Insight

MBCbigP: model based clustering for high dimensional data

Cathal Mullin, Triona Ryan,
Adrian O'Hagan and Claire Gormley

School of Mathematics and Statistics & the Insight Centre for Data Analytics

University College Dublin.

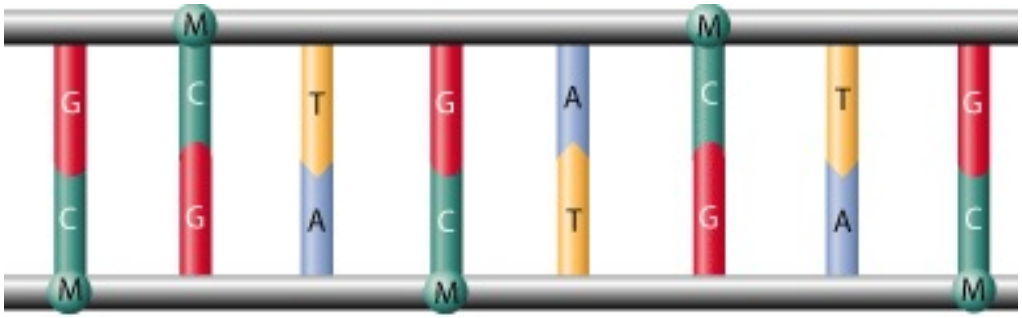
What is MBCbigP?

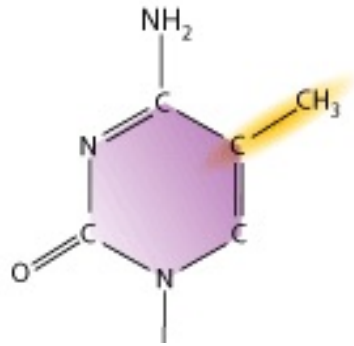
- Model based clustering when the number of variables p is large.
- 'Large' means 100s to 100,000s of variables.
- Clustering when dimension reduction approaches are not practically interpretable.
- Focus is on clustering solution, *and* on interpretability at the variable level.

Epigenetic DNA methylation data

- Epigenetics: heritable changes in phenotype, caused by a mechanism other than mutation to the DNA sequence.

- DNA methylation is the best understood epigenetic mechanism.





DNA methylation is the addition of a methyl group (*M*) to the DNA base cytosine (C).

Epigenetic DNA methylation data

- DNA methylation affects gene transcription and is influenced by environment.
- Current *Illumina* technology reads methylation level at around 0.5 million CpG sites.
- Interest lies in clustering samples based on *all* their methylation data, and on understanding *at a CpG level* the differences between clusters.
- Cluster $N = 597$ breast cancer tumour samples, where tumour subtypes (Basal and non-Basal) are known.

The MBCbigP model

What is MBCbigP?

- MBCbigP employs a finite mixture of probability distributions.
- Approximate the probability distribution within a cluster by the product of conditional distributions.
- Fit MBCbigP in a sequential manner via the EM algorithm.

The MBCbigP model

- Denote observation i's data as the p-vector y_i.
- Unknown number of clusters is G.
- Proportion belonging to cluster g denoted τ_g.
- Assume within a cluster data are $\sim MVN_p(\mu_g, \Sigma_g)$.
- Divide the p variables into $Q = 3$ *segments* (say) denoted $1, 2, 3$, each of length $r = p/Q$:

$$y_i^T = (y_{i1}, y_{i2}, y_{i3})^T$$

The MBCbigP model

$$
\begin{aligned}
f(y_i) &= \sum_{g=1}^{G} \tau_g f[(y_{i1}, y_{i2}, y_{i3})|\theta_g] \\
&= \sum_{g=1}^{G} \tau_g f(y_{i1}|\theta_g) f(y_{i2}|y_{i1}, \theta_g) f(y_{i3}|y_{i2}, y_{i1}, \theta_g)
\end{aligned}
$$

- Given $y_i \sim MVN$, and the properties of the Gaussian distribution, each segment $y_{i\cdot}|\cdots \sim MVN$.

- Introduce the G-vector z_i: $z_{ig} = 1$ if $i \in g$, and 0 otherwise.

$$
\mathcal{L}_C = \prod_{i=1}^{N} \prod_{g=1}^{G} \left[\tau_g f(y_{i1}|\theta_g) f(y_{i2}|y_{i1}, \theta_g) f(y_{i3}|y_{i2}, y_{i1}, \theta_g)\right]^{z_{ig}}
$$

Partitioned Gaussians

- Partition y_i into $Q = 2$ segments, of length r:

$$y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \qquad \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}$$

- Then:

$$y_{i1} \sim MVN_r(\mu_1, \Sigma_{11})$$

and

$$y_{i2}|y_{i1} \sim MVN_r(\mu_{2|1}, \Lambda_{22}^{-1})$$

where

$$\mu_{2|1} = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(y_{i1} - \mu_1)$$

$$\Lambda_{22}^{-1} = (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}$$

The MBCbigP model

- Approximate by conditioning on the previous segment of data only:

$$f(y_{i3}|y_{i2}, y_{i1}, \theta_g) \quad \approx \quad f(y_{i3}|y_{i2}, \theta_g^{3|2})$$

- Run EM on pairs of dependent segments, in stages:

$$\text{Stage 1:} \quad \mathcal{L}_C^1 = \prod_{i=1}^{N} \prod_{g=1}^{G} \left[\tau_g f(y_{i1}|\theta_g^1)\right]^{z_{ig}}$$

$$\text{Stage 2:} \quad \mathcal{L}_C^{12} = \prod_{i=1}^{N} \prod_{g=1}^{G} \left[\tau_g f(y_{i1}|\theta_g^1) f(y_{i2}|y_{i1}, \theta_g^{2|1})\right]^{z_{ig}}$$

$$\text{Stage 3:} \quad \mathcal{L}_C^{23} = \prod_{i=1}^{N} \prod_{g=1}^{G} \left[\tau_g f(y_{i2}|\theta_g^2) f(y_{i3}|y_{i2}, \theta_g^{3|2})\right]^{z_{ig}}$$

$$\vdots \qquad\qquad \vdots$$

The MBCbigP model

- Fit a mixture of (unconstrained) Gaussians at each stage, via an EM algorithm.
- Only requires storage of 2 data segments and their associated r dimensional parameters in memory at a time.
- The Z matrix produced on convergence at each stage is used as a starting value for the next stage.
- Segment specific parameter estimates are passed on to the next stage also.
- Thus the use of conditional probability distributions and the informed starting values, achieves a level of dependence across the large number of dimensions.

Model selection

- Evaluate the conditional Gaussian densities on convergence at each stage.
- Leads to approximation of the likelihood for the p dimensional data:

$$\mathcal{L} \approx \prod_{i=1}^{N} \sum_{g=1}^{G} \tau_g f(y_{i1}|\theta_g^1) f(y_{i2}|y_{i1}, \theta_g^{2|1}) f(y_{i3}|y_{i2}, \theta_g^{3|2})$$

- Use this approximation in standard model selection criteria eg. BIC.

Illustrative examples

Toy example: bank note data

- Six measurements made on 100 genuine and 100 counterfeit Swiss bank notes.
- Divide the $p = 6$ variables into $Q = 3$ segments.
- Only considered VVV models.

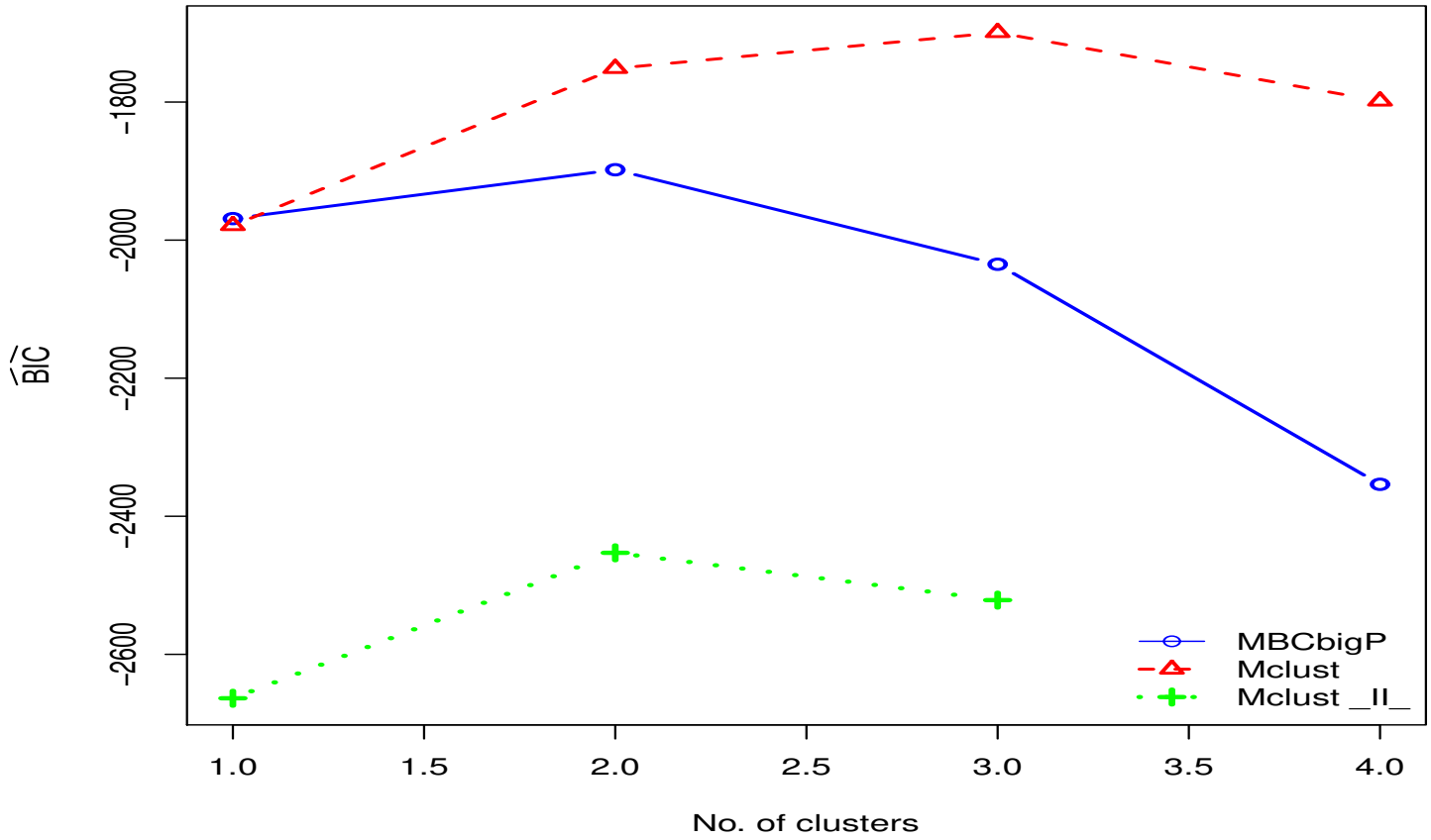
Toy example: bank note data

(Approximate) BIC and adjusted Rand index values:

G	1	2	3	4	adjR
MBCbigP	-1968.81	**-1898.04**	-2034.99	-2353.70	0.96
`mclust` ⫫	-2663.64	**-2452.95**	-2521.47	NA	0.92
`mclust`	-1978.94	-1751.31	**-1699.32**	-1798.55	0.85

- `mclust` with $G = 2$ gives adjR = 0.98.

Toy example: bank note data



DNA methylation data

- Initially examine $p = 2000$ CpG sites known to be differentially methylated between tumour subtypes.
- Use $Q = 25$ segments of size $r = 80$ each.
- Considered $G = 1, 2$. (Not possible to fit $G = 3$.)

	Optimal G	Error Rate	adjR
MBCbigP	2	18%	0.4019
`mclust` ⫫	2	18%	0.4019
`mclust`	2	18%	0.4018

Lots to do...

- M-step for the Σ_{g12} needs to be resolved.
- Examine a 'blended' clustering solution, based on cluster membership at the end of each EM algorithm stage.
- Selection and ordering of segments is influential.
- Variable selection: not all variables contain information.
- Methodological aim:
  develop MBCbigP for a suite of `mclust` models.
- Applied aim:
  cluster cord blood DNA methylation samples with
  $\approx$ 0.5 million CpG sites.

This research was supported by the
Insight Centre for Data Analytics through
Science Foundation Ireland Grant SFI/12/RC/2289.