedback and helped shape the research, analysis and manuscript. All authors read and approved the final manuscript.

Acknowledgements

This publication has emanated from research supported in part by a Grant from Science Foundation Ireland under grant number 18/CRT/6049.

Availability of data and materials

The datasets supporting the conclusions of this article are available in a public repository. Raw data for the prostate cancer dataset are freely available for download on the Gene Expression Omnibus (GEO) repository (GSE119260). Datasets GSM3362390-GSM3362397 were analysed here. Raw data for the ESCC dataset is available in the GEO repository (GSE121930). Datasets GSM3450103-GSM3450106 and GSM3450109-GSM3450112 were used for validation here. All analysis (except stated otherwise) was performed in R 4.1.2. Code for simulations, scripts for creation of figures and original and generated data is available on GitHub.

Author details

[1]School of Mathematics and Statistics, University College Dublin, Belfield, Dublin 4, Ireland. [2]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [3]School of Medicine, UCD Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland.

References

1. Berger, S L and Kouzarides, T and Shiekhattar, R and Shilatifard, A: An operational definition of epigenetics. Genes and Development **23(7)**, 781–783 (2009). doi:10.1101/gad.1787609

2. Moore, L.D., Le, T., Fan, G.: DNA methylation and its basic function. Neuropsychopharmacology **38(1)**, 23–38 (2013). doi:10.1038/npp.2012.112
3. Jin, Z., Liu, Y.: DNA methylation in human diseases. Genes and Diseases **5(1)**, 1–8 (2018). doi.org/10.1016/j.gendis.2018.01.002
4. Das, P., Singal, R.: DNA methylation and cancer. Journal of Clinical Oncology **22(22)**, 4632–4642 (2004). doi:10.1200/JCO.2004.07.151
5. Rakyan, V.K., Down, T.A., Balding, D.J., Beck, S.: Epigenome-wide association studies for common human diseases. Nature Reviews Genetics **12(8)**, 529–541 (2011). doi:10.1038/nrg3000
6. Bird, A.: DNA methylation patterns and epigenetic memory. Genes and Development **16(1)**, 6–21 (2002). doi:10.1101/gad.947102
7. Chen, D.P., Lin, Y.C., Fann, C.S.: Methods for identifying differentially methylated regions for sequence- and array-based data. Briefings in Functional Genomics **15(6)**, 485–490 (2016). doi: 10.1093/bfgp/elw018
8. P de Almeida, B., Apolónio, J.D., Binnie, A., Castelo-Branco, P.: Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers. BMC Cancer **19** (2019). doi:10.1186/s12885-019-5403-0
9. Cho, J.W., Hong, M.H., Ha, S.J., Kim, Y.J., Cho, B.C., Lee, I., Kim, H.R.: Genome-wide identification of differentially methylated promoters and enhancers associated with response to anti-PD-1 therapy in non-small cell lung cancer. Experimental and Molecular Medicine **52(9)**, 1550–1563 (2020). doi:10.1038/s12276-020-00493-8
10. Kim, J.H., Dhanasekaran, S.M., Prensner, J.R., Cao, X., Robinson, D., Kalyana-Sundaram, S., Huang, C., Shankar, S., Jing, X., Iyer, M., Hu, M., Sam, L., Grasso, C., Maher, C.A., Palanisamy, N., Mehra, R., Kominsky, H.D., Siddiqui, J., Yu, J., Qin, Z.S., Chinnaiyan, A.M.: Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. Genome Research **21(7)**, 1028–041 (2011). doi:10.1101/gr.119347.110
11. Pidsley, R., Zotenko, E., Peters, T.J., Lawrence, M.G., Risbridger, G.P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C., Clark, S.J.: Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biology **17(1)**, 208 (2016). doi: 10.1186/s13059-016-1066-1
12. Du, P., Zhang, X., Huang, C.C., Jafari, N., Kibbe, W.A., Hou, L., Lin, S.M.: Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics **11**, 587 (2010). doi:10.1186/1471-2105-11-587
13. Chen, X., Zhang, Q., Chekouo, T.: Filtering high-dimensional methylation marks with extremely small sample size: an application to gastric cancer data. Frontiers in Genetics **12** (2021). 10.3389/fgene.2021.705708
14. Scrucca, L.: A transformation-based approach to Gaussian mixture density estimation for bounded data. Biometrical Journal **61(4)**, 873–888 (2019). doi:10.1002/bimj.201800174
15. Siegmund, K.D., Laird, P.W., Laird-Offringa, I.A.: A comparison of cluster analysis methods using DNA methylation data. Bioinformatics **20(12)**, 1896–1904 (2004). doi:10.1093/bioinformatics/bth176
16. Houseman, E.A., Christensen, B.C., Yeh, R.F., Marsit, C.J., Karagas, M.R., Wrensch, M., Nelson, H.H., Wiemels, J., Zheng, S., Wiencke, J.K., Kelsey, K.T.: Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. BMC Bioinformatics **9**, 365 (2008). doi:10.1186/1471-2105-9-365
17. Koestler, D.C., Christensen, B.C., Marsit, C.J., Kelsey, K.T., Houseman, E.A.: Recursively partitioned mixture model clustering of DNA methylation data using biologically informed correlation structures. Statistical Applications in Genetics and Molecular Biology **12(2)**, 225–240 (2013). doi:10.1515/sagmb-2012-0068
18. P E de Souza, C., Andronescu, M., Masud, T., Kabeer, F., Biele, J., Laks, E., Lai, D., Ye, P., Brimhall, J., Wang, B., Su, E., Hui, T., Cao, Q., Wong, M., Moksa, M., Moore, R.A., Hirst, M., Aparicio, S., Shah, S.P.: Epiclomal: Probabilistic clustering of sparse single-cell DNA methylation data. PLoS Computational Biology **16(9)** (2020). doi:10.1371/journal.pcbi.1008270
19. Ma, Z., Teschendorff, A.E.: A variational Bayes beta mixture model for feature selection in DNA methylation studies. Journal of Bioinformatics and Computational Biology **11(4)** (2013). doi:10.1142/S0219720013500054
20. Zhang, L., Meng, J., Liu, H., Huang, Y.: A nonparametric Bayesian approach for clustering bisulfate-based DNA methylation profiles. BMC Genomics **13(6)**, 20 (2012). doi:10.1186/1471-2164-13-S6-S20
21. Gevaert, O., Tibshirani, R., Plevritis, S.K.: Pancancer analysis of DNA methylation-driven genes using MethylMix. Genome Biology **16(1)**, 17 (2015). doi: 10.1186/s13059-014-0579-8
22. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2022). R Foundation for Statistical Computing. https://www.R-project.org/
23. Wang, D., Yan, L., Hu, Q., Sucheston, L.E., Higgins, M.J., Ambrosone, C.B., Johnson, C.S., Smiraglia, D.J., Liu, S.: IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics **28(5)**, 729–730 (2012). doi:10.1093/bioinformatics/bts013
24. Warden, C.D., Lee, H., Tompkins, J.D., Li, X., Wang, C., Riggs, A.D., Yu, H., Jove, R., Yuan, Y.C.: COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. Nucleic Acids Research [published correction appears in Nucleic Acids Research. 2019 Sep 5;47(15):8335-8336] **41(11)**, 117 (2013). doi:10.1093/nar/gkt242
25. Cuzick, J.: A Wilcoxon-type test for trend. Statistics in Medicine **4(4)**, 543–547 (1985). doi:10.1002/sim.4780040416
26. Chen, Z., Huang, H., Liu, Q.: Detecting differentially methylated loci for multiple treatments based on high-throughput methylation data. BMC Bioinformatics **5**, 142 (2014). doi:10.1186/1471-2105-15-142
27. Li, L.-C., Okino, S.T., Dahiya, R.: DNA methylation in prostate cancer. Biochimica et Biophysica Acta **1704(2)**, 87–102 (2004). doi:10.1016/j.bbcan.2004.06.001
28. Lin, D.C., Wang, M.R., Koeffler, H.P.: Genomic and epigenomic

aberrations in esophageal squamous cell carcinoma and implications for patients. Gastroenterology **154(2)**, 374–389 (2018). doi:10.1053/j.gastro.2017.06.066

29. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **71(3)**, 209–249 (2021). doi:10.3322/caac.21660
30. 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
31. Moran, S., Arribas, C., Esteller, M.: Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics **8(3)**, 389–399 (2016). doi:10.2217/epi.15.114
32. Müller, F., Scherer, M., Assenov, Y., Lutsik, P., Walter, J., Lengauer, T., Bock, C.: RnBeads 2.0: comprehensive analysis of DNA methylation data. Genome Biology **20(1)**, 55 (2019). doi:10.1186/s13059-019-1664-9
33. Chen, Y., Liao, L.D., Wu, Z.Y., Yang, Q., Guo, J.C., He, J.Z., Wang, S.H., Xu, X.E., Wu, J.Y., Pan, F., Lin, D.C., Xu, L.Y., Li, E.M.: Identification of key genes by integrating DNA methylation and next-generation transcriptome sequencing for esophageal squamous cell carcinoma. Aging **12(2)**, 1332–1365 (2020). doi:10.18632/aging.102686
34. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B (Methodological) **39(1)**, 1–38 (1977). doi:10.1111/j.2517-6161.1977.tb01600.x
35. Nocedal, J., Wright, S.J.: Quasi-Newton Methods, pp. 192–221 (1999). doi:10.1007/0-387-22742-3_8
36. Berndt, E., Hall, B., Hall, R., Hausman, J.: Estimation and Inference in Non-linear Structural Models vol. 3(4), pp. 653–665 (1974)
37. 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
38. Akaike, H.: Information Theory and an Extension of the Maximum Likelihood Principle, pp. 199–213 (1998). doi:10.1007/978-1-4612-1694-0_15
39. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics **6**, 461–464 (1978). doi: 10.1214/aos/1176344136
40. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence **22(7)**, 719–725 (2000). 10.1109/34.865189
41. Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E.: mclust 5: clustering, classification and density estimation using gaussian finite mixture models. The R Journal **8(1)**, 289–317 (2016). PMCID: PMC5096736
42. Hubert, L., Arabie, P.: Comparing Partitions. Journal of Classification **2**, 193–218 (1985). doi:10.1007/BF01908075
43. Ameri, A., Alidoosti, A., Hosseini, S.Y., Parvin, M., Emranpour, M.H., Taslimi, F., Salehi, E., Fadavip, P.: Prognostic value of promoter hypermethylation of retinoic acid receptor beta (RARB) and CDKN2 (p16/MTS1) in prostate cancer. Chinese Journal of Cancer Research **23(4)**, 306–311 (2011). doi:10.1007/s11670-011-0306-x
44. Hodges, E., Smith, A.D., Kendall, J., Xuan, Z., Ravi, K., Rooks, M., Zhang, M.Q., Ye, K., Bhattacharjee, A., Brizuela, L., McCombie, W.R., Wigler, M., Hannon, G.J., Hicks, J.B.: High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. Genome Research **19(9)**, 1593–1605 (2009). doi:10.1101/gr.095190.109
45. Nyamundanda, G., Gormley, I.C., Brennan, L.: A Dynamic Probabilistic Principal Components Model for the Analysis of Longitudinal Metabolomics Data. Journal of the Royal Statistical Society, Series C (Applied Statistics) **63(5)**, 763–782 (2014). doi:10.1111/rssc.12060

Additional Files

Appendix to 'betaclust: a family of mixture models for beta valued DNA methylation data' by Majumdar et al.

Appendix to 'betaclust: a family of mixture models for beta valued DNA methylation' data by Majumdar et al.

Appendix 1

K$\cdot\cdot$ Model The complete data log-likelihood for this model is,

$$\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}, \mathbf{Z}|\mathbf{X}) = \sum_{c=1}^{C}\sum_{k=1}^{K} z_{ck}\{\log \tau_k + \sum_{n=1}^{N}\sum_{r=1}^{1} \log[\text{Beta}(x_{cnr}; \alpha_{k\cdot\cdot}, \delta_{k\cdot\cdot})]\}.$$

In the E-step of the EM algorithm the $\hat{z}_{ck}$ is calculated given the current parameter estimates. In the M-step the expected complete data log-likelihood function to be optimized is,

$$\begin{aligned}\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}|\mathbf{X}, \hat{\mathbf{Z}}) = &\sum_{c=1}^{C}\sum_{k=1}^{K} \hat{z}_{ck}\{\log \tau_k + \\ &\sum_{n=1}^{N}\sum_{r=1}^{1}[(\alpha_{k\cdot\cdot} - 1)\log x_{cnr} + (\delta_{k\cdot\cdot} - 1)\log(1 - x_{cnr}) - \log \mathrm{B}(\alpha_{k\cdot\cdot}, \delta_{k\cdot\cdot})]\}.\end{aligned} \tag{1}$$

Differentiating (1) w.r.t $\alpha_{k\cdot\cdot}$ yields,

$$\frac{\partial \ell_C}{\partial \alpha_{k\cdot\cdot}} = \sum_{c=1}^{C} \hat{z}_{ck}\{\log x_{cnr} - [\psi(\alpha_{k\cdot\cdot}) - \psi(\alpha_{k\cdot\cdot} + \delta_{k\cdot\cdot})]\} \tag{2}$$

where ψ is the digamma function.

Similarly, the derivative of $\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}|\mathbf{X}, \hat{\mathbf{Z}})$ w.r.t $\delta_{k\cdot\cdot}$ is,

$$\frac{\partial \ell_C}{\partial \delta_{k\cdot\cdot}} = \sum_{c=1}^{C} \hat{z}_{ck}\{\log(1 - x_{cnr}) - [\psi(\delta_{k\cdot\cdot}) - \psi(\alpha_{k\cdot\cdot} + \delta_{k\cdot\cdot})]\}. \tag{3}$$

The lower bound value of the digamma function ($\psi(y) > \log(y - 1/2)$) is used in (2) and (3) to get closed-form solutions at the M-step of the EM algorithm,

$$\frac{\partial \ell_C}{\partial \alpha_{k\cdot\cdot}} \approx \sum_{c=1}^{C} \hat{z}_{ck} \sum_{n=1}^{N}\sum_{r=1}^{1} \left[\log x_{cnr} - \log \frac{\alpha_{k\cdot\cdot} - 1/2}{\alpha_{k\cdot\cdot} + \delta_{k\cdot\cdot} - 1/2}\right] \tag{4}$$

and

$$\frac{\partial \ell_C}{\partial \delta_{k\cdot\cdot}} \approx \sum_{c=1}^{C} \hat{z}_{ck} \sum_{n=1}^{N}\sum_{r=1}^{1} \left[\log(1 - x_{cnr}) - \log \frac{\delta_{k\cdot\cdot} - 1/2}{\alpha_{k\cdot\cdot} + \delta_{k\cdot\cdot} - 1/2}\right]. \tag{5}$$

Equating (4) and (5) to zero, we get the approximate estimates of $\alpha_{k\cdot\cdot}$ and $\delta_{k\cdot\cdot}$ as,

$$\alpha_{k\cdot\cdot} = 0.5 + \frac{0.5\exp(-y_2)}{\{[\exp(-y_2) - 1][\exp(-y_1) - 1]\} - 1}$$

and

$$\delta_{k\cdot\cdot} = \frac{0.5\exp(-y_2)[\exp(-y_1) - 1]}{\{[\exp(-y_2) - 1][\exp(-y_1) - 1]\} - 1},$$

where $y_1 = (\sum_{c=1}^{C} z_{ck}\log x_{cnr})/(N\sum_{c=1}^{C} z_{ck})$ and $y_2 = (\sum_{c=1}^{C} z_{ck}\log(1 - x_{cnr}))/(N\sum_{c=1}^{C} z_{ck})$.

Appendix 2

KN· Model The complete data log-likelihood for this model is,

$$\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}, \mathbf{Z}|\mathbf{X}) = \sum_{c=1}^{C}\sum_{k=1}^{K} z_{ck}\{\log \tau_k + \sum_{n=1}^{N}\sum_{r=1}^{1} \log[\text{Beta}(x_{cnr}; \alpha_{kn\cdot}, \delta_{kn\cdot})]\}.$$

In the E-step of the EM algorithm the $\hat{z}_{ck}$ is calculated given the current parameter estimates. In the M-step the expected complete data log-likelihood function to be optimized is,

$$\begin{aligned}\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}|\mathbf{X}, \hat{\mathbf{Z}}) = &\sum_{c=1}^{C}\sum_{k=1}^{K} \hat{z}_{ck}\{\log \tau_k + \\ &\sum_{n=1}^{N}\sum_{r=1}^{1}[(\alpha_{kn\cdot} - 1)\log x_{cnr} + (\delta_{kn\cdot} - 1)\log(1 - x_{cnr}) - \log \mathrm{B}(\alpha_{kn\cdot}, \delta_{kn\cdot})]\}.\end{aligned} \tag{6}$$

Differentiating (6) w.r.t $\alpha_{kn\cdot}$ yields,

$$\frac{\partial \ell_C}{\partial \alpha_{kn\cdot}} = \sum_{c=1}^{C} \hat{z}_{ck}\{\log x_{cnr} - [\psi(\alpha_{kn\cdot}) - \psi(\alpha_{kn\cdot} + \delta_{kn\cdot})]\} \tag{7}$$

where ψ is the digamma function.

Similarly, the derivative of $\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}|\mathbf{X}, \hat{\mathbf{Z}})$ w.r.t $\delta_{kn\cdot}$ is,

$$\frac{\partial \ell_C}{\partial \delta_{kn\cdot}} = \sum_{c=1}^{C} \hat{z}_{ck}\{\log(1 - x_{cnr}) - [\psi(\delta_{kn\cdot}) - \psi(\alpha_{kn\cdot} + \delta_{kn\cdot})]\}. \tag{8}$$

The lower bound value of the digamma function ($\psi(y) > \log(y - 1/2)$) is used in (7) and (8) to get closed-form solutions at the M-step of the EM algorithm,

$$\frac{\partial \ell_C}{\partial \alpha_{kn\cdot}} \approx \sum_{c=1}^{C} \hat{z}_{ck} \sum_{n=1}^{N}\sum_{r=1}^{1} \left[\log x_{cnr} - \log \frac{\alpha_{kn\cdot} - 1/2}{\alpha_{kn\cdot} + \delta_{kn\cdot} - 1/2}\right] \tag{9}$$

and

$$\frac{\partial \ell_C}{\partial \delta_{kn\cdot}} \approx \sum_{c=1}^{C} \hat{z}_{ck} \sum_{n=1}^{N}\sum_{r=1}^{1} \left[\log(1 - x_{cnr}) - \log \frac{\delta_{kn\cdot} - 1/2}{\alpha_{kn\cdot} + \delta_{kn\cdot} - 1/2}\right]. \tag{10}$$

Equating (9) and (10) to zero, we get the approximate estimates of $\alpha_{kn\cdot}$ and $\delta_{kn\cdot}$ as,

$$\alpha_{kn\cdot} = 0.5 + \frac{0.5\exp(-y_2)}{\{[\exp(-y_2) - 1][\exp(-y_1) - 1]\} - 1}$$

and

$$\delta_{kn\cdot} = \frac{0.5\exp(-y_2)[\exp(-y_1) - 1]}{\{[\exp(-y_2) - 1][\exp(-y_1) - 1]\} - 1},$$

where $y_1 = (\sum_{c=1}^{C} z_{ck}\log x_{cnr})/(\sum_{c=1}^{C} z_{ck})$ and $y_2 = (\sum_{c=1}^{C} z_{ck}\log(1 - x_{cnr}))/(\sum_{c=1}^{C} z_{ck})$.

Appendix 3

K·R Model The complete data log-likelihood for this model is,

$$\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}, \mathbf{Z}|\mathbf{X}) = \sum_{c=1}^{C}\sum_{k=1}^{K} z_{ck}\{\log \tau_k + \sum_{n=1}^{N}\sum_{r=1}^{R} \log[\text{Beta}(x_{cnr}; \alpha_{k\cdot r}, \delta_{k\cdot r})]\}.$$

In the E-step of the EM algorithm the $\hat{z}_{ck}$ is calculated given the current parameter estimates. In the M-step the expected complete data log-likelihood function to be optimized is,

$$\begin{aligned}\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}|\mathbf{X}, \hat{\mathbf{Z}}) = &\sum_{c=1}^{C}\sum_{k=1}^{K} \hat{z}_{ck}\{\log \tau_k + \\ &\sum_{n=1}^{N}\sum_{r=1}^{R}[(\alpha_{k\cdot r}-1)\log x_{cnr} + (\delta_{k\cdot r}-1)\log(1-x_{cnr}) - \log \mathrm{B}(\alpha_{k\cdot r}, \delta_{k\cdot r})]\}.\end{aligned} \tag{11}$$

Differentiating (11) w.r.t $\alpha_{k\cdot r}$ yields,

$$\frac{\partial \ell_C}{\partial \alpha_{k\cdot r}} = \sum_{c=1}^{C} \hat{z}_{ck}\{\log x_{cnr} - [\psi(\alpha_{k\cdot r}) - \psi(\alpha_{k\cdot r} + \delta_{k\cdot r})]\} \tag{12}$$

where ψ is the digamma function.

Similarly, the derivative of $\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}|\mathbf{X}, \hat{\mathbf{Z}})$ w.r.t $\delta_{k\cdot r}$ is,

$$\frac{\partial \ell_C}{\partial \delta_{k\cdot r}} = \sum_{c=1}^{C} \hat{z}_{ck}\{\log(1-x_{cnr}) - [\psi(\delta_{k\cdot r}) - \psi(\alpha_{k\cdot r} + \delta_{k\cdot r})]\}. \tag{13}$$

The lower bound value of the digamma function ($\psi(y) > \log(y - 1/2)$) is used in (12) and (13) to get closed-form solutions at the M-step of the EM algorithm,

$$\frac{\partial \ell_C}{\partial \alpha_{k\cdot r}} \approx \sum_{c=1}^{C} \hat{z}_{ck} \sum_{n=1}^{N}\sum_{r=1}^{1}\left[\log x_{cnr} - \log \frac{\alpha_{k\cdot r} - 1/2}{\alpha_{k\cdot r} + \delta_{k\cdot r} - 1/2}\right] \tag{14}$$

and

$$\frac{\partial \ell_C}{\partial \delta_{k\cdot r}} \approx \sum_{c=1}^{C} \hat{z}_{ck} \sum_{n=1}^{N}\sum_{r=1}^{1}\left[\log(1-x_{cnr}) - \log \frac{\delta_{k\cdot r} - 1/2}{\alpha_{k\cdot r} + \delta_{k\cdot r} - 1/2}\right]. \tag{15}$$

Equating (14) and (15) to zero, we get the approximate estimates of α_{knr} and δ_{knr} as,

$$\alpha_{k\cdot r} = 0.5 + \frac{0.5\exp(-y_2)}{\{[\exp(-y_2)-1][\exp(-y_1)-1]\} - 1}$$

and

$$\delta_{k\cdot r} = \frac{0.5\exp(-y_2)[\exp(-y_1)-1]}{\{[\exp(-y_2)-1][\exp(-y_1)-1]\} - 1},$$

where $y_1 = (\sum_{c=1}^{C} z_{ck}\log x_{cnr})/(N\sum_{c=1}^{C} z_{ck})$ and $y_2 = (\sum_{c=1}^{C} z_{ck}\log(1-x_{cnr}))/(N\sum_{c=1}^{C} z_{ck})$.

Appendix 4

Table 1: Beta distributions' parameter estimates for sample A in a simulated dataset under the K.. model

Clusters	$\hat{\alpha}$	$\hat{\delta}$	Mean	Std. deviation
1	2.13	21.33	0.091	0.058
2	21.02	2.11	0.909	0.058
3	4.14	3.10	0.571	0.172

Table 2: Beta distributions' parameter estimates for sample A in a simulated dataset under the K·R model

(a) Sample A

Clusters	$\hat{\alpha}$	$\hat{\delta}$	Mean	Std. deviation
1	21.06	2.10	0.909	0.058
2	2.12	2.19	0.09	0.058
3	4.27	3.03	0.585	0.171
4	2.12	2.74	0.089	0.057
5	3.98	3.25	0.551	0.173
6	3.96	2.92	0.576	0.176
7	2.49	2.13	0.910	0.58
8	4.18	3.11	0.574	0.172
9	2.13	21.35	0.091	0.058

(b) Sample B

Clusters	$\hat{\alpha}$	$\hat{\delta}$	Mean	Std. deviation
1	2.08	20.85	0.091	0.059
2	21.14	2.13	0.909	0.059
3	2.13	21.33	0.091	0.058
4	4.04	3.03	0.572	0.174
5	2.10	21.35	0.090	0.058
6	20.53	2.06	0.909	0.059
7	2.82	1.31	0.684	0.205
8	4.2	3.2	0.570	0.172
9	2.10	20.91	0.092	0.059

Table 3: Beta distributions' parameter estimates for benign sample in the PCa dataset under the KN· model.

(a) Patient 1

Clusters	$\hat{\alpha}$	$\hat{\delta}$	Mean	Std. deviation
1	13.774	2.205	0.862	0.084
2	1.491	12.454	0.107	0.080
3	3.970	2.965	0.572	0.176

(c) Patient 2

Clusters	$\hat{\alpha}$	$\hat{\delta}$	Mean	Std. deviation
1	21.434	2.871	0.882	0.064
2	2.166	18.166	0.107	0.067
3	4.111	2.980	0.580	0.174

(b) Patient 3

Clusters	$\hat{\alpha}$	$\hat{\delta}$	Mean	Std. deviation
1	20.158	2.624	0.885	0.065
2	2.183	28.896	.070	0.045
3	3.618	3.023	0.545	0.180

(d) Patient 4

Clusters	$\hat{\alpha}$	$\hat{\delta}$	Mean	Std. deviation
1	26.825	2.644	0.910	0.052
2	2.462	30.940	0.074	0.045
3	3.338	2.237	0.599	0.191

Appendix 5

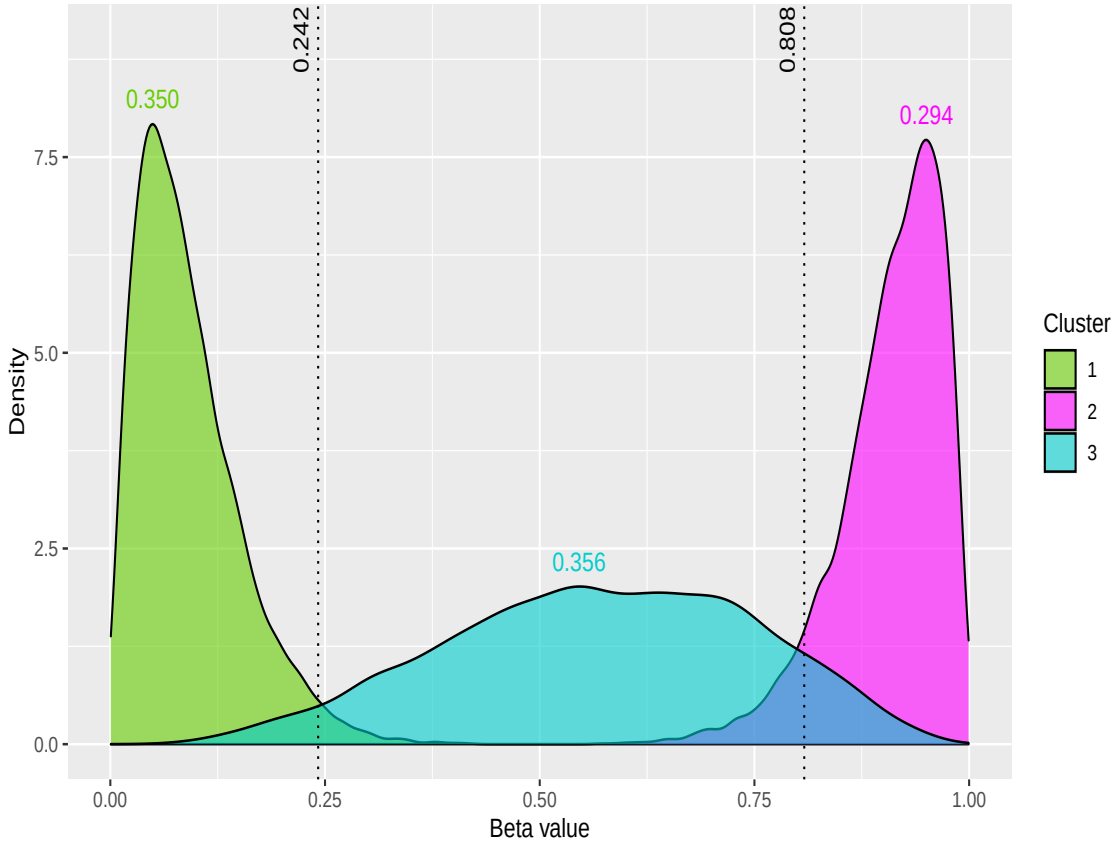

Figure 1: Kernel density estimates under the K·· model fitted to data from sample A in the simulated dataset. The thresholds are 0.242 and 0.808. The proportion of CpG sites belonging to each cluster is displayed.

Appendix 6

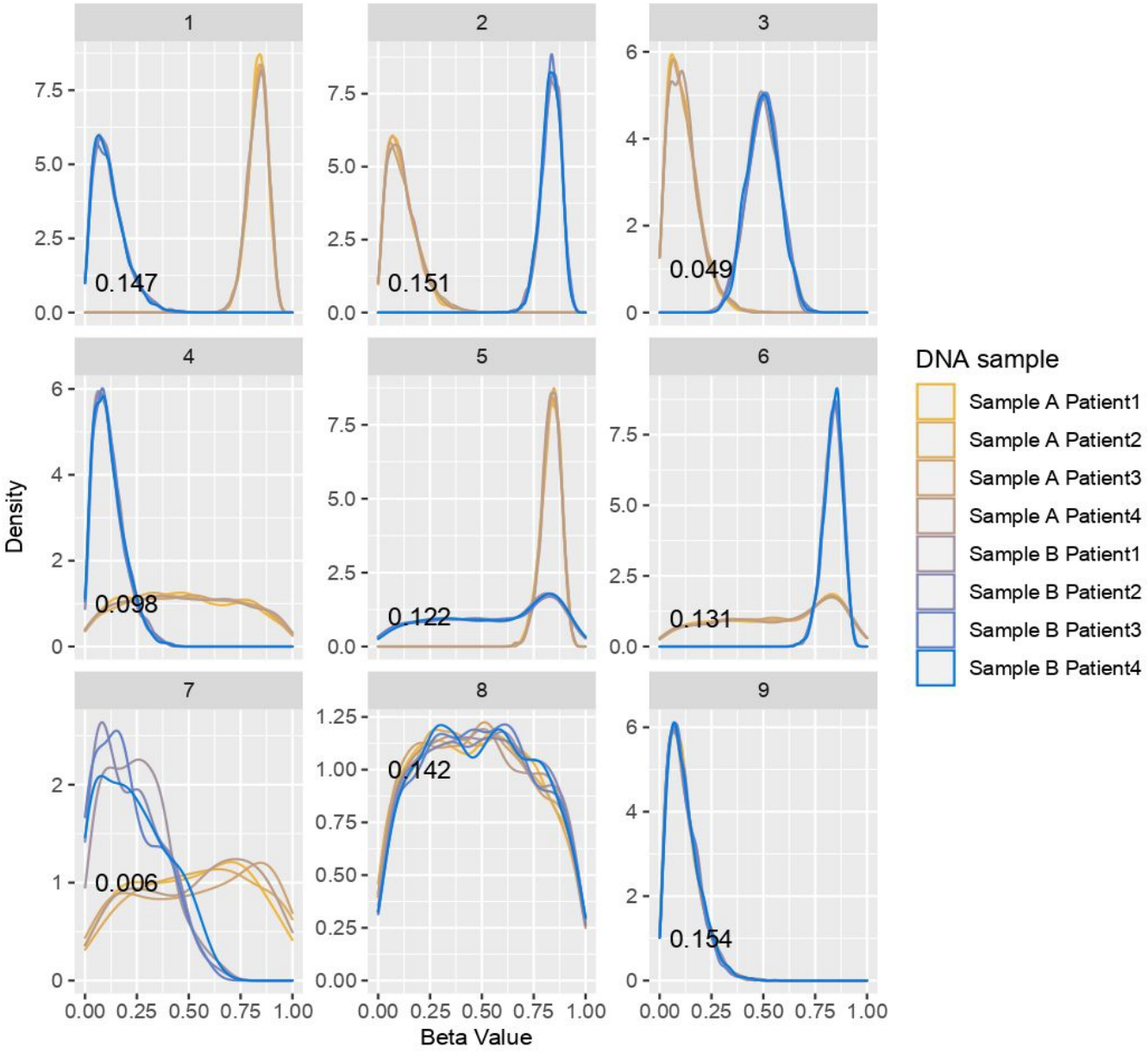

Figure 2: Kernel density estimates under the clustering solution of the K·R model fitted to sample A and sample B from a simulated dataset. The proportion of CpG sites belonging to each of the 9 clusters is displayed in the relevant panel.

Appendix 7

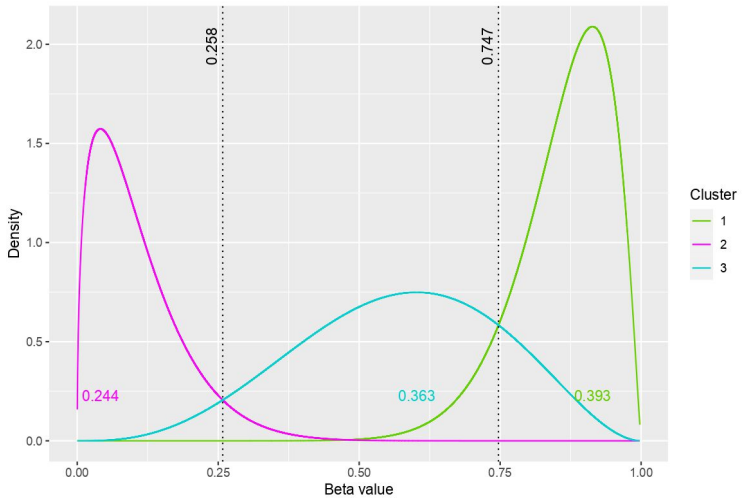

Figure 3: Fitted density estimates under the clustering solution of the KN· model fitted to the benign sample collected from patient 1 in the prostate cancer dataset. The threshold points are illustrated in the graph as 0.258 and 0.747.

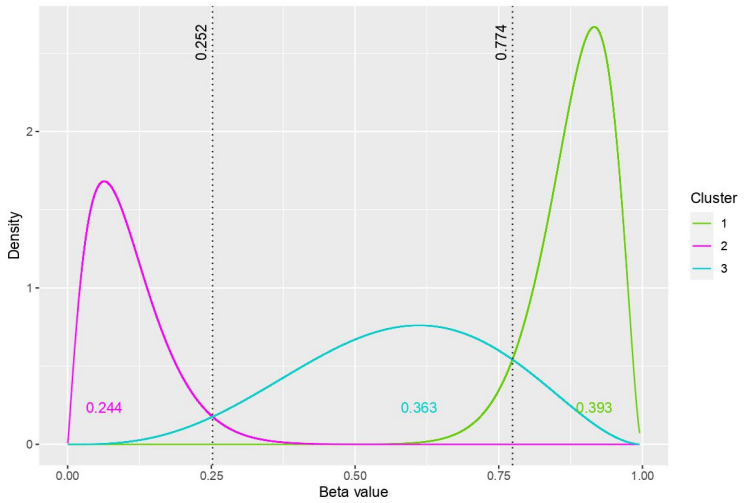

Figure 4: Fitted density estimates under the clustering solution of the KN· model fitted to the benign sample collected from patient 2 in the prostate cancer dataset. The threshold points are illustrated in the graph as 0.252 and 0.774.

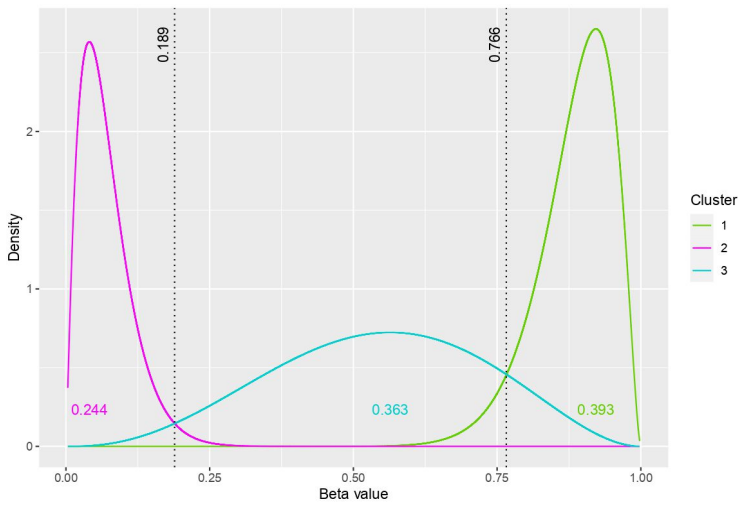

Figure 5: Fitted density estimates under the clustering solution of the KN· model fitted to the benign sample collected from patient 3 in the prostate cancer dataset. The threshold points are illustrated in the graph as 0.189 and 0.766.

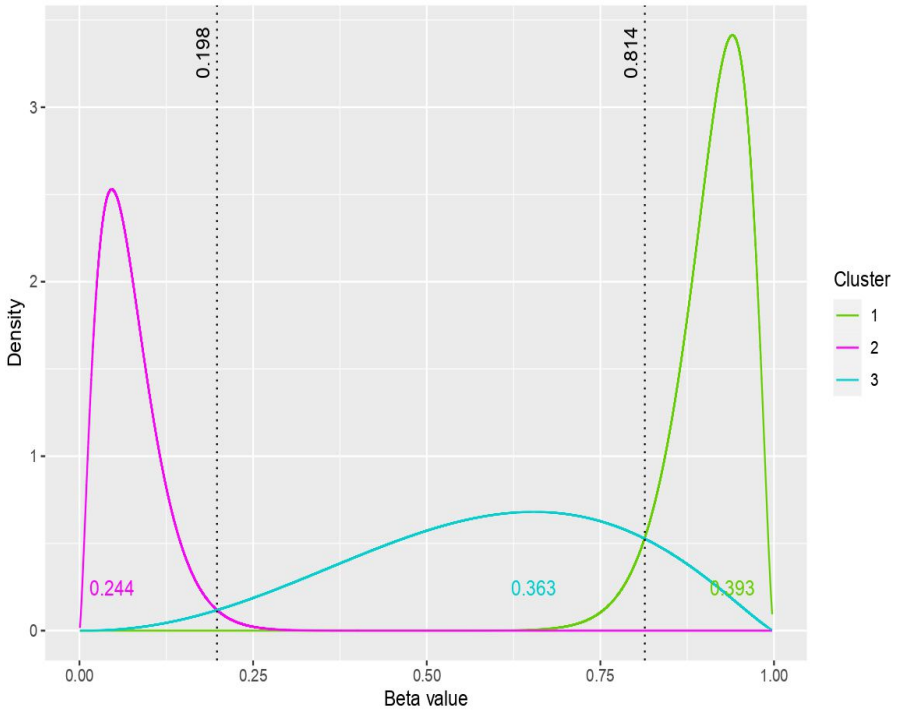

Figure 6: Fitted density estimates under the clustering solution of the KN· model fitted to the benign sample collected from patient 4 in the prostate cancer dataset. The threshold points are illustrated in the graph as 0.198 and 0.814.

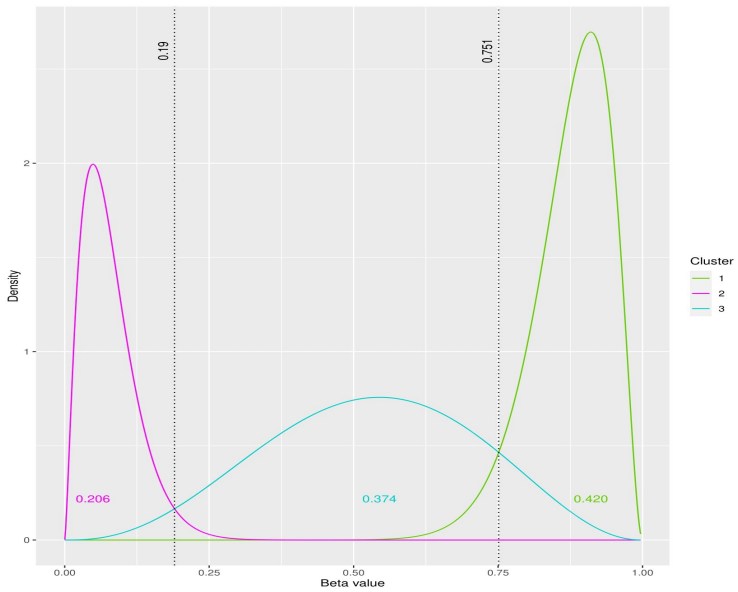

Figure 7: Fitted density estimates under the clustering solution of the KN· model fitted to the tumour sample collected from patient 1 in the prostate cancer dataset. The threshold points are illustrated in the graph as 0.19 and 0.751.

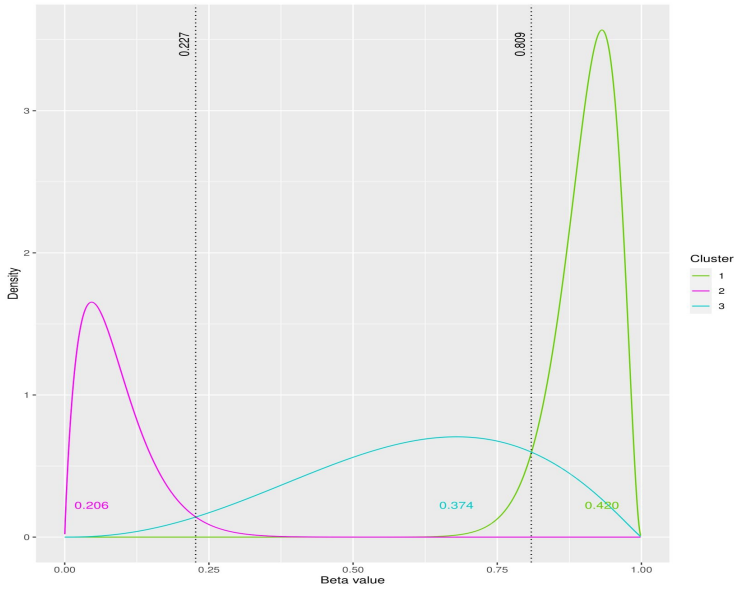

Figure 8: Fitted density estimates under the clustering solution of the KN· model fitted to the tumour sample collected from patient 2 in the prostate cancer dataset. The threshold points are illustrated in the graph as 0.227 and 0.809.

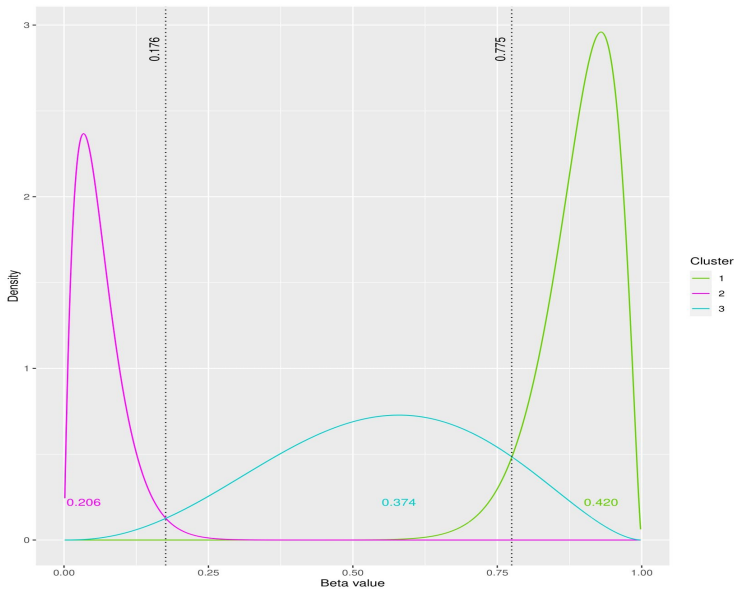

Figure 9: Fitted density estimates under the clustering solution of the KN· model fitted to the tumour sample collected from patient 3 in the prostate cancer dataset. The threshold points are illustrated in the graph as 0.176 and 0.775.

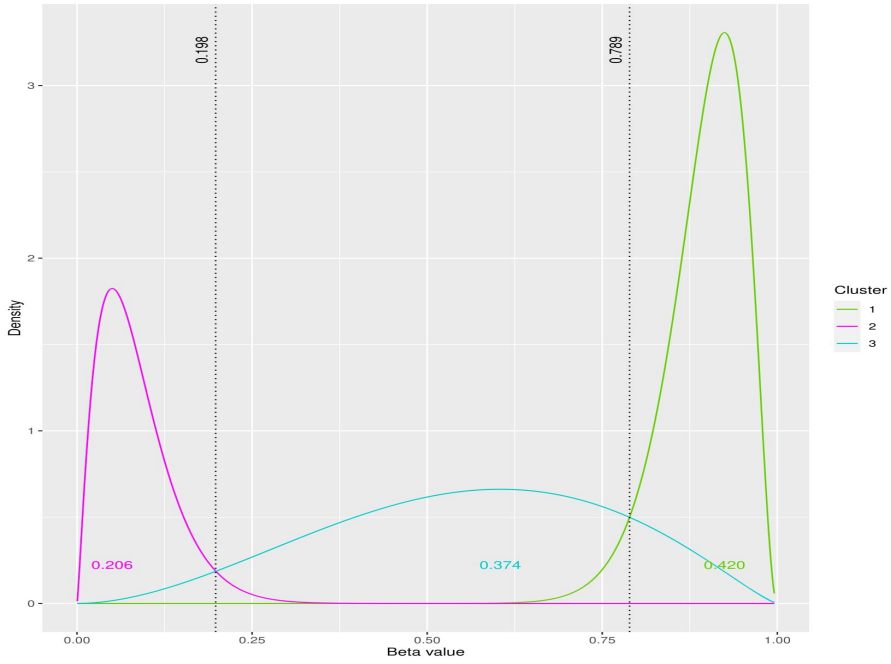

Figure 10: Fitted density estimates under the clustering solution of the KN· model fitted to the tumour sample collected from patient 4 in the prostate cancer dataset. The threshold points are illustrated in the graph as 0.198 and 0.789.

Appendix 8

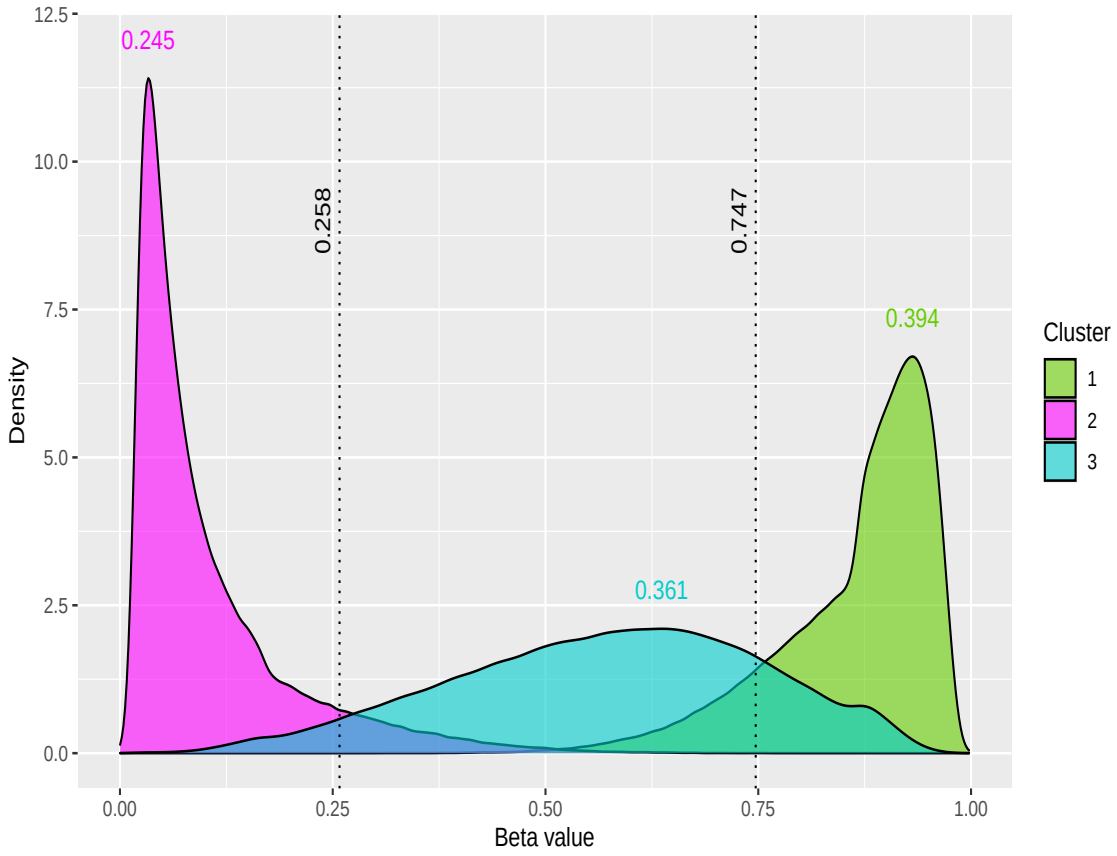

Figure 11: Kernel density estimates under the clustering solution of the KN· model fitted to DNA methylation data from the benign sample collected from patient 1 in the prostate cancer dataset. The thresholds are illustrated along with the proportion of CpG sites belonging to each cluster.

Appendix 9

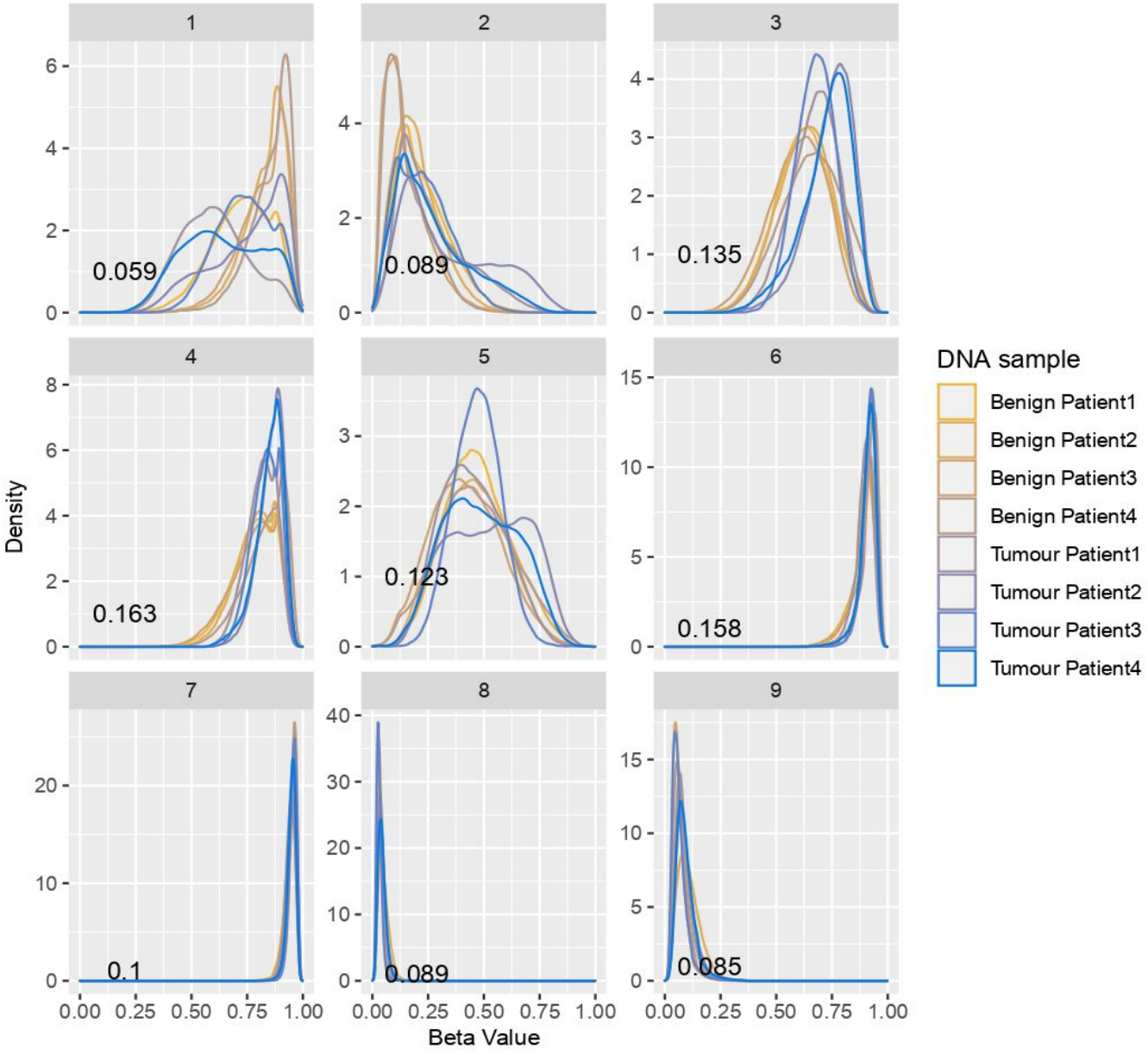

Figure 12: Kernel density estimates under the clustering solution of the K·R model fitted to the DNA methylation data from benign and tumour prostate cancer samples. The proportion of CpG sites belonging to each of the 9 clusters is displayed in the relevant panel.

Appendix 10

Esophageal squamous cell carcinoma data

Esophageal squamous cell carcinoma (ESCC) is a subtype of esophageal cancer characterized by aberrant DNA methylation. A study was conducted to investigate abnormal genes in ESCC, and DNA samples were collected from 15 patients' benign and tumour tissues [1]. Paired samples from 4 randomly selected patients were considered. The ESCC dataset contains *beta* values for each of the $C = 474,869$ CpG sites from $R = 2$ DNA samples collected from $N = 4$ patients.

Estimating methylation state thresholds

The methylation states of each CpG site and the threshold points between these states are to be inferred. The K·· and KN· models are used to achieve this objective by clustering the CpG sites from the benign sample into 3 methylation states, allowing objective inference of the thresholds. The KN· model was selected as the optimal model by BIC and the fitted density estimates of the clustering solution for patient 1 are displayed in Figure 13. As the KN· model estimates different parameters for each patient, different pairs of thresholds are calculated for each patient. The methylation state thresholds for patient 1 are inferred to be 0.326 and 0.789 under the KN· model. A summary of the parameter estimates under the KN· model is presented in Table 4.

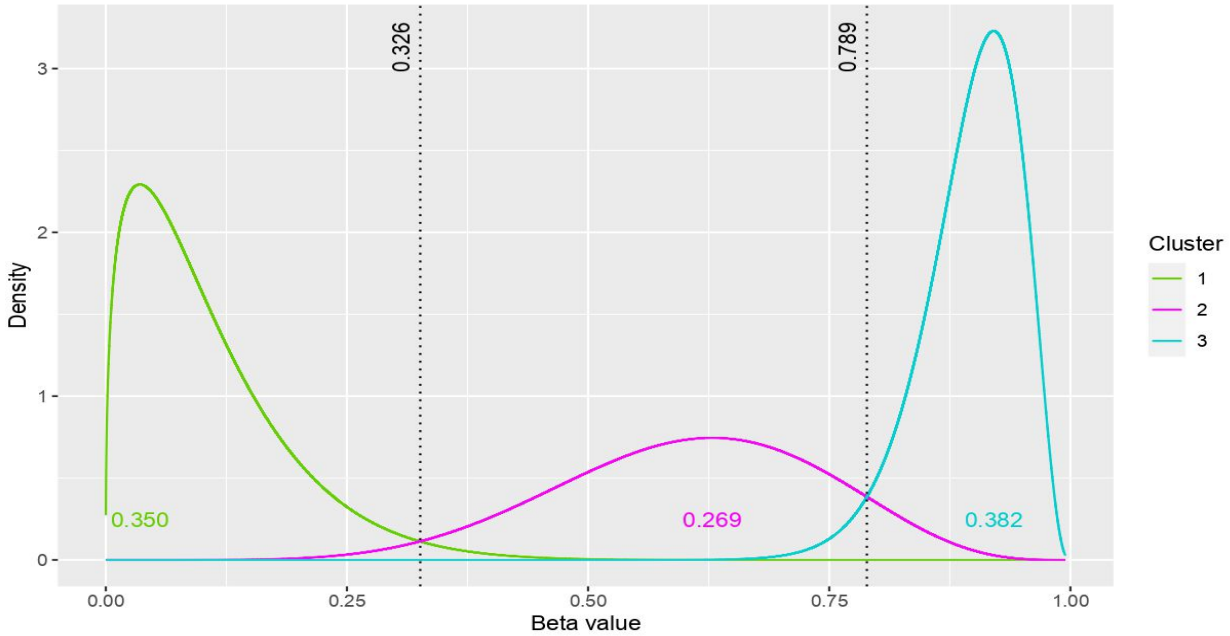

Figure 13: Fitted density estimates under the clustering solution of the KN· model for DNA methylation data from the benign sample collected from patient 1 in the ESCC dataset. The methylation state thresholds are illustrated by the black dotted lines along with the proportion of CpG sites belonging to each cluster.

Table 4: Beta distributions' parameter estimates for benign samples in the ESCC dataset under the KN· model.

(a) Patient 1

Clusters	$\hat{\alpha}$	$\hat{\delta}$	Mean	Std. deviation
1	1.399	11.885	0.105	0.081
2	7.179	4.651	0.607	0.136
3	31.438	3.629	0.897	0.051

(b) Patient 3

Clusters	$\hat{\alpha}$	$\hat{\delta}$	Mean	Std. deviation
1	1.438	11.782	0.109	0.083
2	7.294	4.924	0.597	0.135
3	28.286	3.868	0.880	0.056

(c) Patient 2

Clusters	$\hat{\alpha}$	$\hat{\delta}$	Mean	Std. deviation
1	1.355	12.059	0.101	0.079
2	6.421	4.215	0.604	0.143
3	31.344	3.620	0.896	0.051

(d) Patient 4

Clusters	$\hat{\alpha}$	$\hat{\delta}$	Mean	Std. deviation
1	1.408	10.347	0.120	0.091
2	7.022	4.879	0.590	0.137
3	29.343	3.973	0.881	0.055

Identifying DMCs in the ESCC data

The CpG sites that are differentially methylated between the benign and tumour samples are identified by fitting the K·R model with $K = 9$. The fitted densities are shown in Figure 14. The CpG sites for which the methylation state varies between benign and tumour samples are clustered in clusters 1–5. Clusters 6 and 7 capture CpG sites that are hypermethylated in both samples; clusters 8 and 9 identify CpG sites that are hypomethylated in both samples. A summary of the parameter estimates under the K·R model is presented in Table 5.

Table 5: Beta distributions' parameter estimates for the ESCC dataset under the K·R model.

(a) Benign samples

Clusters	$\hat{\alpha}$	$\hat{\delta}$	Mean	Std. deviation
1	2.101	10.274	0.170	0.103
2	12.159	17.882	0.405	0.088
3	74.220	10.663	0.874	0.036
4	30.532	19.358	0.612	0.068
5	41.069	12.186	0.771	0.057
6	114.976	4.681	0.961	0.018
7	173.782	4.725	0.922	0.019
8	9.044	96.170	0.086	0.027
9	2.195	87.67	0.025	0.016

(b) Tumour samples

Clusters	$\hat{\alpha}$	$\hat{\delta}$	Mean	Std. deviation
1	1.209	2.984	0.288	0.199
2	3.198	3.168	0.502	0.184
3	16.696	3.217	0.838	0.080
4	5.673	3.176	0.641	0.153
5	8.624	2.937	0.746	0.123
6	29.278	1.814	0.942	0.041
7	85.562	7.729	0.917	0.028
8	8.059	83.097	0.088	0.030
9	2.038	76.552	0.026	0.018

The maximum possible uncertainty when clustering the CpG sites into K clusters is $1 - 1/K = 8/9$. Figure 15 illustrates the clustering uncertainties for all CpG sites and demonstrates that there is low uncertainty in the CpG site's cluster memberships under the K·R model.

The K·R model identified DMCs related to genes implicated in esophageal squamous cell carcinogenesis. For example, the expression of the PRSS27 gene has been observed

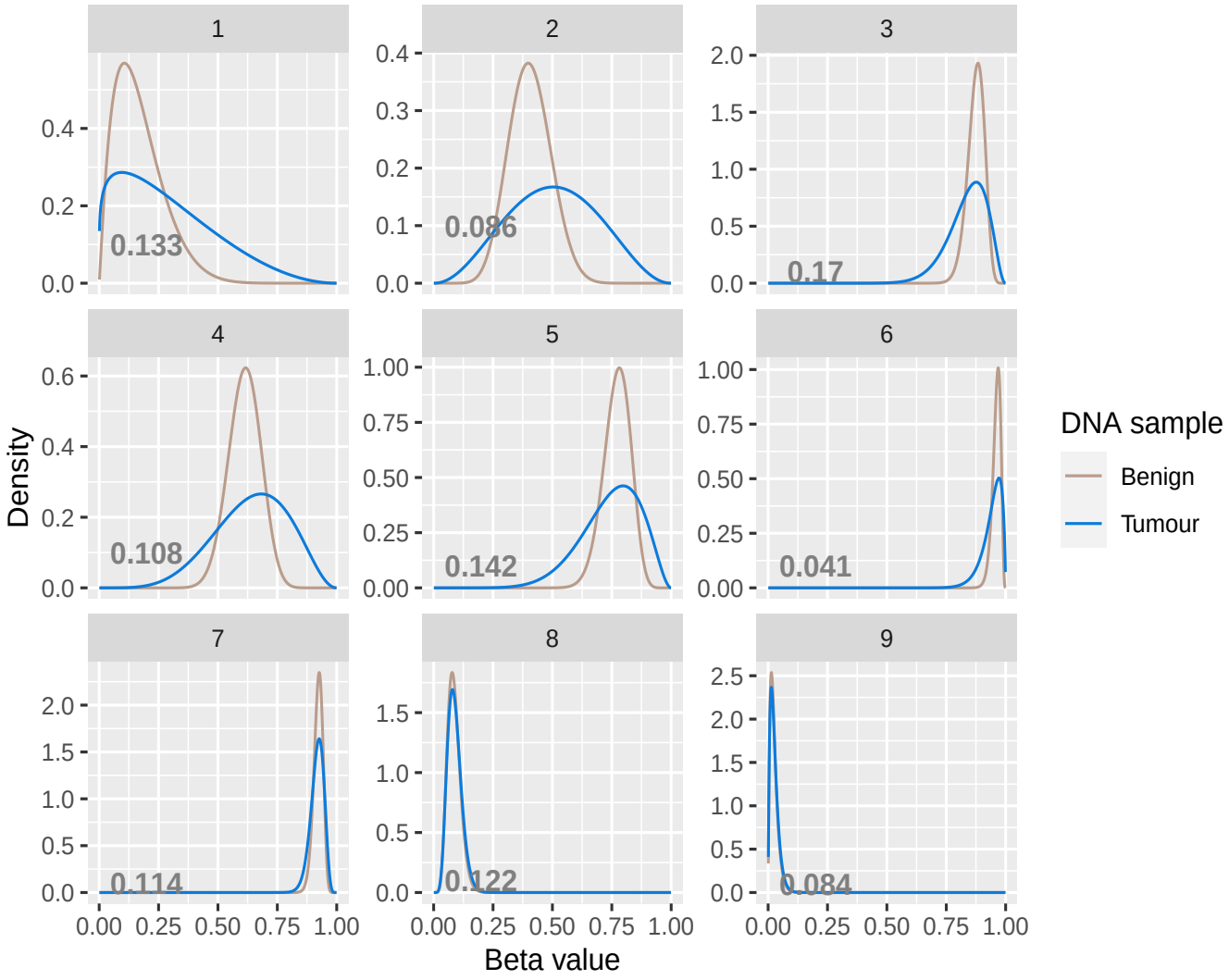

Figure 14: Fitted density estimates under the clustering solution of the K·R model for DNA methylation data from benign and tumour ESCC samples. The proportion of CpG sites belonging to each of the 9 clusters is displayed in the relevant panel.

to be downregulated in ESCC [2]. Figure 16 shows the empirical cumulative distribution function (ECDF) for the DMCs related to the PRSS27 genes for all patient samples. The ECDF illustrates that the identified DMCs have increased *beta* values in the tumour samples compared to the benign samples.

The expression of the GPX3 gene has been shown to be downregulated in ESCC when compared with normal esophageal mucosa [3]. The promoter methylation results in the silencing of the GPX3 genes in ESCC. The ECDF plot in Figure 17 illustrates hypermethylation of the identified DMCs related to the GPX3 gene in the tumour samples.

The CRABP2 gene is expressed in ESCC tissues and normal esophageal squamous ep-

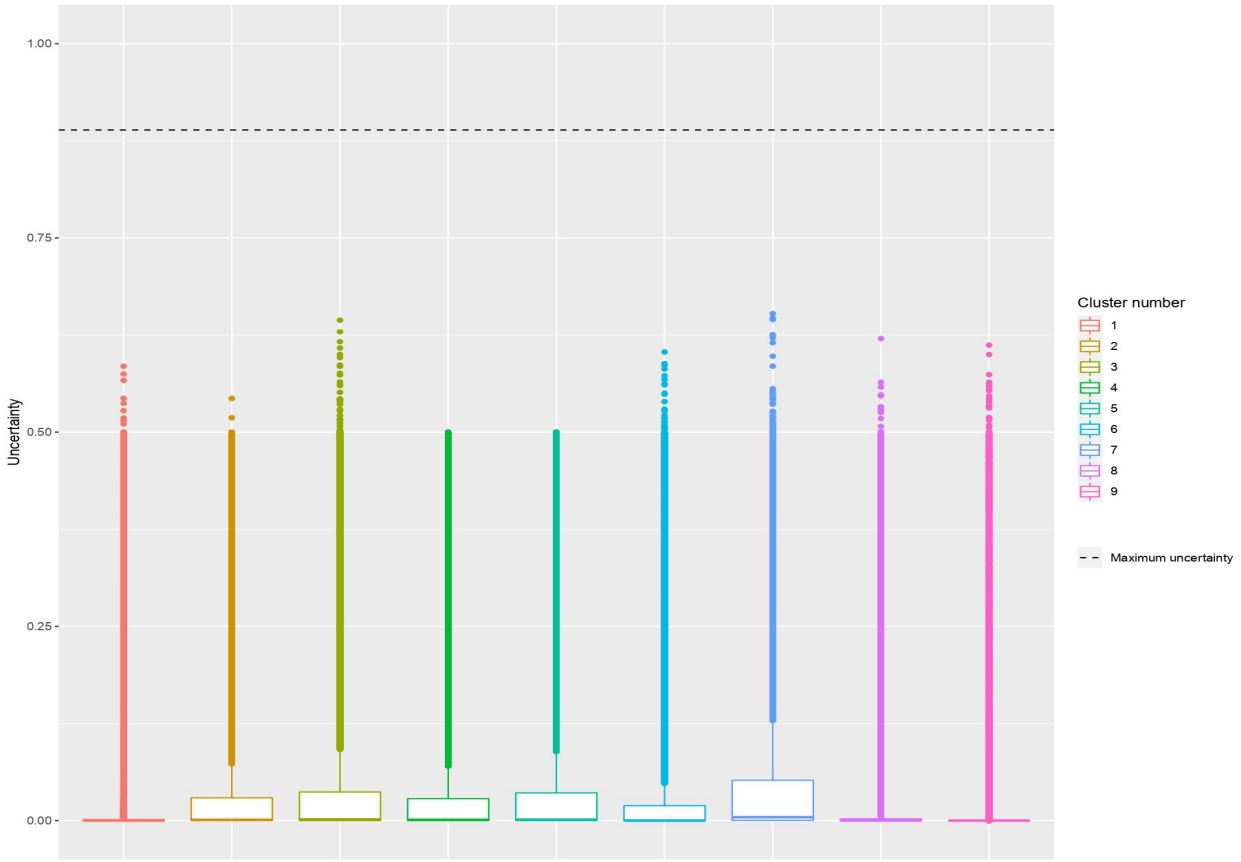

Figure 15: Clustering uncertainties for CpG sites in each clustering group under the clustering solution of the K·R model for the ESCC dataset.

ithelium tissues, but gene expression was shown to be significantly higher in normal tissues than in malignant tissues [4]. Hypermethylation of the DMCs related to the CRABP2 gene is observed in the ECDF plot in Figure 18.

The MFAP2 gene expression, on the other hand, has been observed to be upregulated in ESCC tissues in comparison to normal esophageal squamous epithelium tissues [5]. In Figure 19, the ECDF plot shifts to the left for the tumour samples for most patients, suggesting hypomethylation of the CpG sites related to the MFAP2 gene.

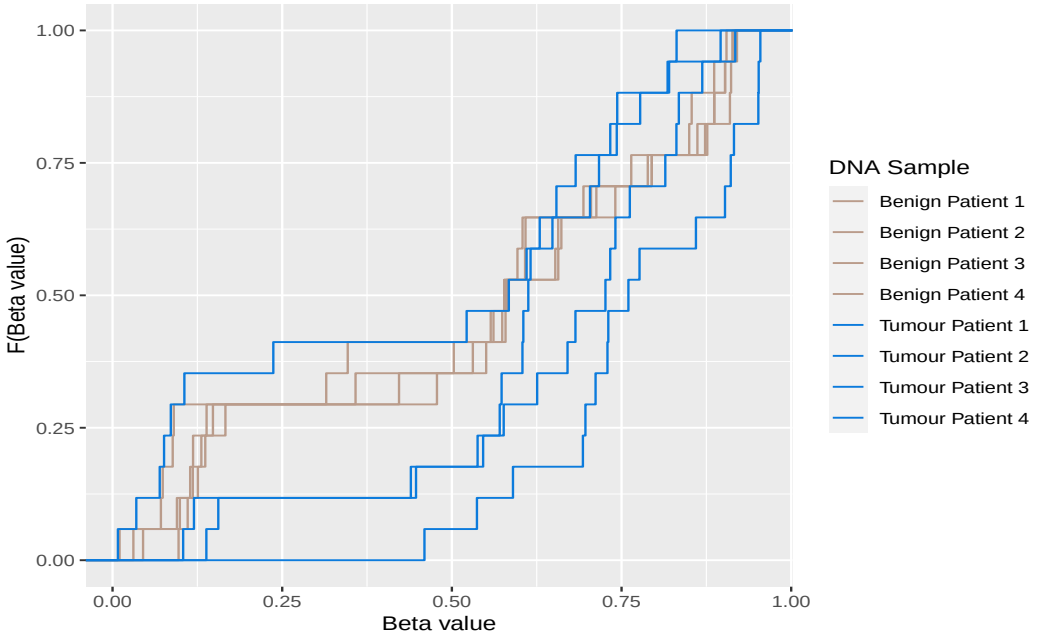

Figure 16: ECDFs for all the CpG sites identified as differentially methylated and related to the PRSS27 genes for all patient samples.

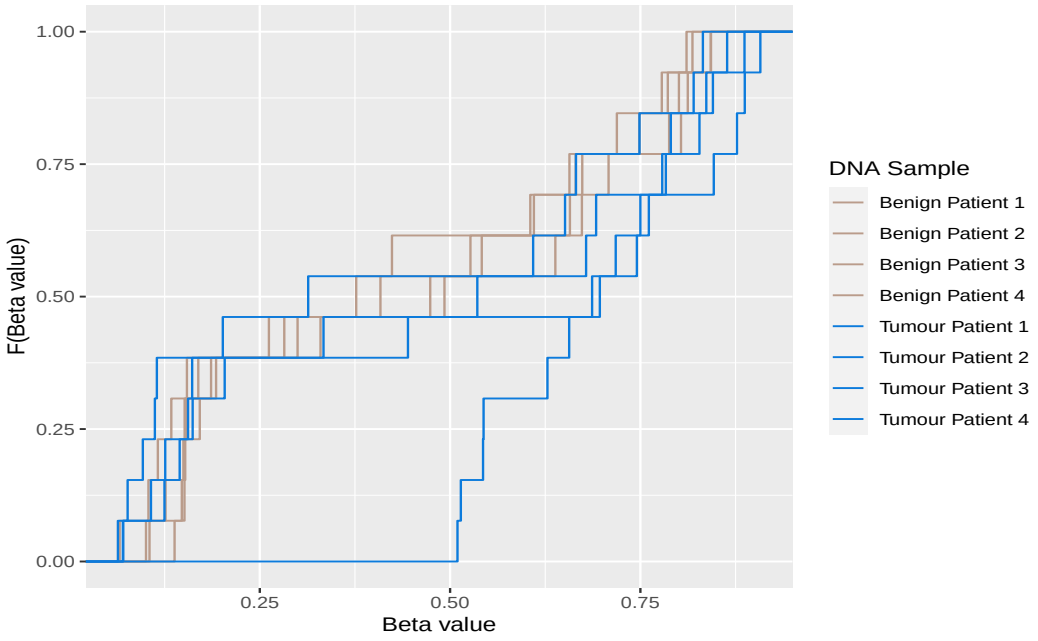

Figure 17: ECDFs for all the CpG sites identified as differentially methylated and related to the GPX3 genes for all patient samples.

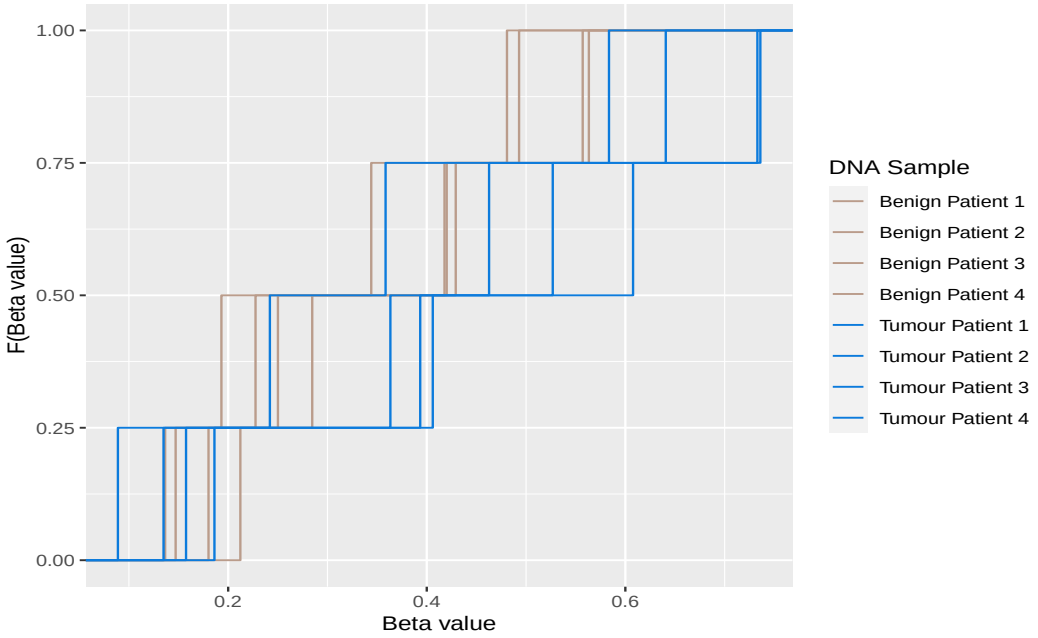

Figure 18: ECDFs for all the CpG sites identified as differentially methylated and related to the CRABP2 genes for all patient samples.

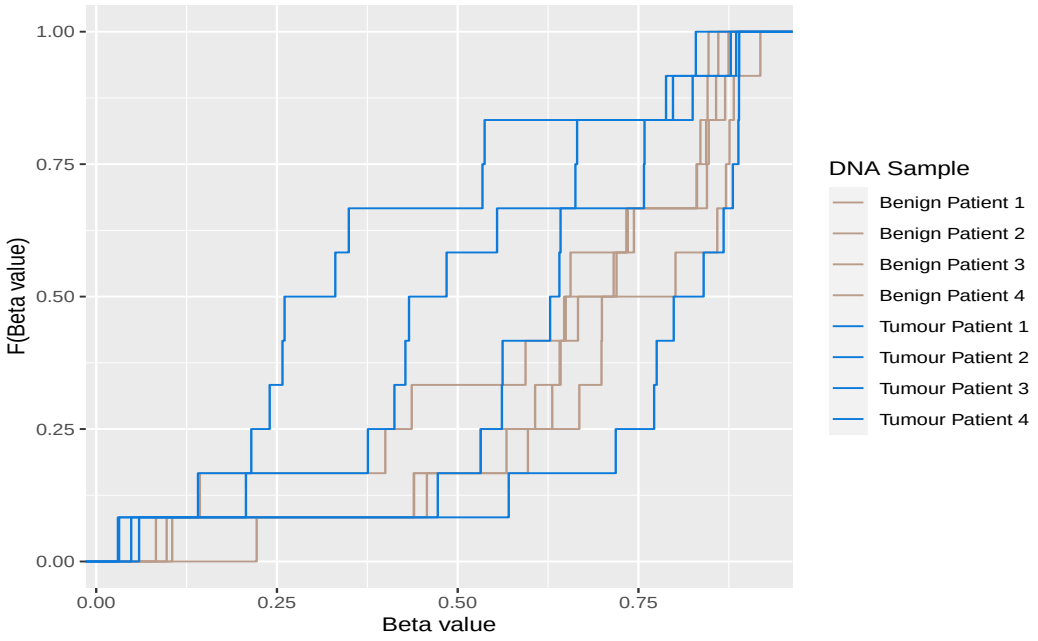

Figure 19: ECDFs for all the CpG sites identified as differentially methylated and related to the MFAP2 genes for all patient samples.

References

[1] Chen, Y., Liao, L. D., Wu, Z. Y., Yang, Q., Guo, J. C., He, J. Z., Wang, S. H., Xu, X. E., Wu, J. Y., Pan, F., Lin, D. C., Xu, L. Y., Li, E. M.: Identification of key genes by integrating DNA methylation and next-generation transcriptome sequencing for esophageal squamous cell carcinoma. Aging **12(2)**, 1332–1365 (2020). doi:10.18632/aging.102686.

[2] Kataoka, A., Yamada, K., Hagiwara, T., Terayama, M., Sugimoto, T., Nohara, K., Igari, T., Yokoi, C., Kawamura, Y.I.: Expression status of Serine Protease 27: a prognostic marker for esophageal squamous cell carcinoma treated with preoperative chemotherapy/chemoradiotherapy. Annals of Surgical Oncology **28(9)**, 5373-5381 (2021). doi:10.1245/s10434-020-09550-y.

[3] Lin, Y., Zhang, Y., Chen, Y., Liu, Z.: Promoter methylation and clinical significance of GPX3 in esophageal squamous cell carcinoma. Pathology - Research and Practice **215(11)**, 152676 (2019). doi: 10.1016/j.prp.2019.152676.

[4] Li, M., Li, C., Lu, P., Wang, B., Gao, Y., Liu, W., Shi, Y., Ma, Y.: Expression and function analysis of CRABP2 and FABP5, and their ratio in esophageal squamous cell carcinoma. Open Medicine (Warsaw, Poland) **16(1)**. 1444-1458 (2021). doi: 10.1515/med-2021-0350.

[5] Shen, Z., Chen, M., Luo, F., Xu, H., Zhang, P., Lin, J., Kang, M.: Identification of key genes and pathways associated with paclitaxel resistance in esophageal squamous cell carcinoma based on bioinformatics analysis. Frontiers in Genetics. **12**, 671639 (2021). doi: 10.3389/fgene.2021.671639.