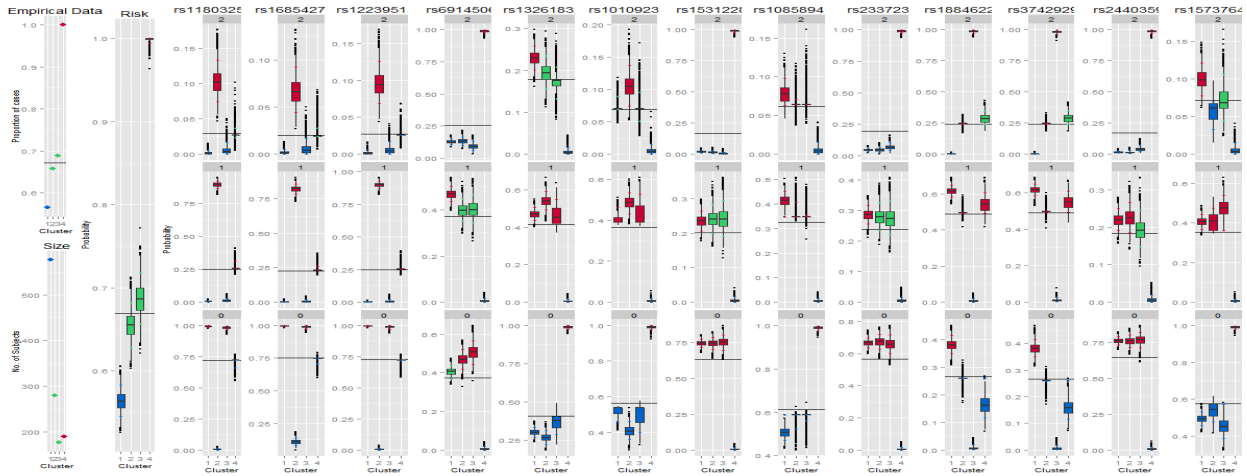
Exploring the presence of complex dependence structures in epidemiological and genetic data through flexible clustering

Sylvia Richardson
MRC Biostatistics Unit



Bayesian week Luminy 29-02-2016

People

Joint work with:

- Michail Papathomas (St Andrews University): relation between the DP and log-linear modelling
- Paul Kirk (MRC Biostatistics): integrative clustering

Collaborators involved in the development of profile regression [based on the Dirichlet process (DP)]:

- Silvia Liverani
- David Hastie
- John Molitor

Motivation

In modern epidemiology, study designs typically include (large sets of) measurements of many types of variables:

- socio-demographic and life style
- environmental/occupational exposure
- genetic variants.

The aim is to **understand the joint effect of risk factors and host characteristics** on complex phenotypes such as cancer and cardiovascular diseases.

Motivation

In modern epidemiology, study designs typically include (large sets of) measurements of many types of variables:

- socio-demographic and life style
- environmental/occupational exposure
- genetic variants.

The aim is to **understand the joint effect of risk factors and host characteristics** on complex phenotypes such as cancer and cardiovascular diseases.

In precision medicine, multiple layers of genetic and genomics information are collected to better characterise the disease state of patients.

The aim is **integrate the different layers of genomics information** to discover groups of patients with distinct molecular phenotypes and clinical outcomes.

Questions

Faced with the tasks

- to explore complex dependence structures among sets of inter-related variables
- to evaluate the effect of combinations of multiple risk factors/molecular characteristics for stratification
- to detect interactions

Fitting linear (or log-linear) models with many interaction parameters becomes quickly unfeasible

Questions

Faced with the tasks

- to explore complex dependence structures among sets of inter-related variables
- to evaluate the effect of combinations of multiple risk factors/molecular characteristics for stratification
- to detect interactions

Fitting linear (or log-linear) models with many interaction parameters becomes quickly unfeasible

Dimensionality reduction using clustering

- Partitions the subjects into groups according to covariate profile
- Flexible Bayesian clustering based on the Dirichlet Process
- Can jointly model covariate patterns and health outcome: **profile regression** (Molitor et al, 2010)

The many aspects of clustering

Unsupervised

- *Exploratory aim*
  - how to uncorporate uncertainty with regards to the clustering
  - how to provide tractable output from post-processing
  - how to discover among many variables those that 'drive' the clustering?
- *Modelling dependence*
  - can clustering provide clues as to the CI structure?
  - can clustering help in the search for interactions?

The many aspects of clustering

Unsupervised

- *Exploratory aim*
  - how to uncorporate uncertainty with regards to the clustering
  - how to provide tractable output from post-processing
  - how to discover among many variables those that 'drive' the clustering?
- *Modelling dependence*
  - can clustering provide clues as to the CI structure?
  - can clustering help in the search for interactions?

Supervised

- *Predictive aim*: covariate patterns corresponding to clusters are linked to outcome, e.g. survival
- *Integrative aim*: exploit multiple data types to improve 'useful' clustering

Outline

Notation

Consider categorical covariates $x_{.p}$, $p = 1, ..., P$. For example,
$x_{.1}$: smokes, does not smoke
$x_{.2}$: drinks, does not drink
$x_{.3}$: exercises, does not exercise

For individual i, denote covariate profile with $x_i = (x_{i1}, ..., x_{iP})$.
For example, x_i=(smokes, drinks, does not exercise).

Notation

For individual i

- $z_i = c$ allocates subject, i, to cluster c.
- $\phi_p^c(x)$ the probability that covariate $x_{.p} = x$, for $z_i = c$.
- Given $z_i = c$, $x_{.p}$ has a multinomial distribution with cluster specific parameters $\phi_p^c = [\phi_p^c(1), ..., \phi_p^c(M_p)]$
- A priori, $\phi_p^c \sim \text{Dirichlet}(\lambda_1, ..., \lambda_{M_p})$
- ψ_c denotes the probability that a subject is assigned to cluster c.

Statistical Framework

For $\phi = \{\phi_p^c,\ c \in N, p = 1, ..., P\}$,

- 'stick-breaking' prior on the allocation weights ψ_c
- $x_i|z, \phi \sim \prod_{p=1}^{P} \phi_p^{z_i}(x_{ip})$ for $i = 1, 2, ..., n$.
- This implies

$$Pr(x_i|\phi, \psi) = \sum_{c=1}^{\infty} \Pr(z_i = c|\psi) \prod_{p=1}^{P} \Pr(x_{ip}|z_i = c) = \sum_{c=1}^{\infty} \psi_c \prod_{p=1}^{P} \phi_p^c(x_{ip}).$$

Using 0-1 variable selection switches

- Identify covariates that contribute more than others to the formation of clusters. [Tadesse et al. (2005), Chung and Dunson (2009); Papathomas et al. (2012)]
- Cluster specific binary indicators, γ_p^c, so that $\gamma_p^c = 1$ when covariate $x_{.p}$ is important for allocating subjects to cluster c; otherwise $\gamma_p^c = 0$.
- Denote with $\pi_p(x)$ the marginal probability that covariate $x_{.p}$ takes the value x in the whole sample.

$$Pr(x_{.p} = x \mid c) = \phi_p^{\#c}(x) = [\phi_p^c(x)]^{\gamma_p^c} \times [\pi_p(x)]^{(1-\gamma_p^c)}. \quad (1)$$

- Prior for switches: given ρ_p, $\gamma_p^c \sim \text{Bernoulli}(1, \rho_p)$.
- We consider a sparsity inducing prior for ρ with an atom at zero:

$$\rho_p \sim 1_{\{w_p=0\}}\delta_0(\rho_p) + 1_{\{w_p=1\}}\text{Beta}(\alpha_\rho, \beta_\rho)$$

where $w_j \sim \text{Bernoulli}(0.5)$.

Similar to Chung and Dunson (JASA, 2009), but in their set up, covariate observations contribute to the likelihood through a regression model. In our case, covariate observations contribute directly to the likelihood, and we introduce $\pi_p(x)$.

Implementation

Implement DP clustering with variable selection in C++ within the R package PReMiuM; Liverani et al. (2015, JSS)

- Normal and discrete covariates

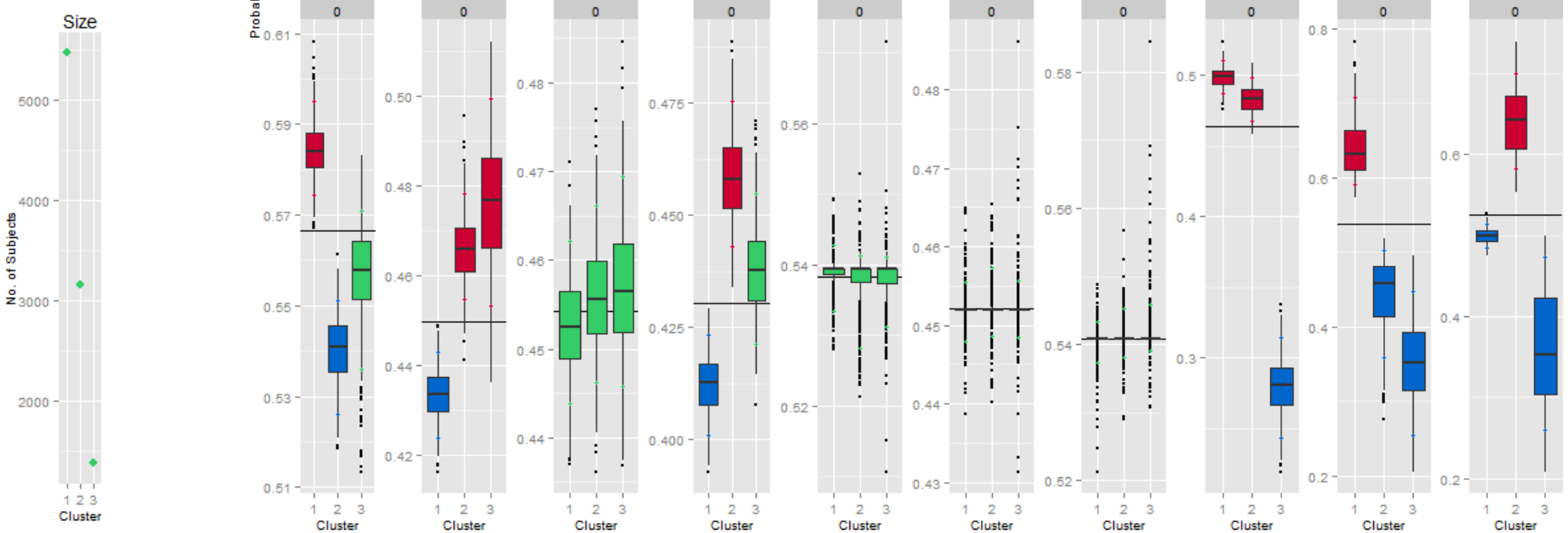
Offers a selection of advanced or standard samplers,

- dependent or independent slice sampling (Kalli et al., 2011)
- or truncated Dirichlet process model (Ishwaran and James, 2001)

Also,

- straightforward to handle missing data (Bayesian paradigm)
- extends to joint modelling of covariates and outcome: **Profile regression** (see later)

Example: 10,000 subjects, 10 binary covariates



Outline

Clustering and log-linear interaction models

- It is not clear how clustering output may translate into interactions in a log-linear regression modelling framework.
- Can we assist the process of comparing a large number of log-linear models with the clustering variable selection results?
- The important aspect of a model that combines clustering and variable selection is that **covariates are not chosen in accordance with size of marginal effect**. They are selected because **they combine to create distinct groups of subjects**. Consequently, we expect that this type of modelling should be able to **inform on interactions** in a log-linear model setting.

Theoretical results

Theorem 1: Consider random variables $x_{.p}$ and $x_{.q}$, $1 \leq p, q \leq P$, $p \neq q$. If $\sum_{c=1}^{C} \gamma_p^c \times \gamma_q^c = 0$ then $x_{.p}$ and $x_{.q}$ are independent.

Theorem 2: Consider a set of random variables $\{x_{.1}, \ldots, x_{.P}\}$. If, for some $p \in \{1, ..., P\}$, $\sum_{c=1}^{C} \gamma_p^c \times \gamma_q^c = 0$, for all $q \neq p$, then $x_{.p}$ is independent of $\{x_{.1}, \ldots, x_{.P}\} \setminus x_{.p}$.

Proofs: See Papathomas and Richardson (2015). Note that the converse is not true.

The previous Theorems imply the following Corollary,

Corollary: Consider covariate $x_{.p}$. If $\sum_{c=1}^{C} \gamma_p^c = 0$ then $x_{.p}$ is independent from all other covariates.

Theoretical results

Therefore,

- if the selection probability ρ_p for $x_{.p}$ is zero or close to zero, something that implies that $\sum_{c=1}^{C} \gamma_p^c$ is also zero or close to zero, we can assume that $x_{.p}$ is **independent from all other covariates**.

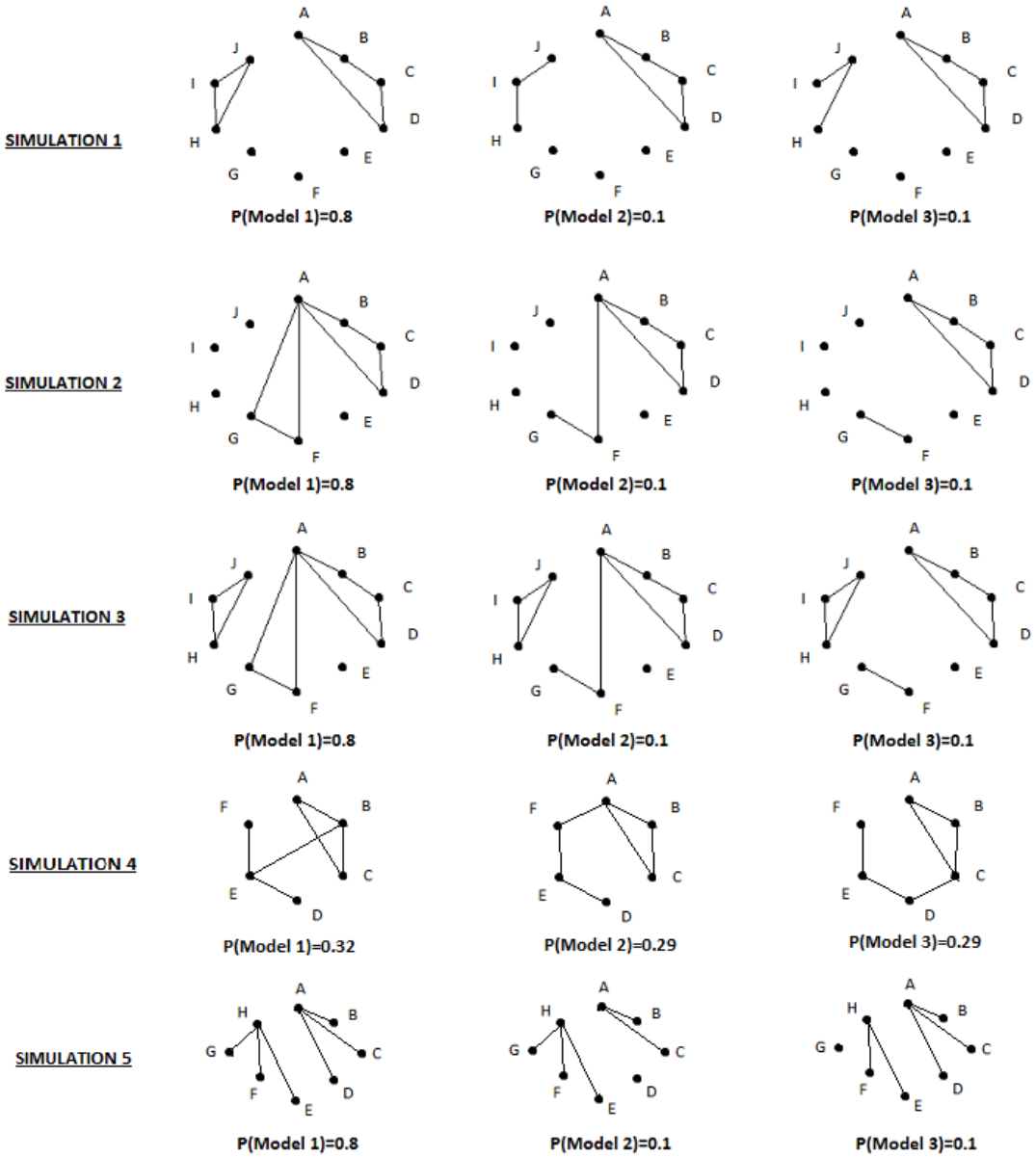
- Assuming that our interest lies in exploring interactions, to reduce the dimensionality of the problem when fitting linear models to sparse contingency tables, $x_{.p}$ **could be removed from the analysis**.

Graphical models for contingency tables

- They allow to visualize and build complex dependence structures for the covariates (classification factors) under consideration
- Models can be interpreted in terms of C.I. using Markov property (local & global)
- Neighbourhoods of models are easily defined, and it is straightforward to move in the space of models by adding, removing or replacing edges.
- Graphical models correspond to a subclass of log-linear models:

$$\text{graph} \rightarrow \text{cliques} \rightarrow \text{interactions of higher order}$$

- The number of possible undirected graphs is 2^H, where $H = P!/(2(P-2)!)$,assuming the intercept and all factor main effects are included in the model. For example, the number of possible graphical models for six covariates is 32768.



e.g. log-linear model corresponding to Simulation 1, Model 1 contains the following interaction terms:
AB + BC + CD + AD + I*J*H
(i.e. IJ + IH + JH + IJH)

Exploiting the output of the clustering

The clustering output can be used in several ways:

- Dimension reduction: eliminate variables from the search for interactions
- Improving the proposals when running an MCMC strategy for graphical model search

We propose

- to summarise the relevant information via a matrix $\boldsymbol{T}_\gamma$, built from **joint selection probabilities** of pairs of variables in clusters
- modify the vanilla MCMC search algorithm by informing addition, removal or swop moves through $\boldsymbol{T}_\gamma$

Construction of matrix $\boldsymbol{T}_\gamma$

- For iteration i_t and for each cluster c with more than one subject, form matrix $\boldsymbol{T}^{c,i_t}$, so that element (p_1, p_2), $1 \leq p_1 < p_2 \leq P$ is either zero or one, and equal to $\gamma^c_{p_1}(i_t) \times \gamma^c_{p_2}(i_t)$. All other matrix cells are empty.
- Sum up all matrices $\boldsymbol{T}^{c,i_t}$, weighing by cluster size, to create an information matrix $\boldsymbol{T}_\gamma$,

$$\boldsymbol{T}_\gamma = \sum_{i_t} \sum_{c} n_{c,i_t} \times \boldsymbol{T}^{c,i_t}.$$

  where n_{c,i_t} is the size of cluster c at iteration i_t.
- For ease of interpretation reweight the elements of $\boldsymbol{T}_\gamma$ so that the maximum element is one, $\boldsymbol{T}_\gamma = (max\{\boldsymbol{T}_\gamma\})^{-1} \times \boldsymbol{T}_\gamma$.

Construction of matrix $\boldsymbol{T}_\gamma$

- For iteration i_t and for each cluster c with more than one subject, form matrix $\boldsymbol{T}^{c,i_t}$, so that element (p_1, p_2), $1 \leq p_1 < p_2 \leq P$ is either zero or one, and equal to $\gamma^c_{p_1}(i_t) \times \gamma^c_{p_2}(i_t)$. All other matrix cells are empty.
- Sum up all matrices $\boldsymbol{T}^{c,i_t}$, weighing by cluster size, to create an information matrix $\boldsymbol{T}_\gamma$,

$$\boldsymbol{T}_\gamma = \sum_{i_t} \sum_{c} n_{c,i_t} \times \boldsymbol{T}^{c,i_t}.$$

  where n_{c,i_t} is the size of cluster c at iteration i_t.
- For ease of interpretation reweight the elements of $\boldsymbol{T}_\gamma$ so that the maximum element is one, $\boldsymbol{T}_\gamma = (max\{\boldsymbol{T}_\gamma\})^{-1} \times \boldsymbol{T}_\gamma$.

Matrix $\boldsymbol{T}_\gamma$

$\boldsymbol{T}_\gamma$ is constructed in such a manner so that if element $t_\gamma(p_1, p_2)$, $1 \leq p_1 < p_2 \leq P$, is close to zero, this implies that an edge between $x_{.p_1}$ and $x_{.p_2}$ is not likely to be present in a highly supported graphical model.

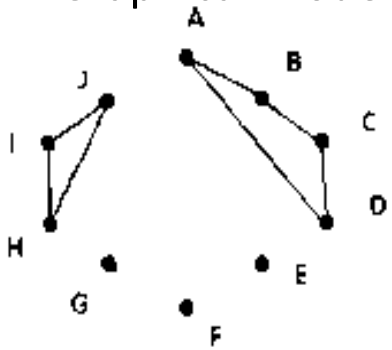
Example. Simulation 1

Consider Simulation 1, with 10000 subjects clustered in accordance with 10 binary covariates. Then,

Table 1: Cluster profiles. In parenthesis the number of subjects typically allocated to each group.

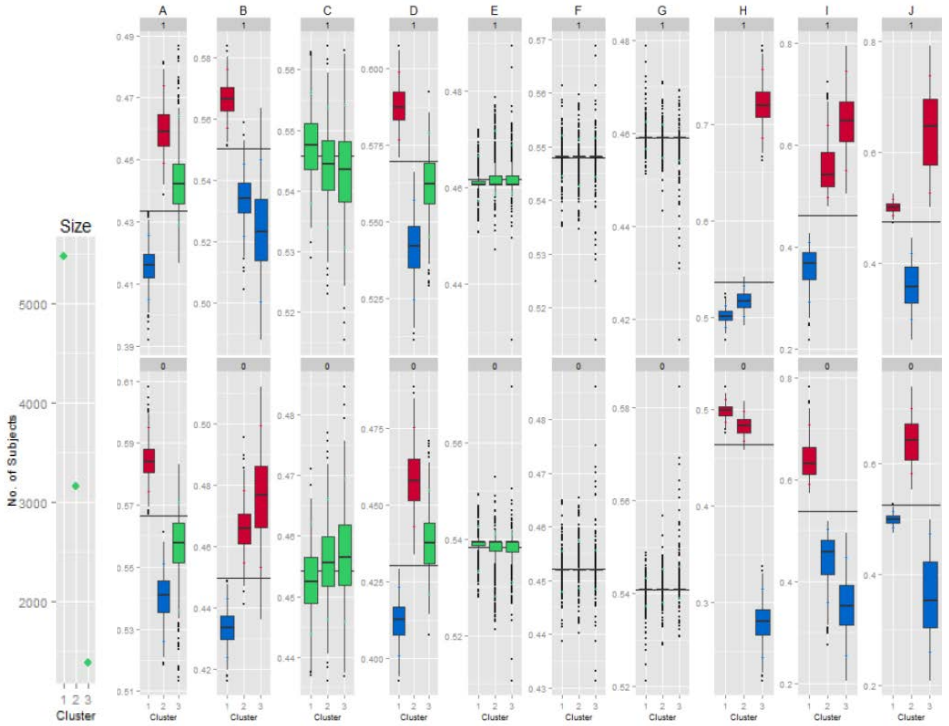
	Simulation 1									
	A	B	C	D	E	F	G	H	I	J
Median(ρ_p)	0.36	0.78	0.32	0.75	0.06	0.05	0.00	0.48	0.57	0.50
Group 1 (5465)	><	<>	00	<>	00	00	00	><	><	<>
Group 2 (3159)	<>	><	00	><	00	00	00	><	<>	><
Group 3 (1376)	00	><	00	00	00	00	00	<>	<>	<>

Simulation 1
Main Graphical Model



Prior prob =0.8
Post prob= 0.55

Clustering output



$$T_{\gamma}^{sim1} = \begin{pmatrix} & A & B & C & D & E & F & G & H & I & J \\ A & & \mathbf{.52} & .08 & \mathbf{.50} & .04 & .02 & .02 & .20 & .27 & .15 \\ B & & & \mathbf{.45} & 1 & .06 & .04 & .03 & .47 & .64 & .47 \\ C & & & & \mathbf{.45} & .02 & .02 & .009 & .12 & .23 & .16 \\ D & & & & & .06 & .04 & .03 & .45 & .65 & .48 \\ E & & & & & & .003 & .003 & .03 & .04 & .03 \\ F & & & & & & & .002 & .02 & .03 & .03 \\ G & & & & & & & & .02 & .02 & .02 \\ H & & & & & & & & & \mathbf{0.61} & \mathbf{.56} \\ I & & & & & & & & & & \mathbf{.74} \end{pmatrix}$$

-- Clear indication that E, F, G, have no interactions
-- Noticeable `spilling effects'

Example. Simulation 5

Consider Simulation 5, with 10000 subjects clustered in accordance with 100 covariates.

- Only 8 out of 100 covariates are involved in the log linear model which generate the variability associated with the cell counts
- After running the clustering, all other 92 covariates have selection probabilities less than 0.01 $\Rightarrow$ can be eliminated from search for most supported graphical models.

Table 2: Cluster profiles. In parenthesis the number of subjects typically allocated to each group. Posterior median selection probabilities for the remaining 92 covariates in Simulation 5 were either equal to zero or smaller than 0.01

	Simulation 5							
	A	B	C	D	E	F	G	H
Median(ρ_p)	0.96	0.95	0.97	0.93	0.97	0.96	0.97	0.96
Group 1 (4036)	><	<>	<>	<>	<>	><	<>	<>
Group 2 (3813)	><	<>	<>	<>	><	<>	><	><
Group 3 (399)	><	00	<>	><	<>	><	><	<>
Group 4 (720)	<>	><	><	><	<>	><	<>	<>
Group 5 (902)	<>	><	><	><	><	<>	><	><
Group 5 (130)	<>	><	><	><	><	<>	><	<>

Example. Simulation 5

$$\boldsymbol{T}_{\gamma}^{sim5} = \begin{pmatrix} & A & B & C & D & E & F & G & H \\ A & & \mathbf{.99} & \mathbf{1} & \mathbf{1} & 1 & 1 & 1 & 1 \\ B & & & .97 & .99 & .99 & .99 & .99 & .99 \\ C & & & & .99 & .99 & .99 & .99 & .99 \\ D & & & & & .99 & 1 & 1 & 1 \\ E & & & & & & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ F & & & & & & & \mathbf{1} & \mathbf{1} \\ G & & & & & & & & \mathbf{1} \end{pmatrix}$$

Mixing performance of samplers

MCMC model search for graphical log-linear models using difference combinations of vanilla (uniformly random) and informed proposal for addition, deletion and swop.

	Simulation 1		
	Acceptance rate as a percentage	Iterations (median) to highest posterior probability model	Posterior probability for highest probability model
(a) Uniformly random	5.1	590 (452,821)	0.55
(b) Cluster specific	3.8	247 (164,369)	0.55
(c) Combined (30%,10%)	5.3	540 (290,674)	0.53
(d) Combined (20%,20%)	4.9	403 (312,493)	0.55
	Simulation 2		
	Acceptance rate as a percentage	Iterations (median) to highest posterior probability model	Posterior probability for highest probability model
(a) Uniformly random (PDV)	4.4	717 (475,990)	0.60
(b) Cluster specific	4.4	189 (147,238)	0.58
(c) Combined (30%,10%)	4.4	417 (346,354)	0.60
(d) Combined (20%,20%)	4.5	257 (181,314)	0.59

Strategy (d) offers a good balance between performance and safeguard against clustering missing an existing edge.

Real data example

- 30 single nucleotide polymorphisms (SNPs) in chromosomes 6 and 15.

  (Data from 4260 subjects in a genome-wide association study of lung cancer presented in Hung *et al.* (2008).)

- 12 SNPs were indicated as important by variable selection within clustering.

  (Two from chromosome 15 and ten from chromosome 6.)

- SNPs were highly correlated. 3 SNPs included in the competing log-linear graphical models as representatives.

  rs8034191 from chromosome 15 and {rs4324798,rs1950081} from chromosome 6.

- Also include age, gender and smoking status in the competing log-linear graphical models, to search for gene-environment interactions.

Real data example

$$\boldsymbol{T}_{\gamma}^{\text{Real data (GE)}} = \begin{pmatrix} & A & B & C & D & E & F \\ A & & 0.002 & .01 & .06 & 0.06 & .06 \\ B & & & .001 & .02 & 0.02 & .02 \\ C & & & & .09 & .07 & .08 \\ D & & & & & \mathbf{1} & \mathbf{.98} \\ E & & & & & & \mathbf{.88} \end{pmatrix}$$

- The highest posterior probability model (P=0.8) from the MCMC search is

  'SNP1+SNP2+SNP3+AGE*GENDER*SMOKING'

  which does not support the presence of gene-gene or gene-environment interactions.

Outline

Mixture modelling for stratified medicine

- In stratified medicine applications, we seek subpopulations (clusters) of patients that have **diagnostic**, **prognostic**, or **theranostic** meaning.

Mixture modelling for stratified medicine

- In stratified medicine applications, we seek subpopulations (clusters) of patients that have **diagnostic**, **prognostic**, or **theranostic** meaning.
- A common strategy is to cluster patients on the basis of 'omics data, and then to see if the clusters correspond to useful strata.
  - e.g. are individuals with poor survival outcomes over-represented in Cluster c?

Mixture modelling for stratified medicine

- In stratified medicine applications, we seek subpopulations (clusters) of patients that have **diagnostic**, **prognostic**, or **theranostic** meaning.
- A common strategy is to cluster patients on the basis of 'omics data, and then to see if the clusters correspond to useful strata.
  - e.g. are individuals with poor survival outcomes over-represented in Cluster c?
- Alternatively, we can include the diagnostic/prognostic/theranostic response in our model, and use this to "guide" the clustering.
- This can help to determine **better quality, more meaningful** strata.

Mixture modelling for stratified medicine

- In stratified medicine applications, we seek subpopulations (clusters) of patients that have **diagnostic**, **prognostic**, or **theranostic** meaning.
- A common strategy is to cluster patients on the basis of 'omics data, and then to see if the clusters correspond to useful strata.
  - e.g. are individuals with poor survival outcomes over-represented in Cluster c?
- Alternatively, we can include the diagnostic/prognostic/theranostic response in our model, and use this to "guide" the clustering.
- This can help to determine **better quality, more meaningful** strata.

Profile regression

We refer to this supervised approach as **profile regression**.

Profile regression: notation

For individual i

y_i	outcome of interest (e.g. case/control, survival, response to therapy).
$\mathbf{w}_i$	fixed effects

and, as before,

$\mathbf{x}_i = (x_{i1}, \ldots, x_{iP})$	covariate profile
$z_i = c$	the allocation variable indicates the cluster to which individual i belongs

Statistical Framework

- Joint covariate and response model

$$f(\mathbf{x}_i, y_i|\phi, \theta, \psi, \beta) = \sum_c \psi_c f(\mathbf{x}_i|z_i = c, \phi_c) f(y_i|z_i = c, \theta_c, \beta, \mathbf{w}_i)$$

Statistical Framework

- Joint covariate and response model

$$f(\mathbf{x}_i, y_i|\phi, \theta, \psi, \beta) = \sum_c \psi_c f(\mathbf{x}_i|z_i = c, \phi_c) f(y_i|z_i = c, \theta_c, \beta, \mathbf{w}_i)$$

- Mixture model jointly for covariate and response
- For example, for Bernoulli outcome

$$\text{logit}\{p(y_i = 1|\theta_c, \beta, \mathbf{w}_i)\} = \theta_c + \beta^T \mathbf{w}_i$$

Statistical Framework

- Joint covariate and response model

$$f(\mathbf{x}_i, y_i|\phi, \theta, \psi, \beta) = \sum_c \psi_c f(\mathbf{x}_i|z_i = c, \phi_c) f(y_i|z_i = c, \theta_c, \beta, \mathbf{w}_i)$$

- Mixture model jointly for covariate and response
- For example, for Bernoulli outcome

$$\text{logit}\{p(y_i = 1|\theta_c, \beta, \mathbf{w}_i)\} = \theta_c + \beta^T \mathbf{w}_i$$

- The association of the profiles with the response are characterised by the risk effect parameters θ_c

Statistical Framework

- Joint covariate and response model

$$f(\mathbf{x}_i, y_i|\phi, \theta, \psi, \beta) = \sum_c \psi_c f(\mathbf{x}_i|z_i = c, \phi_c) f(y_i|z_i = c, \theta_c, \beta, \mathbf{w}_i)$$

- Mixture model jointly for covariate and response
- For example, for Bernoulli outcome

$$\text{logit}\{p(y_i = 1|\theta_c, \beta, \mathbf{w}_i)\} = \theta_c + \beta^T \mathbf{w}_i$$

- The association of the profiles with the response are characterised by the risk effect parameters θ_c
- Note: the above framework, adopted in Molitor et al. (Biostatistics, 2010), is similar in spirit to Bigelow and Dunson (JASA, 2009).

Profile regression applications

- Implemented in R package: **PReMiuM**
- In Molitor et al. (2010): epidemiological application to NCSH data exploring the link between a child mental health and socio demographic, family and living conditions
- In Hastie et al. (2013): epidemiological application to estimate risk functions associated with multidimensional exposure profiles (e.g. smoking and lung cancer)
- In Papathomas et al. (2012): investigation of how profile regression combined with variable selection can *highlight combinations of SNPs* associated with higher disease risk in a genetic association study of lung cancer (Hung et al 2008).

But how to cope with multiple datasets, each possessing a (potentially) different clustering structure?

Multiple dataset challenges

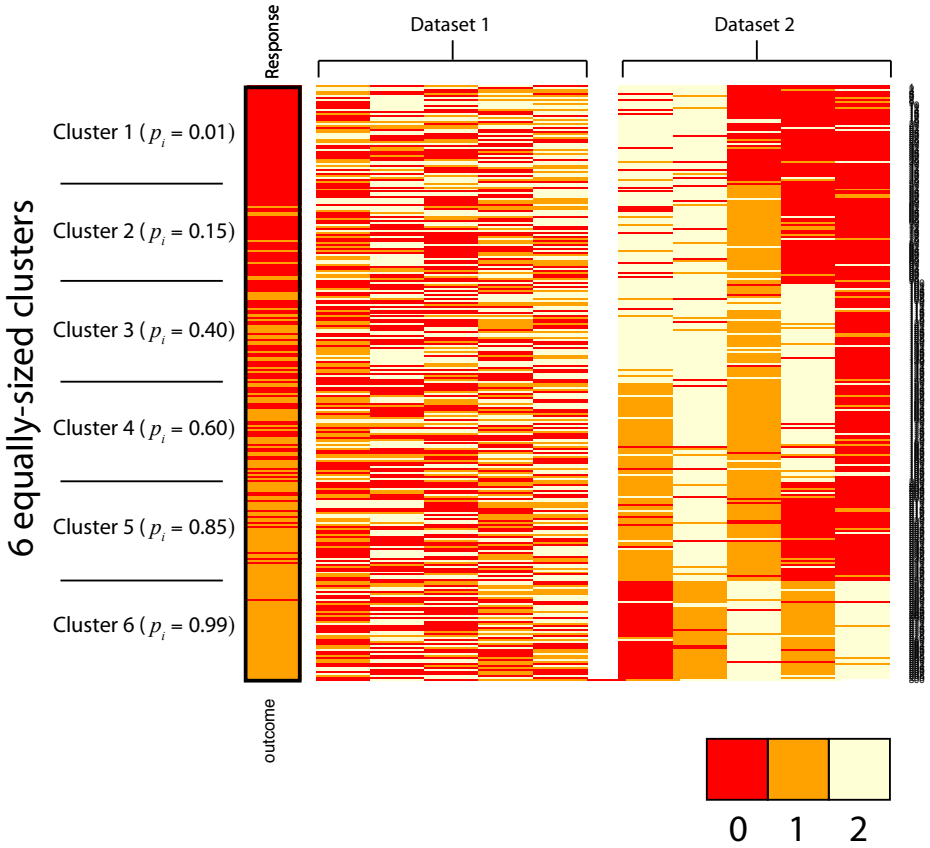
- It is increasingly common that we have multiple 'omics measurements for each patient
  - mRNA, miRNA, genotype, methylation, protein, ...
  - Several unsupervised methods for their integration, e.g. Bayesian correlated clustering (Kirk et al., 2012), iCluster (Shen et al., 2009)

Multiple dataset challenges

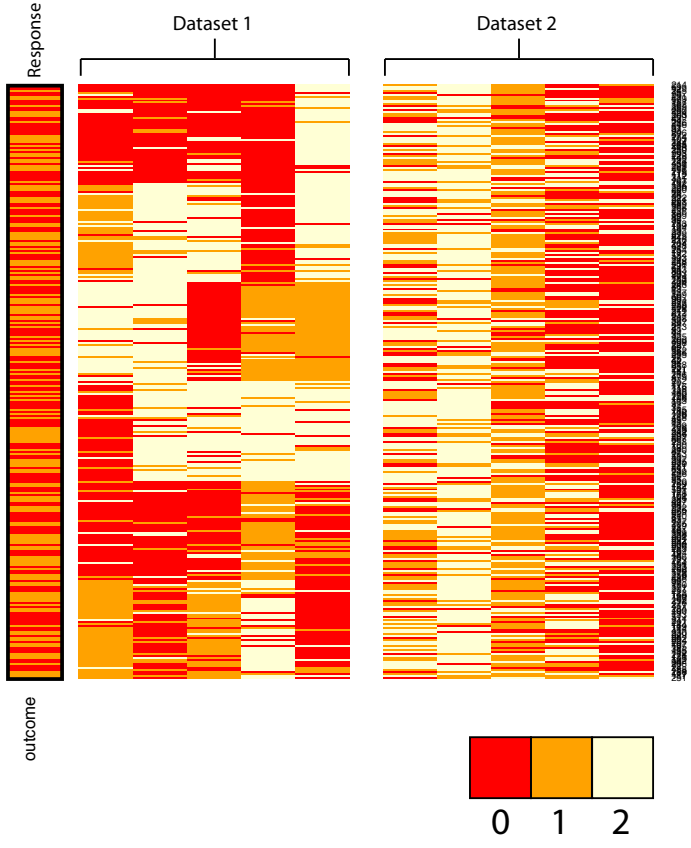
- It is increasingly common that we have multiple 'omics measurements for each patient
  - mRNA, miRNA, genotype, methylation, protein, ...
  - Several unsupervised methods for their integration, e.g. Bayesian correlated clustering (Kirk et al., 2012), iCluster (Shen et al., 2009)
- Some datasets will be more/less relevant for stratifying patients.
  - We wish to **assess dataset relevance**.
- There may be multiple strong clustering structures within each dataset, but not all will be useful for stratification.
  - We wish to determine the **relevant clustering structure**.

Illustration

Two categorical datasets:



Rows ordered to show Dataset 2 structure.



Rows ordered to show Dataset 1 structure.

Dataset 2 has a "relevant" clustering structure; Dataset 1 does not.

Outline

Assessing dataset relevance: pragmatic approach

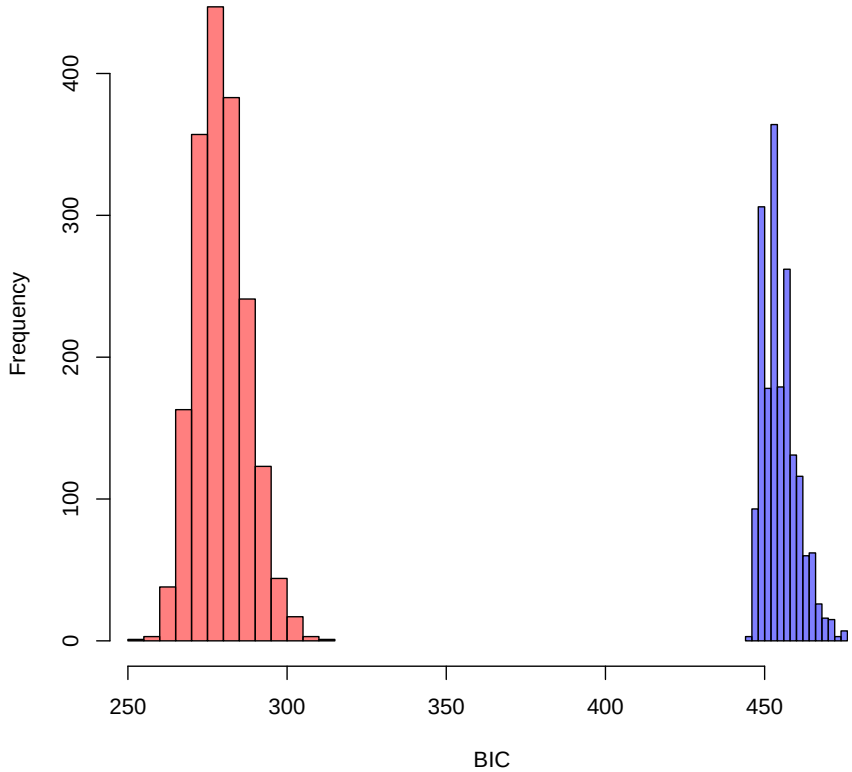
- In profile regression, each draw from the posterior defines a clustering of the data, and an associated predictive model.
- A **less effective** (!) alternative to profile regression would be a 2-step approach:
  1. Obtain a clustering, $\boldsymbol{z}$, of the data; and then
  2. Estimate the remaining parameters of the predictive model.
- In either case, we can score the relative prediction quality of different partitions, $\boldsymbol{z}$, e.g.

$$\text{BIC} \mid \boldsymbol{z} = -2\ln(L) + a_{\boldsymbol{z}} \ln(n)$$

  - L is the likelihood associated with the observed $y \mid \boldsymbol{z}$, e.g. logit;
  - a is the number of parameters in the predictive model (which will grow with the number of clusters);
  - n is the number of observations.

Assessing dataset relevance: illustration

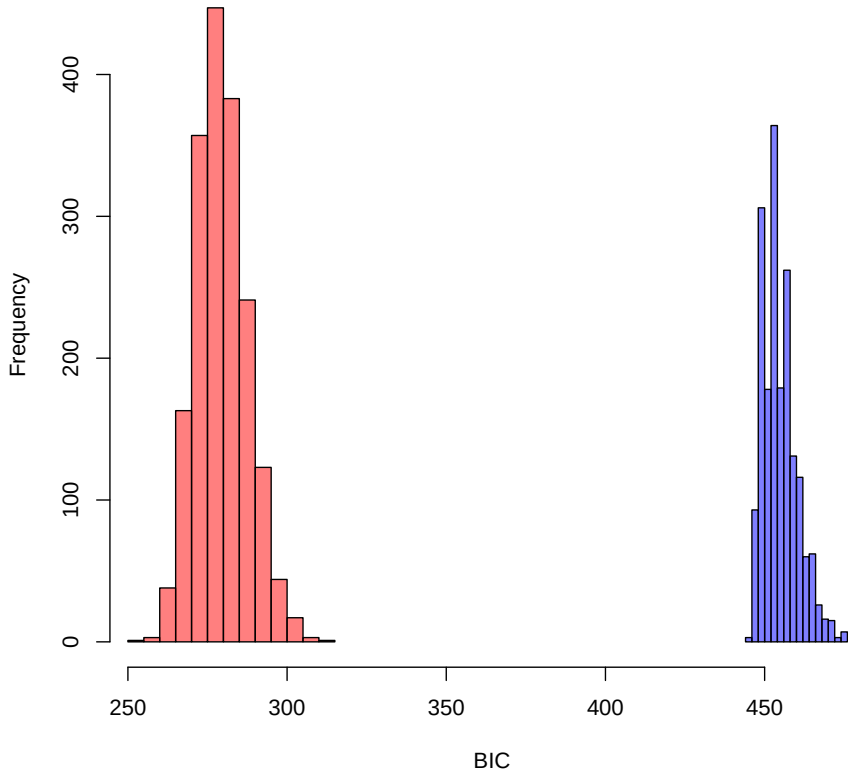
- For our 2 categorical datasets with the relevant and irrelevant clusterings, use PReMiuM to fit profile regression models to each.
- For clusterings sampled from the resulting posteriors, calculate BIC scores:



Clearly the partitions derived from Dataset 2 (red) result in much better models for the response than those from Dataset 1 (blue).

Assessing dataset relevance: illustration

- For our 2 categorical datasets with the relevant and irrelevant clusterings, use PReMiuM to fit profile regression models to each.
- For clusterings sampled from the resulting posteriors, calculate BIC scores:

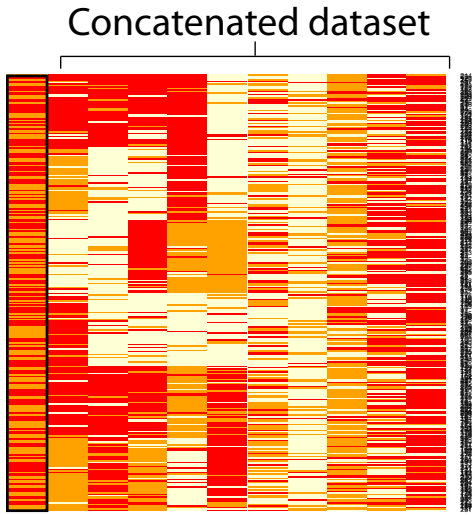


Clearly the partitions derived from Dataset 2 (red) result in much better models for the response than those from Dataset 1 (blue).

Dataset 2 is **more relevant** for this stratification task.

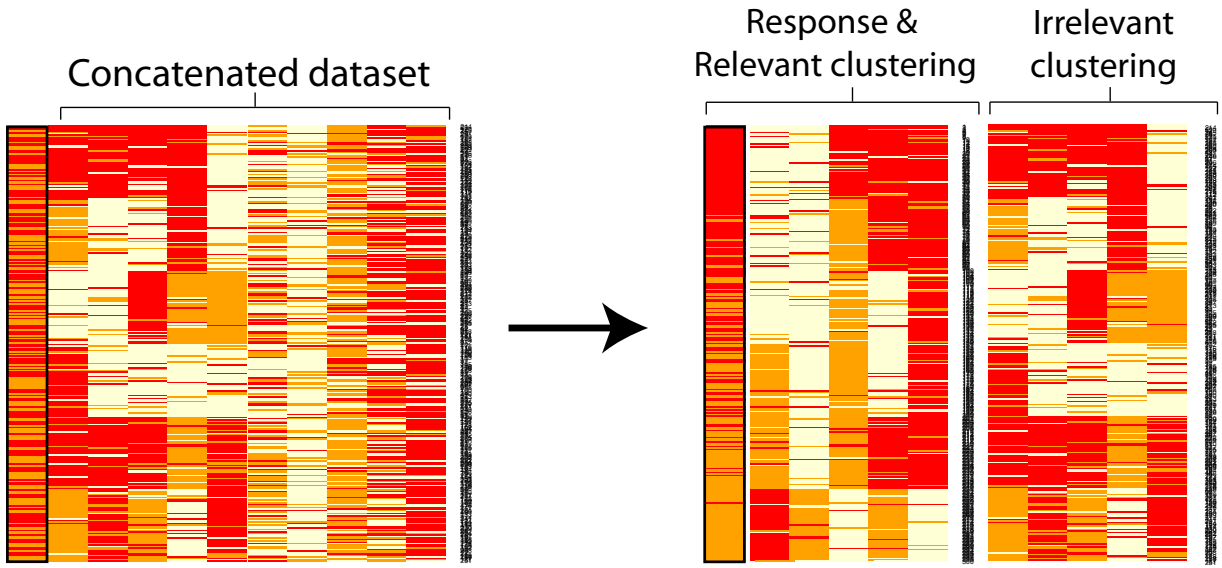
Confusing clusterings

- In practice, our datasets are much more complex (!)
- Typically, individual datasets will possess both relevant and irrelevant clustering structures
  - e.g. a concatenation of Datasets 1 and 2.



Confusing clusterings

- In practice, our datasets are much more complex (!)
- Typically, individual datasets will possess both relevant and irrelevant clustering structures
  - e.g. a concatenation of Datasets 1 and 2.
- This presents a challenge for profile regression:
  - Can we ensure that we identify the **relevant** clustering structure?

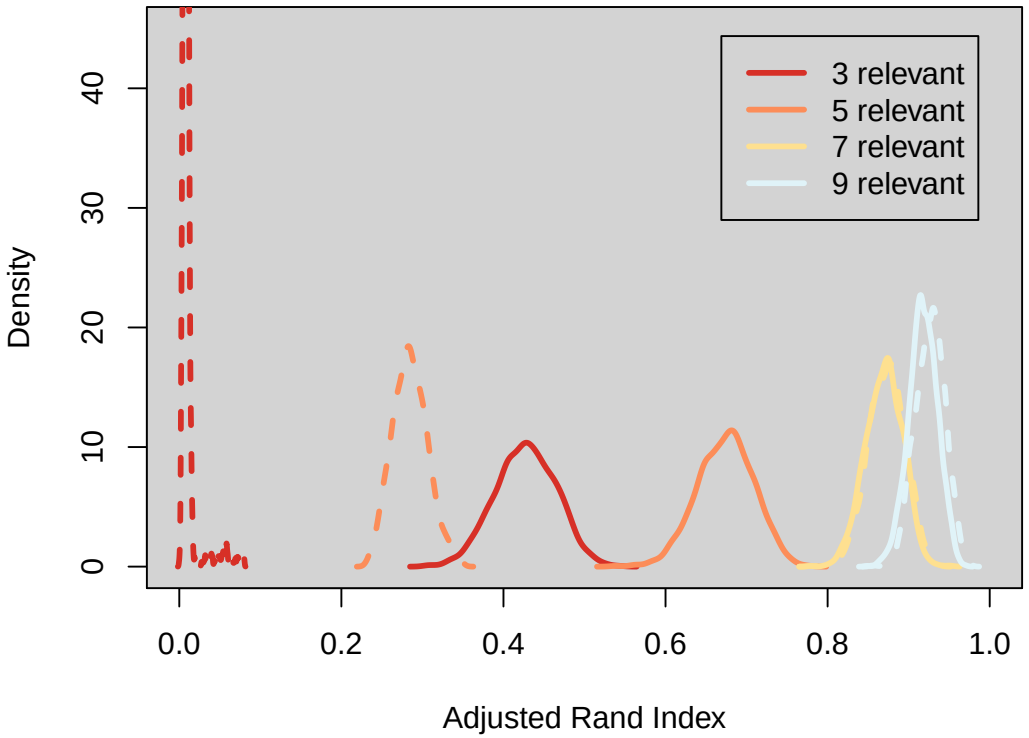


Confusing clusterings: effects on profile regression

Simulation study

- We simulate datasets similar to the concatenation of Datasets 1 and 2 ($p = 10$ covariates + 1 response), but vary the numbers of covariates contributing to the relevant clustering structure.
- We calculate the adjusted Rand index (ARI) between the true relevant clustering structure, and:
  1. The structure inferred using PReMiuM, if we first remove the irrelevant covariates (**gold standard**).
  2. The structure inferred using PReMiuM with variable selection.

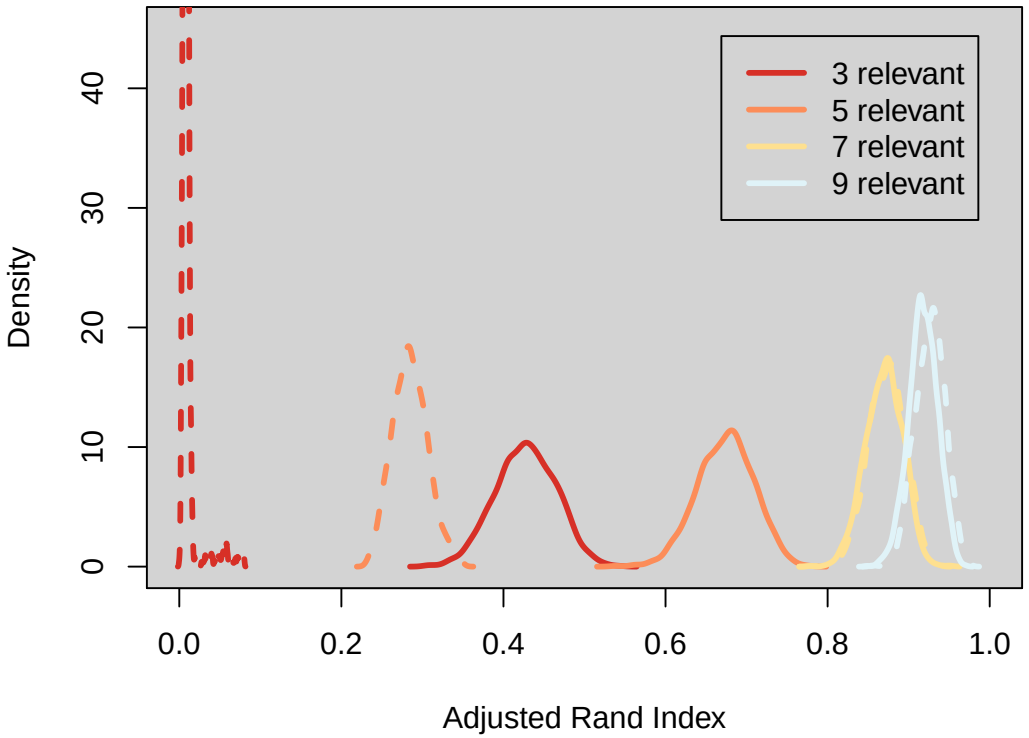
Confusing clusterings: effects on profile regression



Solid lines: gold standard ARI values.

Dashed lines: ARI values obtained using PReMiuM with variable selection.

Confusing clusterings: effects on profile regression

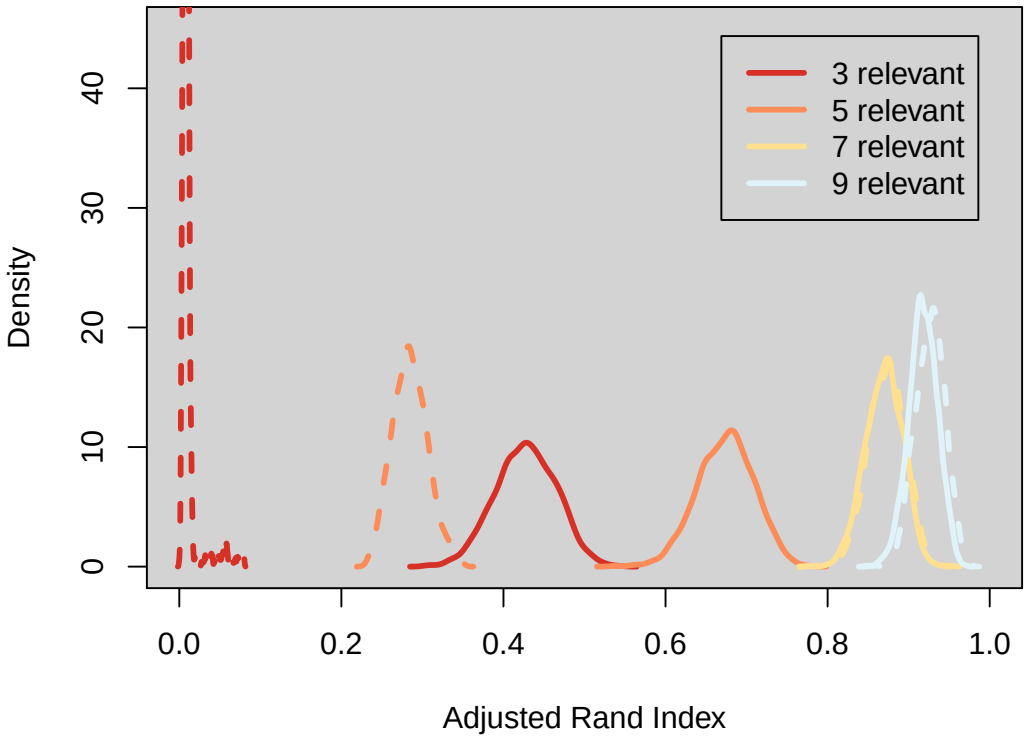


If there are many more irrelevant than relevant covariates, we will home in on the **irrelevant** clustering structure.

Solid lines: gold standard ARI values.

Dashed lines: ARI values obtained using PReMiuM with variable selection.

Confusing clusterings: effects on profile regression



Solid lines: gold standard ARI values.

Dashed lines: ARI values obtained using PReMiuM with variable selection.

If there are many more irrelevant than relevant covariates, we will home in on the **irrelevant** clustering structure.

Can we resolve this?

Outline

Multiclustering: work in progress

- We refer to the problem of picking out multiple clustering structures from a single dataset as **multiclustering**.
  - A form of **biclustering**, in which similarity between covariates is defined in terms of the similarity in their clustering structure.

Multiclustering: work in progress

- We refer to the problem of picking out multiple clustering structures from a single dataset as **multiclustering**.
  - A form of **biclustering**, in which similarity between covariates is defined in terms of the similarity in their clustering structure.
- In unsupervised clustering, we may wish to identify **all** clustering structures in the dataset.

Multiclustering: work in progress

- We refer to the problem of picking out multiple clustering structures from a single dataset as **multiclustering**.
  - A form of **biclustering**, in which similarity between covariates is defined in terms of the similarity in their clustering structure.
- In unsupervised clustering, we may wish to identify **all** clustering structures in the dataset.
- For profile regression, the main aim is to identify the clustering structure that is **most predictive** of the response.
  - We will call the associated covariates 'relevant'.

Multiclustering

- Suppose we know that the covariates can be partitioned into 3 disjoint sets: $V^{(1)}$, $V^{(2)}$ and $V^{(0)}$.
  - Those in $V^{(1)}$ define the **relevant** clustering structure;
  - Those in $V^{(2)}$ define an **irrelevant** clustering structure; and
  - Those in $V^{(0)}$ **do not contribute** to any clustering structure.

Multiclustering

- Suppose we know that the covariates can be partitioned into 3 disjoint sets: $V^{(1)}$, $V^{(2)}$ and $V^{(0)}$.
  - Those in $V^{(1)}$ define the **relevant** clustering structure;
  - Those in $V^{(2)}$ define an **irrelevant** clustering structure; and
  - Those in $V^{(0)}$ **do not contribute** to any clustering structure.
- We wish to allocate each covariate to **precisely one** of these sets.

Multiclustering

- Suppose we know that the covariates can be partitioned into 3 disjoint sets: $V^{(1)}$, $V^{(2)}$ and $V^{(0)}$.
  - Those in $V^{(1)}$ define the **relevant** clustering structure;
  - Those in $V^{(2)}$ define an **irrelevant** clustering structure; and
  - Those in $V^{(0)}$ **do not contribute** to any clustering structure.
- We wish to allocate each covariate to **precisely one** of these sets.
- Consider categorical indicators, $\gamma_j \in \{0, 1, 2\}$, so that

$$\gamma_j = k \iff x_j \in V^{(k)}, \qquad k = 0, 1, 2$$

Multiclustering

- We model two separate clustering structures associated with $V^{(1)}$ and $V^{(2)}$ using Dirichlet process mixture models (as previously).
- Define $z_i^{(\ell)}$ to be the latent allocation variable describing the allocation of the i-th individual in the ℓ-th mixture model, $\ell = 1, 2$.
- Similarly:
  - $\phi_c^{(\ell)}$ are the parameters associated with the c-th component in the ℓ-th mixture model; and
  - $\psi^{(\ell)}$ are the mixture weights for the ℓ-th mixture model.

  The discrete covariate model is then

$$f(\mathbf{x}_i \,|\gamma, \{z^{(\ell)}\}, \{\phi^{(\ell)}\}, \phi^{(0)}) =$$
$$\prod_{j=1}^{J} \left(\phi^{(1)}_{z_i^{(1)},\, j}(x_{ij})\right)^{1_{\gamma_j=1}} \left(\phi^{(2)}_{z_i^{(2)},\, j}(x_{ij})\right)^{1_{\gamma_j=2}} \left(\phi_{0,\, j}(x_{ij})\right)^{1_{\gamma_j=3}}$$

Multiclustering

The joint model for $\mathbf{x}$ and y given the parameters and fixed effects is:

$$f(\mathbf{x}, y \mid \gamma, \{z^{(\ell)}\}, \{\phi^{(\ell)}\}, \phi^{(0)}, \theta, \beta, \mathbf{w}) =$$
$$\prod_{i=1}^{n} p(y_i|\theta, \beta, \mathbf{w}_i, z_i^{(1)}) f(\mathbf{x}_i \mid \gamma, \{z^{(\ell)}\}, \{\phi^{(\ell)}\}, \phi^{(0)}),$$

Multiclustering

The joint model for $\mathbf{x}$ and y given the parameters and fixed effects is:

$$f(\mathbf{x}, y \mid \gamma, \{z^{(\ell)}\}, \{\phi^{(\ell)}\}, \phi^{(0)}, \theta, \beta, \mathbf{w}) = \prod_{i=1}^{n} p(y_i|\theta, \beta, \mathbf{w}_i, z_i^{(1)}) f(\mathbf{x}_i \mid \gamma, \{z^{(\ell)}\}, \{\phi^{(\ell)}\}, \phi^{(0)}),$$

where $f(\mathbf{x}_i \mid \gamma, \{z^{(\ell)}\}, \{\phi^{(\ell)}\}, \phi^{(0)})$ is as just given, and

$$f(y_i|\theta, \beta, \mathbf{w}_i, z_i^{(1)}) = q^{y_i}(1-q)^{y_i}, \text{ with } q = \text{logit}^{-1}(\theta_{z_i^{(1)}} + \beta^T \mathbf{w}_i).$$

Multiclustering

The joint model for $\mathbf{x}$ and y given the parameters and fixed effects is:

$$f(\mathbf{x}, y \mid \gamma, \{z^{(\ell)}\}, \{\phi^{(\ell)}\}, \phi^{(0)}, \theta, \beta, \mathbf{w}) = \prod_{i=1}^{n} p(y_i|\theta, \beta, \mathbf{w}_i, z_i^{(1)}) f(\mathbf{x}_i \mid \gamma, \{z^{(\ell)}\}, \{\phi^{(\ell)}\}, \phi^{(0)}),$$

where $f(\mathbf{x}_i \mid \gamma, \{z^{(\ell)}\}, \{\phi^{(\ell)}\}, \phi^{(0)})$ is as just given, and

$$f(y_i|\theta, \beta, \mathbf{w}_i, z_i^{(1)}) = q^{y_i}(1-q)^{y_i}, \text{ with } q = \text{logit}^{-1}(\theta_{z_i^{(1)}} + \beta^T \mathbf{w}_i).$$

Crucially, note that we model y_i as being conditionally dependent on $z_i^{(1)}$, but **not** on $z_i^{(2)}$.

Thus, the model for y_i depends on the cluster allocation of the i-th individual only in the **relevant** clustering structure, and **not** in the **irrelevant** clustering structure.

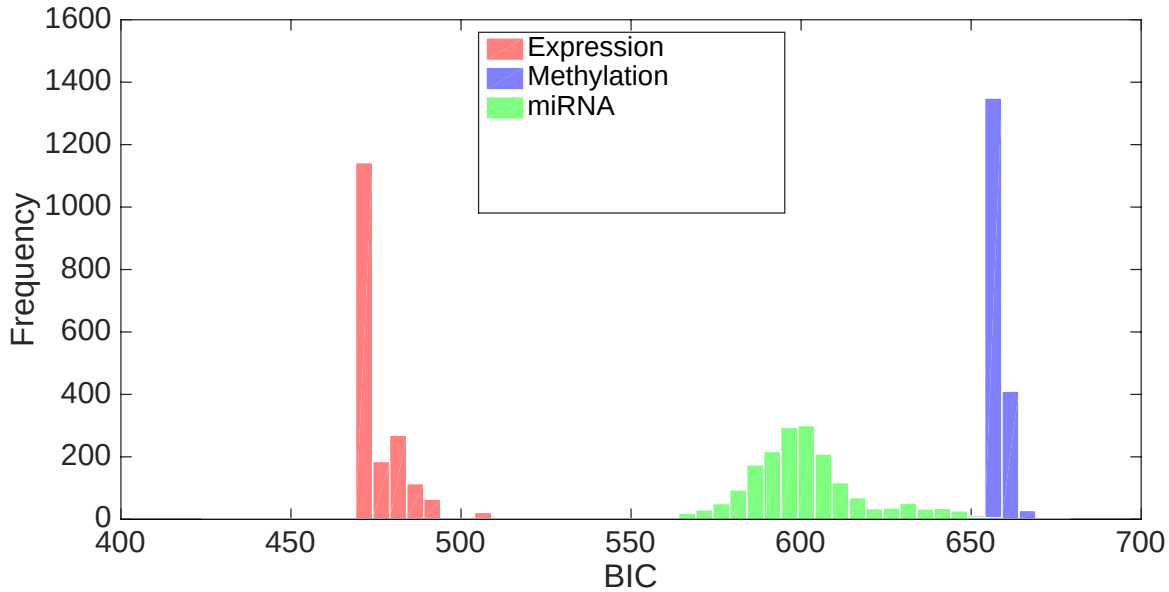
Application: tumour subtype characterisation

- A motivating application for multiclustering is **tumour subtyping**.
- Tumours are often subtyped on the basis of morphological features, with **different subtypes having different prognoses**.
- There is considerable interest in:
  1. Characterising the molecular profiles of existing subtypes; and
  2. Determining novel subtypes from molecular data.
- We would like to be able to exploit all available molecular datasets, and avoid the *confusing clusterings* problem.
- TCGA (The Cancer Genome Atlas) is a repository of molecular datasets for different cancers.
- We wish to use these datasets to identify molecular profiles for 4 known breast cancer subtypes (Luminal A/B, Her2, basal-like).

Application: tumour subtype characterisation

Multiclustering is a work in progress, but we can still use the method described before to **assess the relevance of different molecular datasets** for the breast cancer subtype prediction problem.

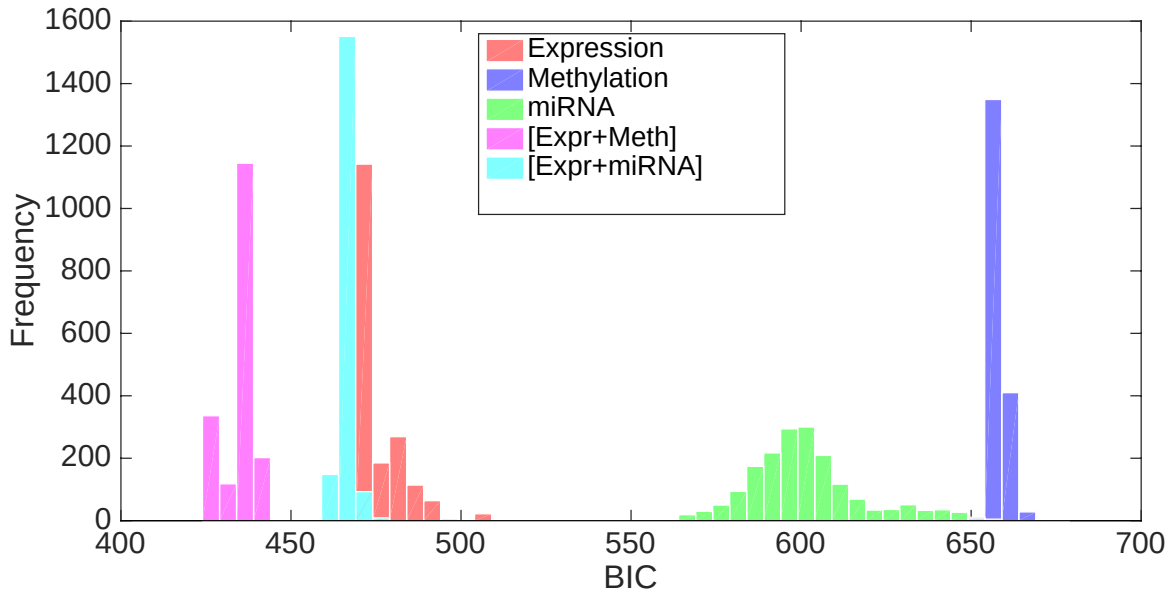
Individual datasets $n = 348$



Application: tumour subtype characterisation

Multiclustering is a work in progress, but we can still use the method described before to **assess the relevance of different molecular datasets** for the breast cancer subtype prediction problem.

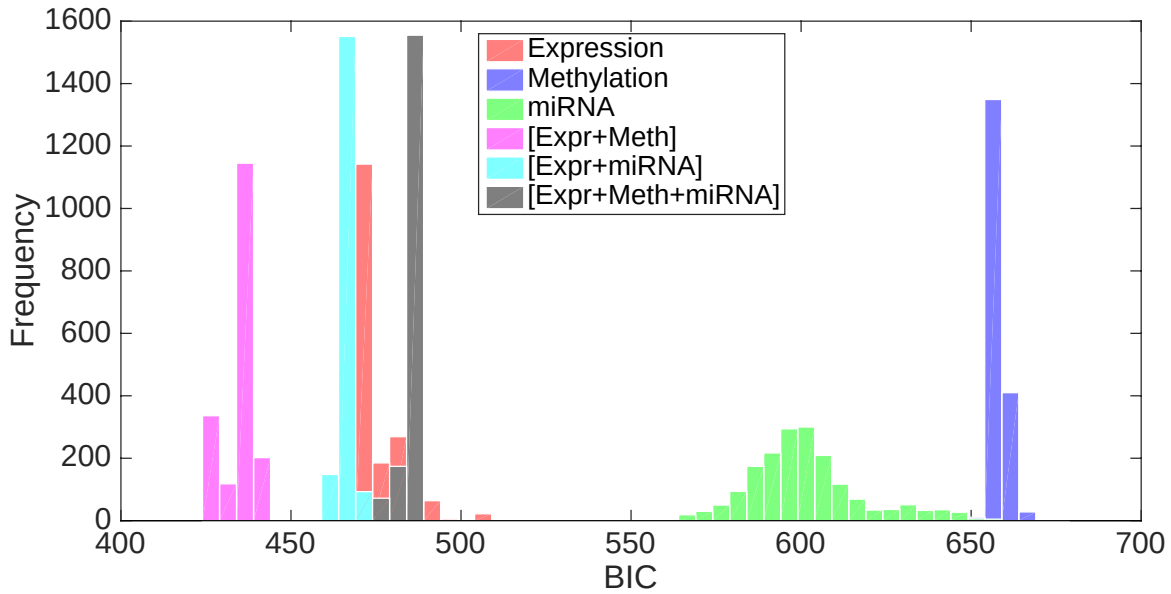
Individual datasets + 2 concatenated pairs



Application: tumour subtype characterisation

Multiclustering is a work in progress, but we can still use the method described before to **assess the relevance of different molecular datasets** for the breast cancer subtype prediction problem.

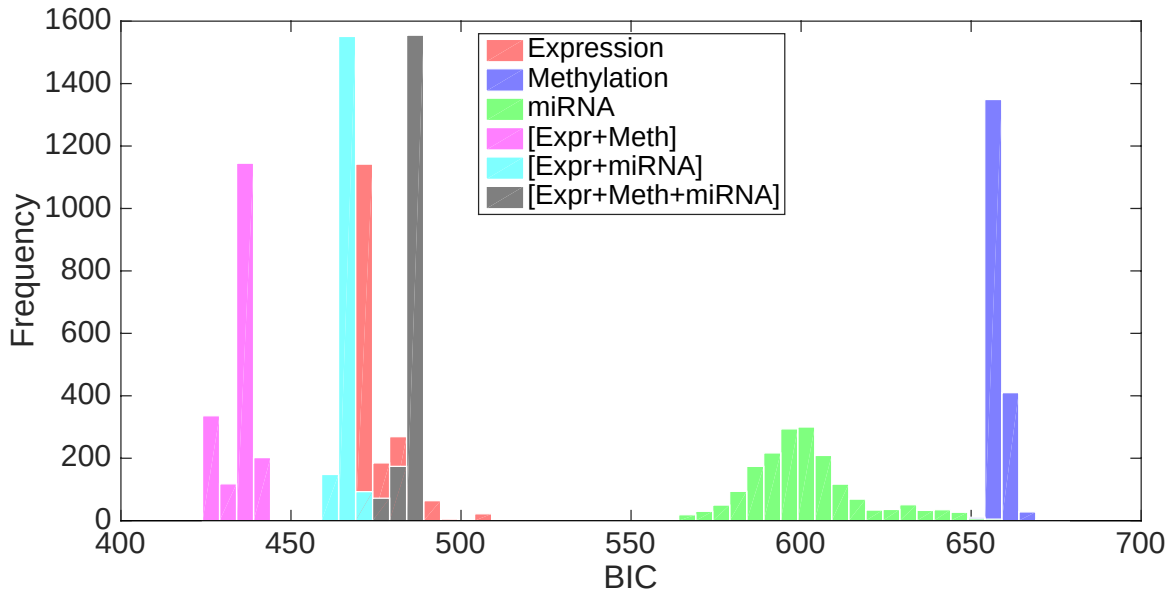
Individual datasets + 2 concatenated pairs + 1 concatenated triple



Application: tumour subtype characterisation

Multiclustering is a work in progress, but we can still use the method described before to **assess the relevance of different molecular datasets** for the breast cancer subtype prediction problem.

Individual datasets + 2 concatenated pairs + 1 concatenated triple



Crudely concatenating all datasets is not always the best option!

Concluding comments

- DP Clustering is a natural and flexible approach to investigate dependence and complex multifactorial effects
- There is no direct correspondence between variables that co-cluster and existence of interactions **BUT**
  - useful summaries can be extracted from clustering output
  - for categorical variables, can eliminate covariates irrelevant for clustering in the search for interactions:

    $\rightarrow$ huge reduction of model space
  - other useful summaries?

Concluding comments

- In biomedicine, supervised clustering is particularly appealing
- There are clear benefits to use *joint modelling* of cluster structure and outcome as implemented in profile regression
- Increasing number of complementary data sets are becoming available for better characterisation of patients profiles, and for treatment stratification
- Current integrative clustering models (e.g. iCluster, MDI, BCC, ...) are unsupervised and treat all data sources on an equal footing
  - We show that this can lead to a 'confused' (unhelpful for prediction) clustering structure
  - We propose a new supervised clustering model which distinguishes 2 types of clustering structures (or more?)

Some References

- Chung, Y., Dunson, D.B. (2009) Nonparametric Bayes conditional distribution modelling with variable selection. J. Am. Stat. Assoc. 104, 1646-60.
- Kirk, P., Griffin, J.E., Savage, R., Ghahramani, Z. and Wild, D.L. (2012) Bayesian correlated clustering to integrate multiple datasets. Bioinformatics, 28(24), 3290–3297.
- Liverani, S., Hastie, D. I., Papathomas, M., Richardson, S. (2013) PReMiuM: An R package for Profile Regression Mixture Models using Dirichlet Processes. Journal of Statistical Software, 64, Issue 7.
- J. T. Molitor, M. Papathomas, M. Jerrett and S. Richardson (2010) Bayesian Profile Regression with an Application to the National Survey of Children's Health, Biostatistics, 11, 484-498.
- Papathomas, M., Molitor, J., Richardson, S., Riboli E. and Vineis P. (2011) Examining the joint effect of multiple risk factors using exposure risk profiles: lung cancer in non smokers. Environmental Health Perspectives, 119, 84-91.
- Papathomas, M , Molitor, J, Hoggart, C, Hastie, D and Richardson, S (2012) Exploring data from genetic association studies using Bayesian variable selection and the Dirichlet process: application to searching for gene-gene patterns. Genetic Epidemiology. 36:663-674.
- Papathomas, M. and Richardson, S. (2016): Exploring dependence between categorical variables: benefits and limitations of using variable selection within Bayesian clustering in relation to log-linear modelling with interaction terms. Journal of Statistical Planning and Inference