MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

PAUL DW KIRK

INTEGRATIVE CLUSTERING APPROACHES FOR MULTI-OMICS DATASETS

INTEGRATIVE CLUSTERING:

- Identifying meaningful subgroups is a key task in statistical omics and precision medicine, e.g.
  - **Clustering genes:** identifying groups of genes that are functionally related.
  - **Clustering patients:** identifying people who will respond similarly to treatments, or have similar disease progression.

INTEGRATIVE CLUSTERING:

- Omics datasets are characteristically:
  - **High dimensional**.
    - Large p (1,000s or 10,000s of variables), small n (often)
    - How do we decide which variables are relevant for defining clusters?
  - **Diverse**
    - Multiple different types of data (transcriptomic, proteomic, …)
- We would like to be able to share information across multiple different data types when identifying subgroups.

INTEGRATIVE CLUSTERING:

- Many methods have been proposed to identify latent structure *shared across multiple omics layers*
    - iCluster
        - Shen et al. 2009
    - COCA (Cluster-of-Clusters Analysis)
        - Hoadley et al. 2014
    - MDI (Multiple Dataset Integration)
        - Kirk et al. 2012
    - …

INTEGRATIVE CLUSTERING:

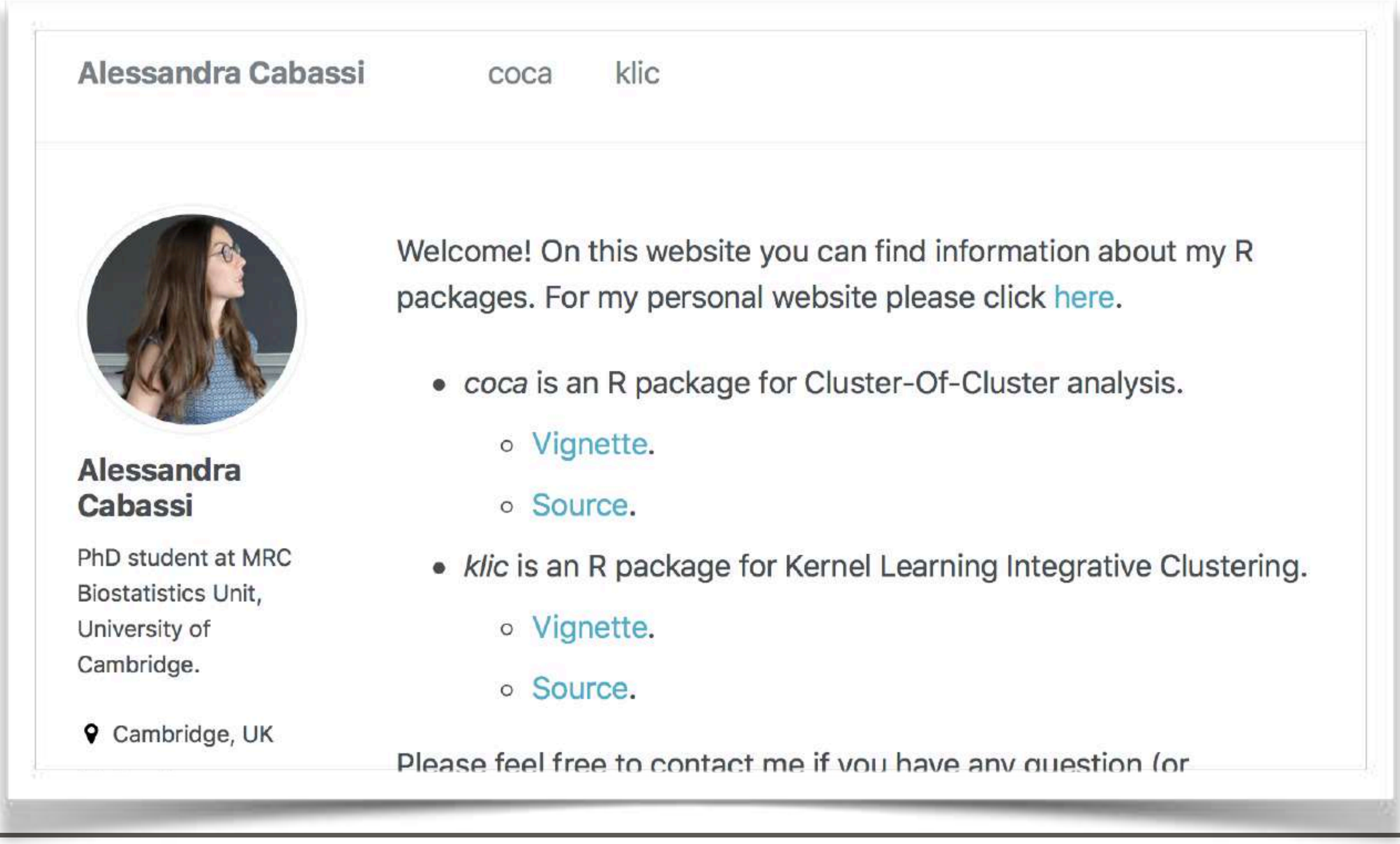
- Many methods have been proposed to identify latent structure *shared across multiple omics layers*
    - iCluster
        - Shen et al. 2009
    - COCA (Cluster-of-Clusters Analysis)
        - Hoadley et al. 2014
    - MDI (Multiple Dataset Integration)
        - Kirk et al. 2012
    - …

But actually there can be multiple latent structures *within each dataset*: how can we find these, and identify “useful” structure?

AD BREAK

KLIC: KERNEL LEARNING FOR INTEGRATIVE CLUSTERING

Alessandra Cabassi



Alessandra Cabassi

PhD student at MRC Biostatistics Unit, University of Cambridge.

Cambridge, UK

Welcome! On this website you can find information about my R packages. For my personal website please click here.

- *coca* is an R package for Cluster-Of-Cluster analysis.
  - Vignette.
  - Source.
- *klic* is an R package for Kernel Learning Integrative Clustering.
  - Vignette.
  - Source.

Please feel free to contact me if you have any question (or

- Cabassi & **Kirk** (2018), Multiple kernel learning for integrative consensus clustering. In preparation.
  - https://acabassi.github.io

TALK OUTLINE:

- Part 1: Illustrations and intuition
- Part 2: Profile regression (semi-supervised clustering)
- Part 3: Semi-supervised *multiview* clustering
- Part 4: Examples
- Part 5: Wrap up

PART 1:

PART 1:
ILLUSTRATIONS AND INTUITION

ILLUSTRATION 1: COLOURFUL PEOPLE

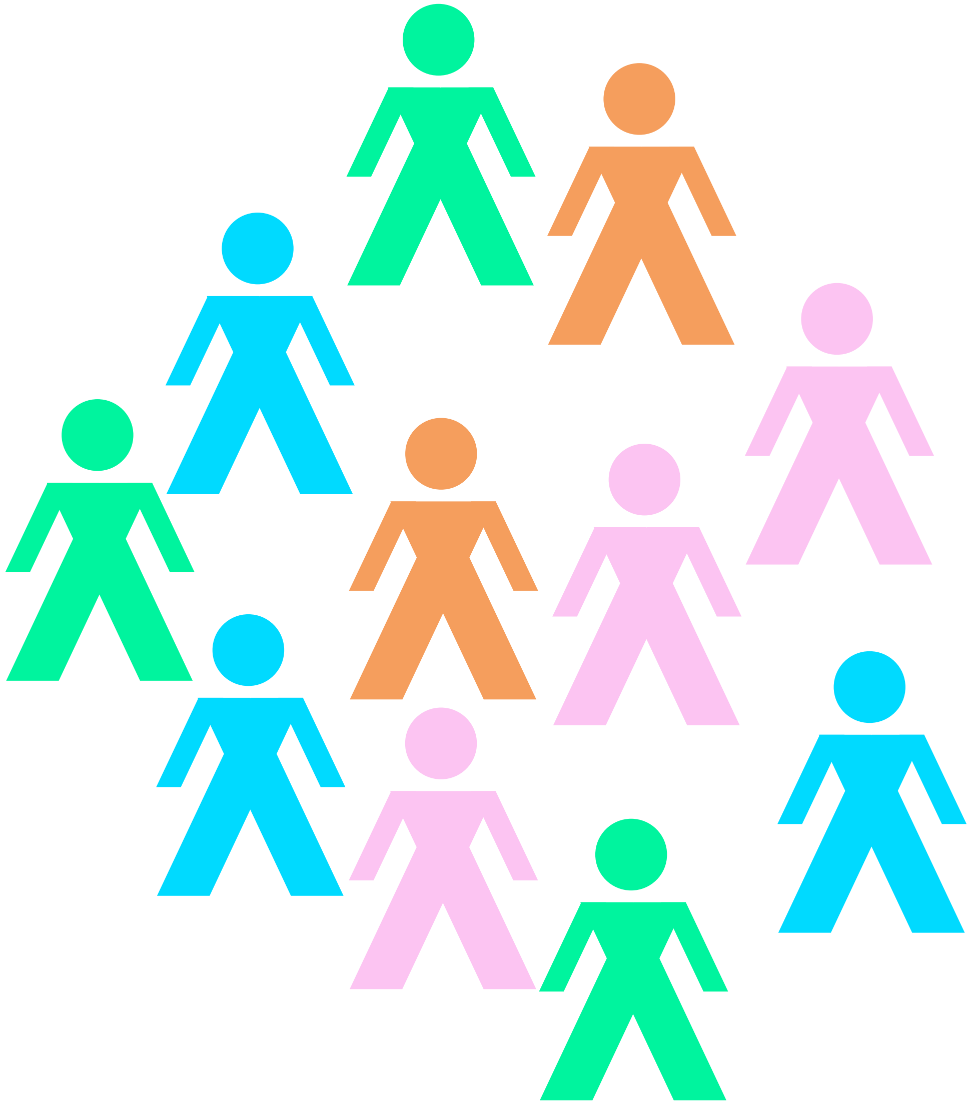


ILLUSTRATION 1: COLOURFUL PEOPLE WEARING HATS

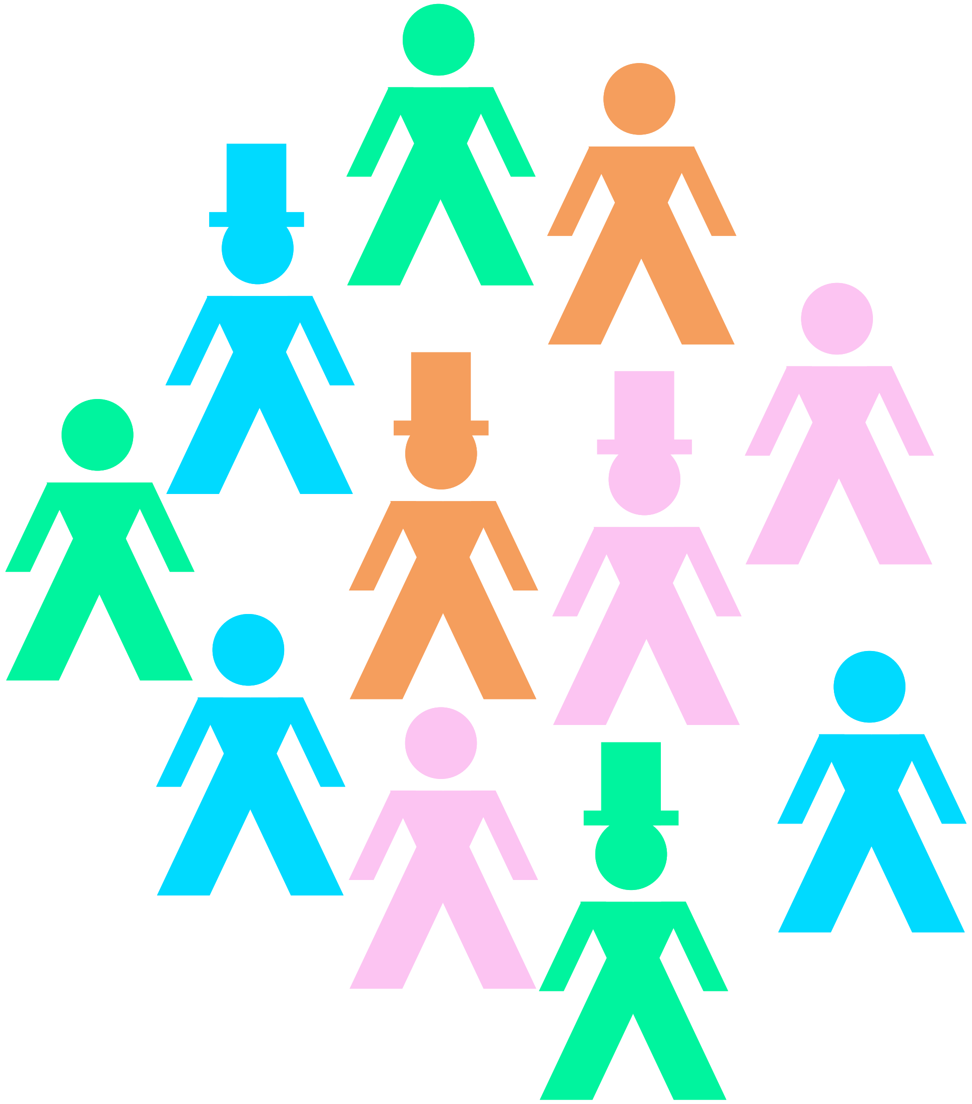
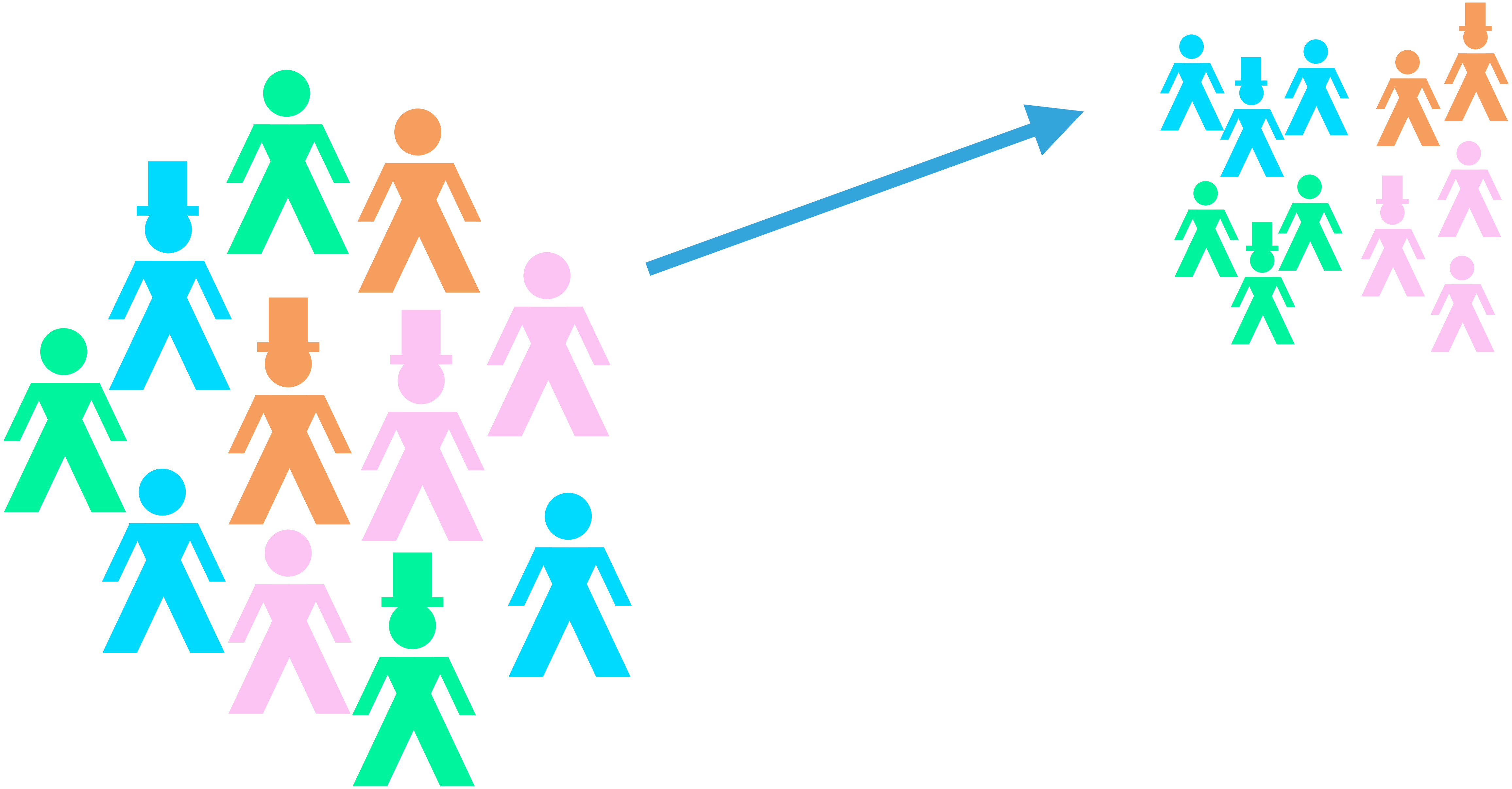
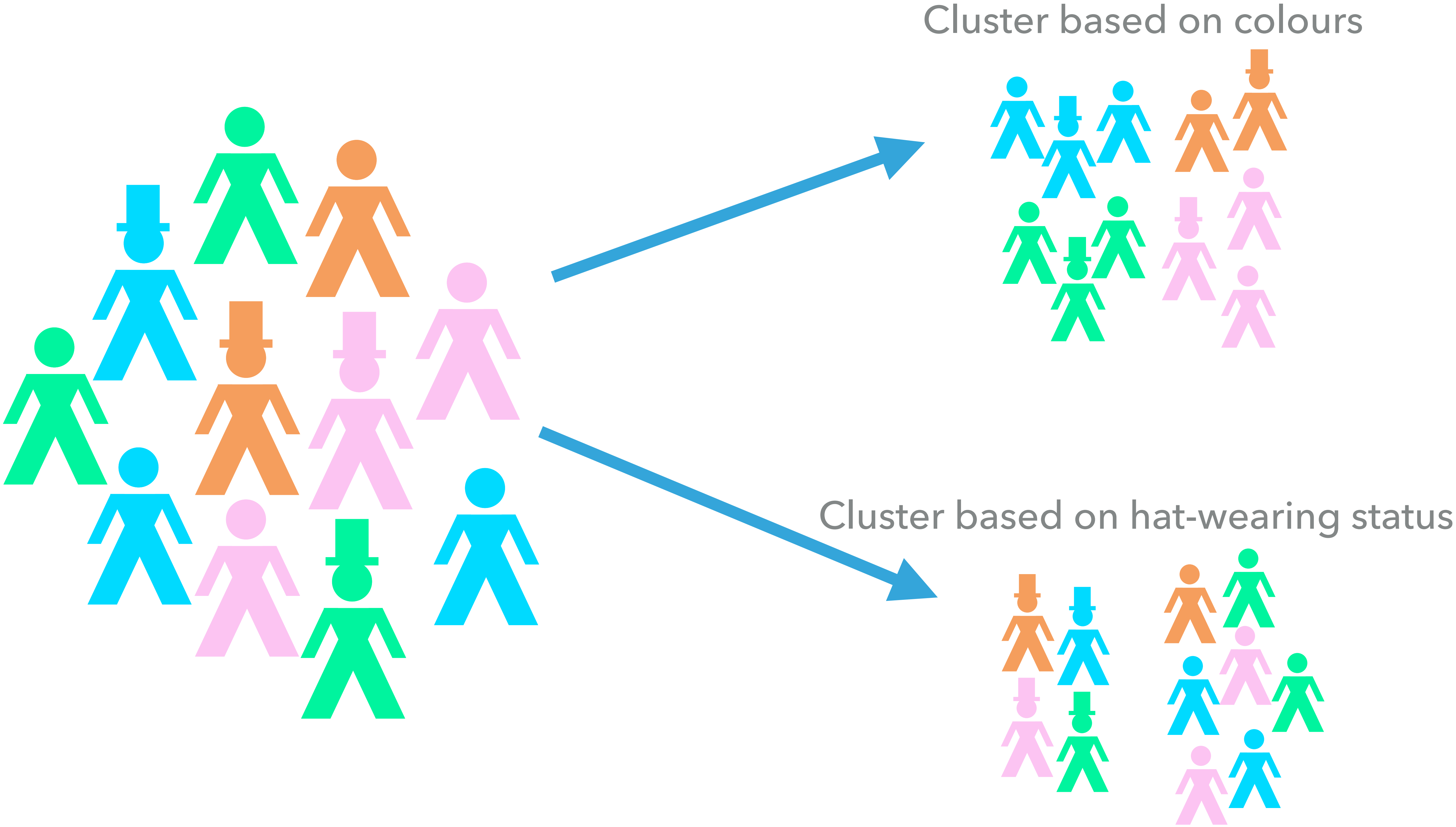


ILLUSTRATION 1: COLOURFUL PEOPLE WEARING HATS



Cluster based on colours

ILLUSTRATION 1: COLOURFUL PEOPLE WEARING HATS



Cluster based on colours

Cluster based on hat-wearing status

ILLUSTRATION 1: COLOURFUL PEOPLE WEARING HATS

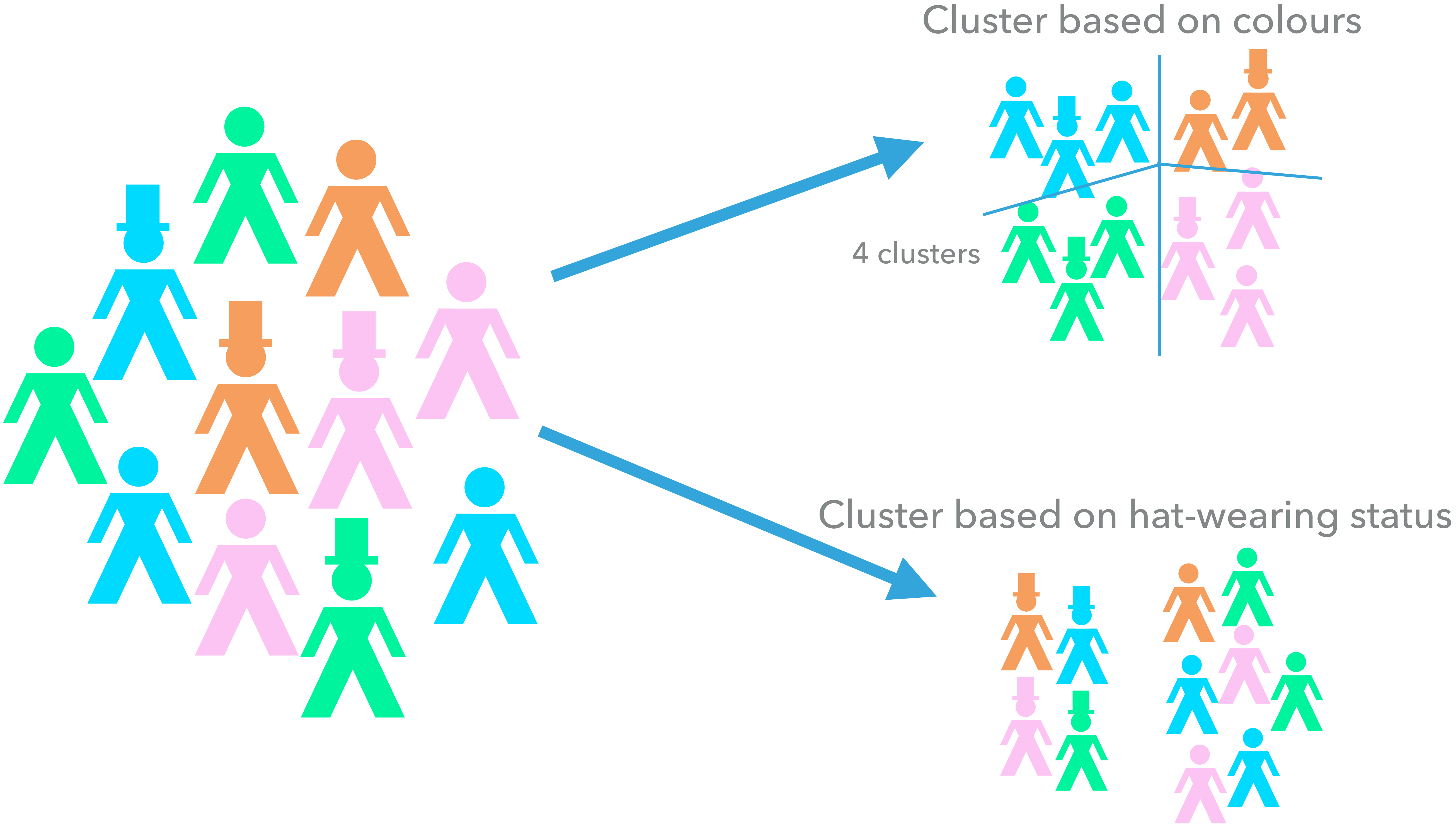
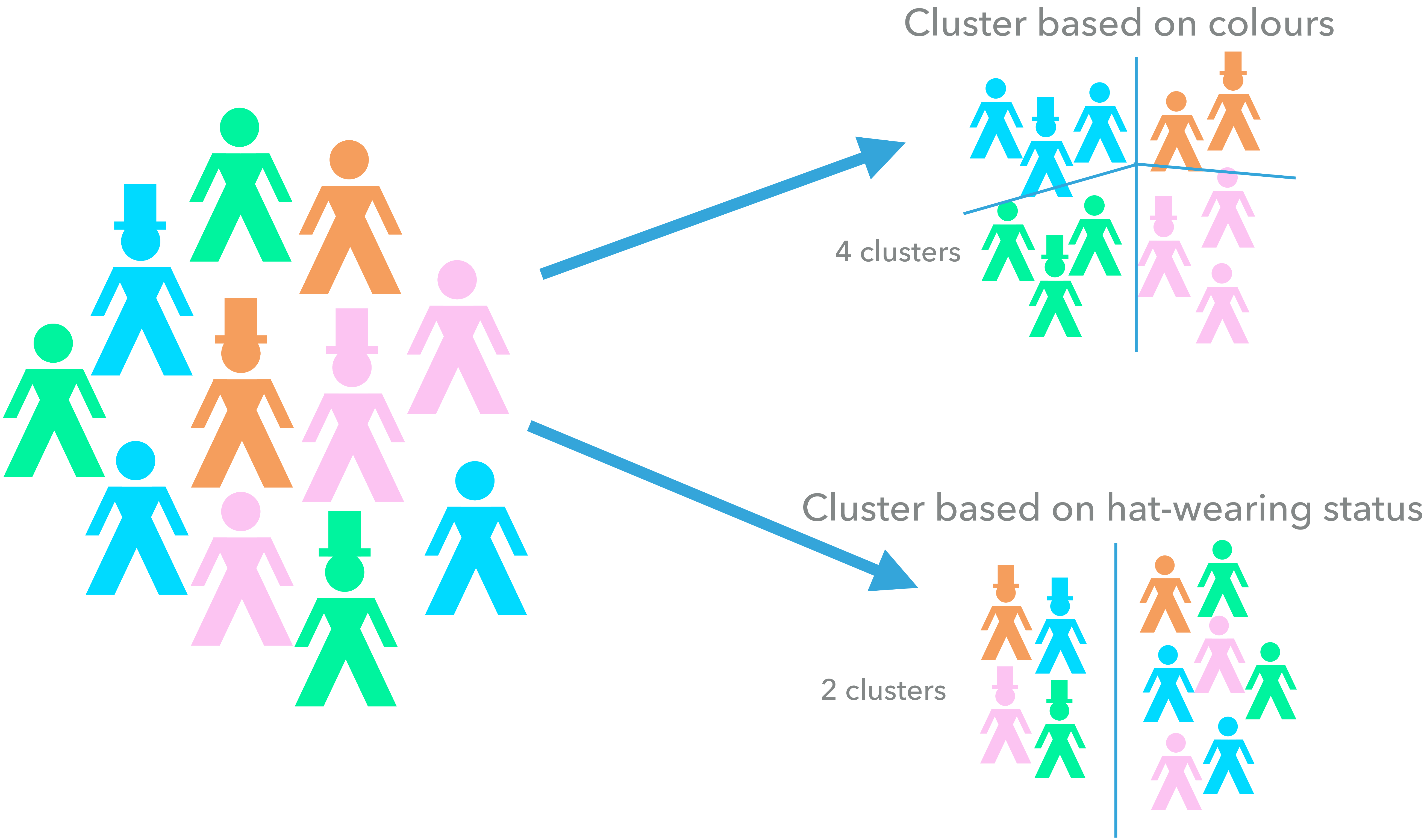


ILLUSTRATION 1: COLOURFUL PEOPLE WEARING HATS



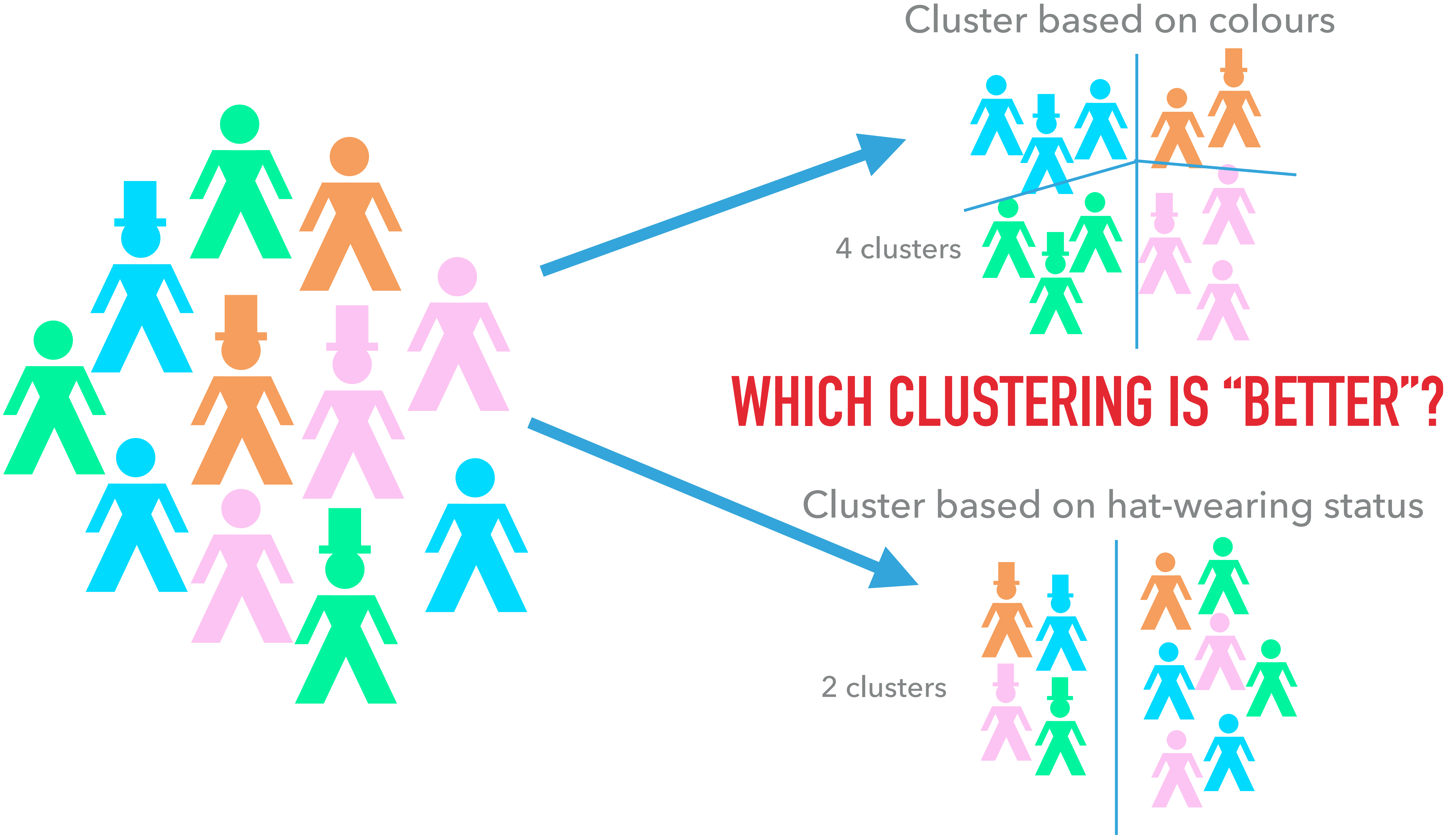
Cluster based on colours

4 clusters

Cluster based on hat-wearing status

2 clusters

ILLUSTRATION 1: COLOURFUL PEOPLE WEARING HATS



Cluster based on colours

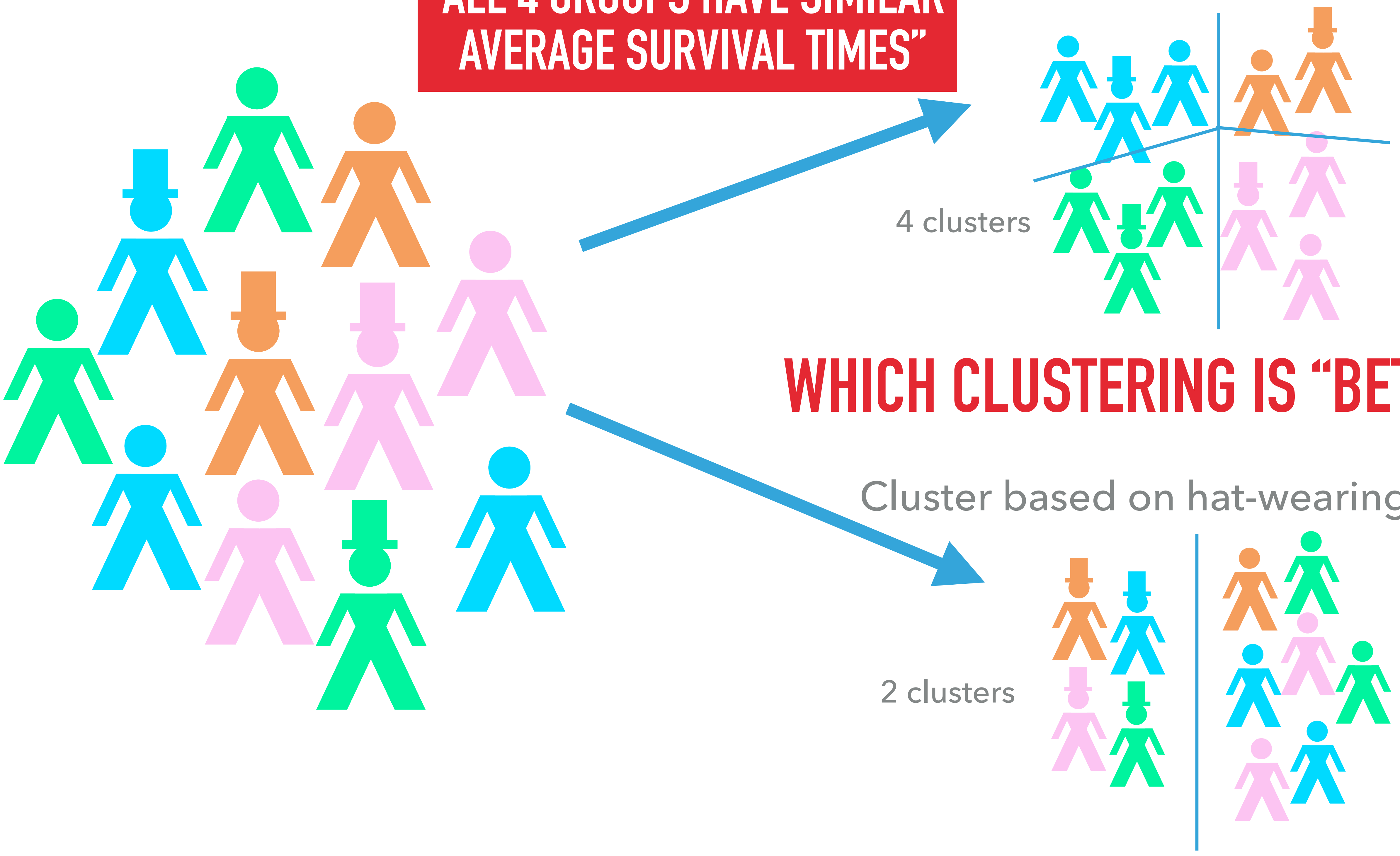
4 clusters

WHICH CLUSTERING IS “BETTER”?

Cluster based on hat-wearing status

2 clusters

ILLUSTRATION 1: COLOURFUL PEOPLE WEARING HATS



"ALL 4 GROUPS HAVE SIMILAR AVERAGE SURVIVAL TIMES"

Cluster based on colours

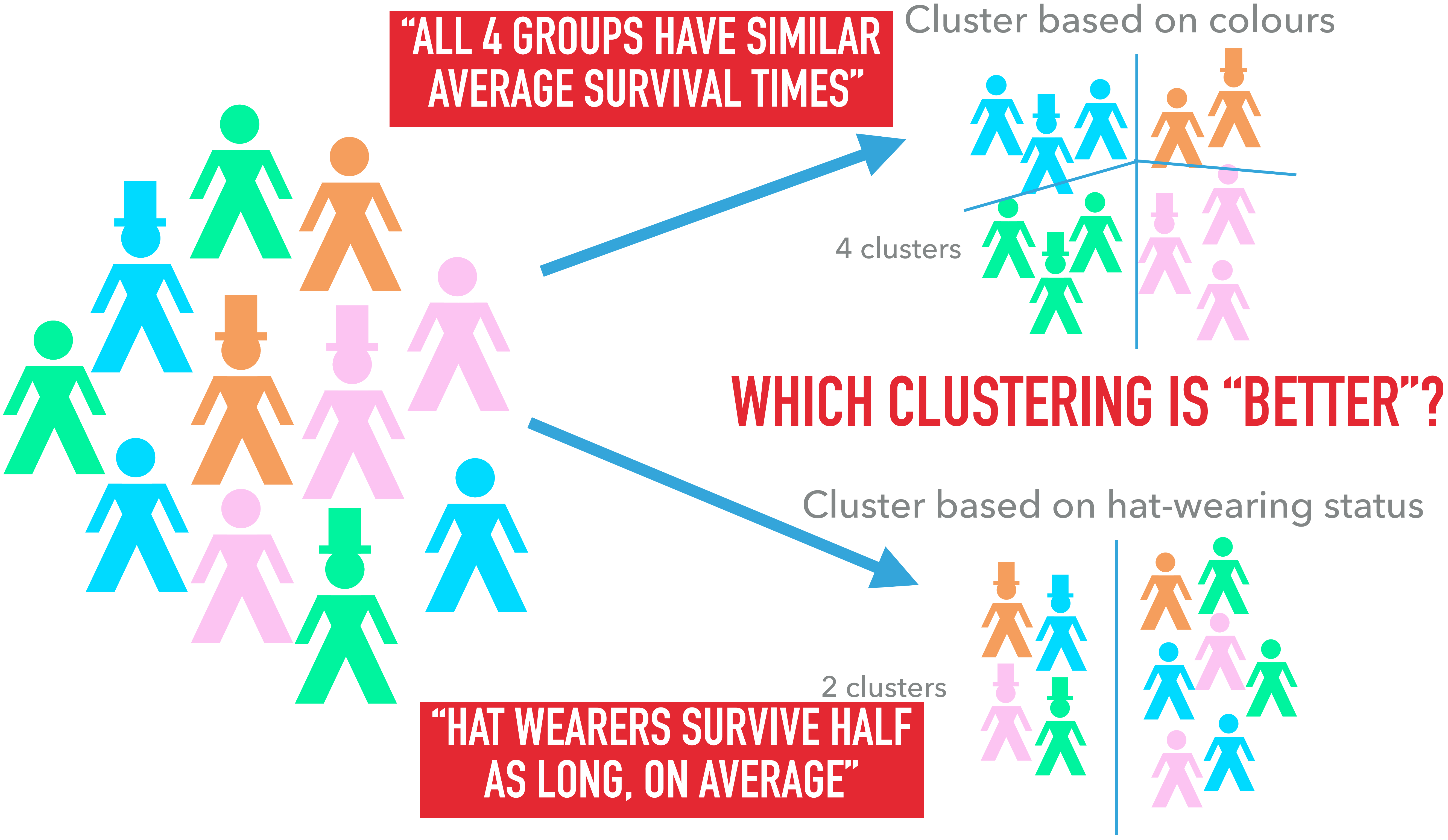
4 clusters

WHICH CLUSTERING IS "BETTER"?

Cluster based on hat-wearing status

2 clusters

ILLUSTRATION 1: COLOURFUL PEOPLE WEARING HATS



“ALL 4 GROUPS HAVE SIMILAR AVERAGE SURVIVAL TIMES”

Cluster based on colours

4 clusters

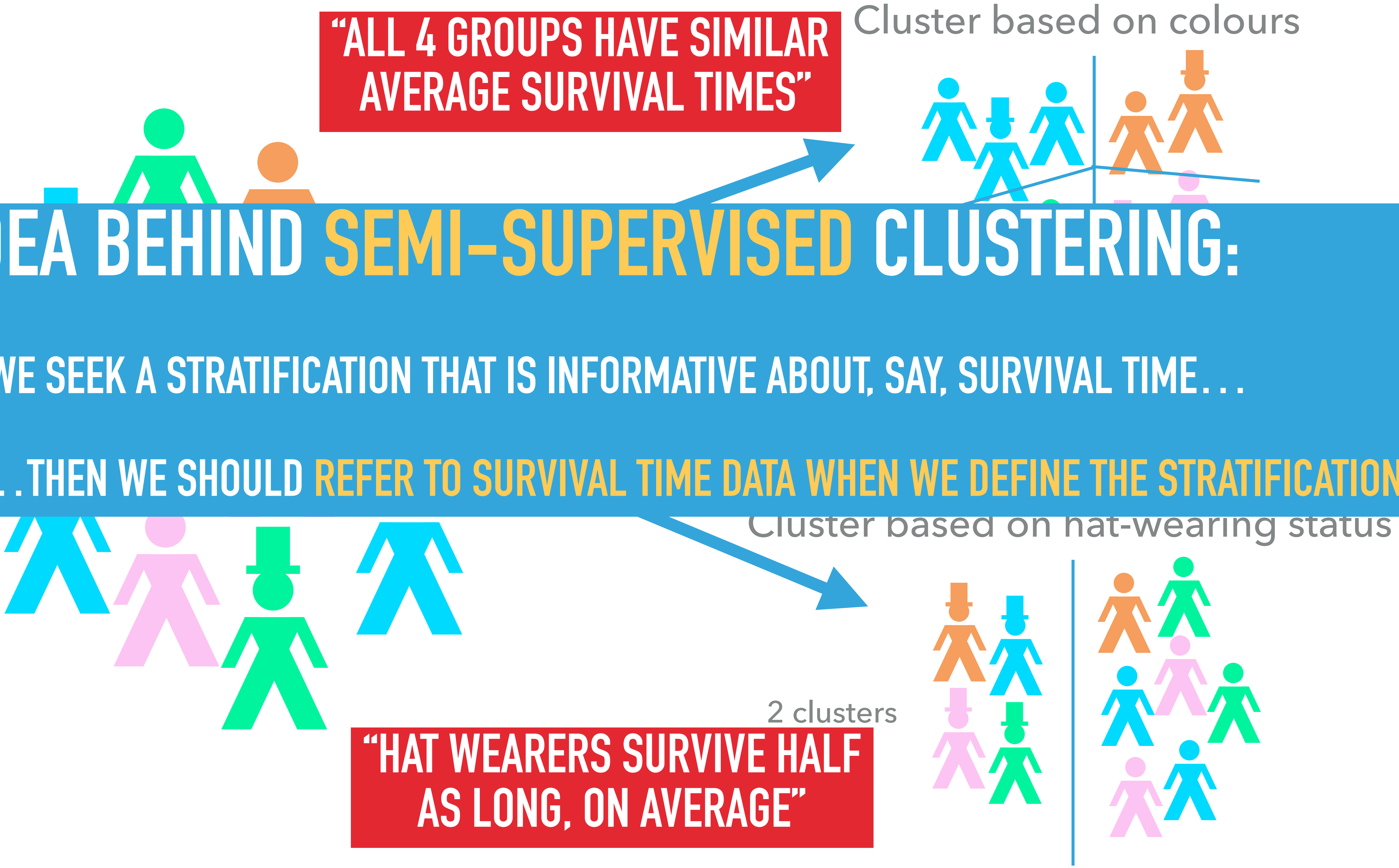
WHICH CLUSTERING IS “BETTER”?

Cluster based on hat-wearing status

2 clusters

“HAT WEARERS SURVIVE HALF AS LONG, ON AVERAGE”

ILLUSTRATION 1: COLOURFUL PEOPLE WEARING HATS



"ALL 4 GROUPS HAVE SIMILAR AVERAGE SURVIVAL TIMES"

Cluster based on colours

Cluster based on hat-wearing status

2 clusters

"HAT WEARERS SURVIVE HALF AS LONG, ON AVERAGE"

IDEA BEHIND SEMI-SUPERVISED CLUSTERING:

IF WE SEEK A STRATIFICATION THAT IS INFORMATIVE ABOUT, SAY, SURVIVAL TIME…

…THEN WE SHOULD REFER TO SURVIVAL TIME DATA WHEN WE DEFINE THE STRATIFICATION.

IDEA BEHIND SEMI-SUPERVISED CLUSTERING:

IF WE SEEK A STRATIFICATION THAT IS INFORMATIVE ABOUT, SAY, SURVIVAL TIME…

…THEN WE SHOULD REFER TO SURVIVAL TIME DATA WHEN WE DEFINE THE STRATIFICATION.

ILLUSTRATION 2: CONTINUOUS DATA

- Why we need additional information

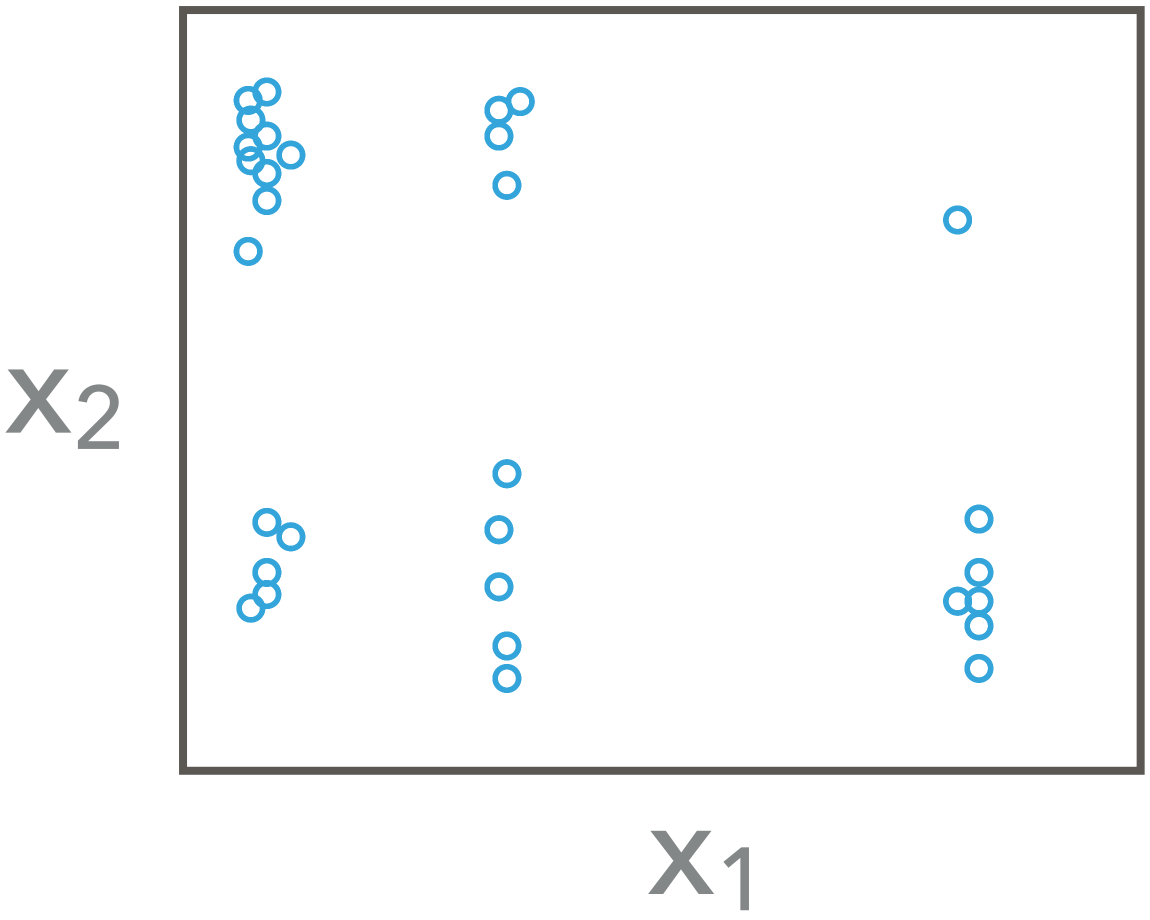


ILLUSTRATION 2: CONTINUOUS DATA

- Why we need additional information
  - First cluster on the basis of both variables

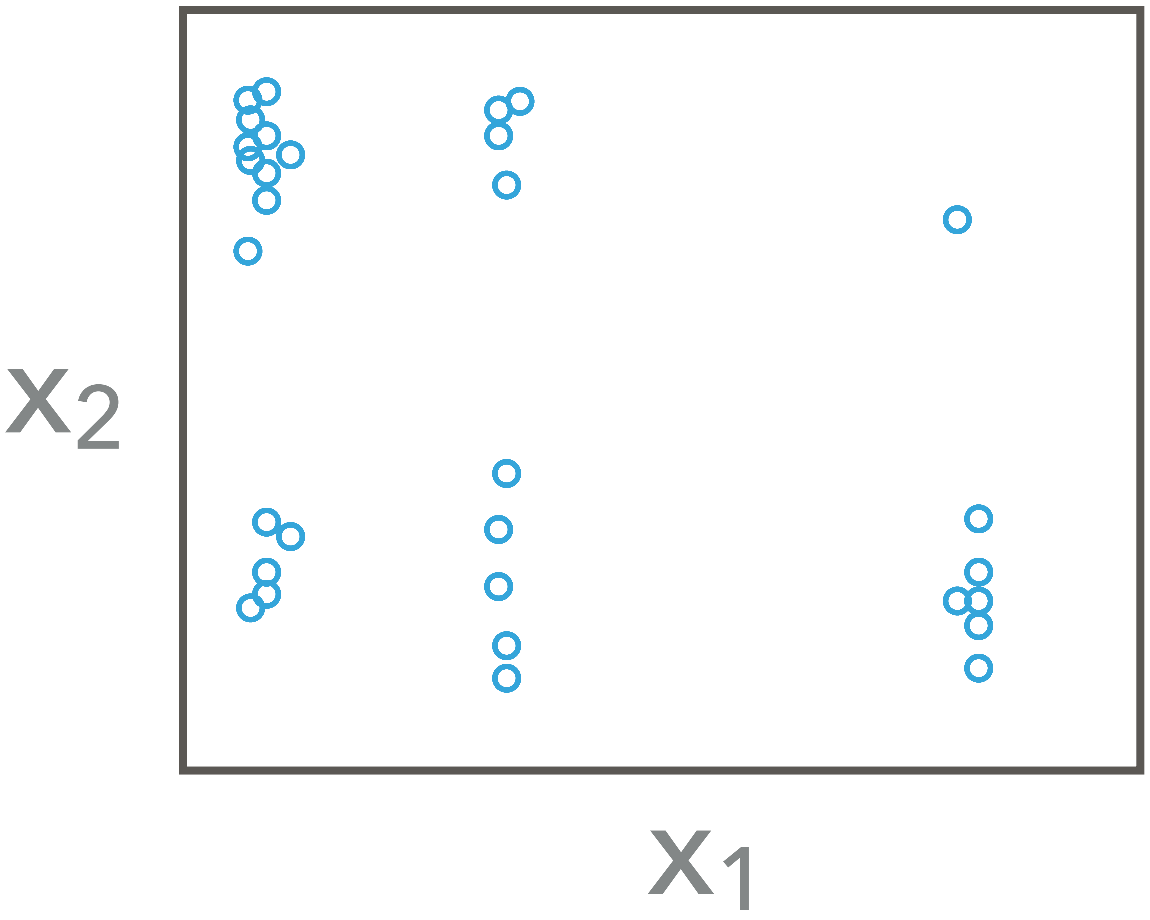


ILLUSTRATION 2: CONTINUOUS DATA

- Why we need additional information
  - First cluster on the basis of both variables

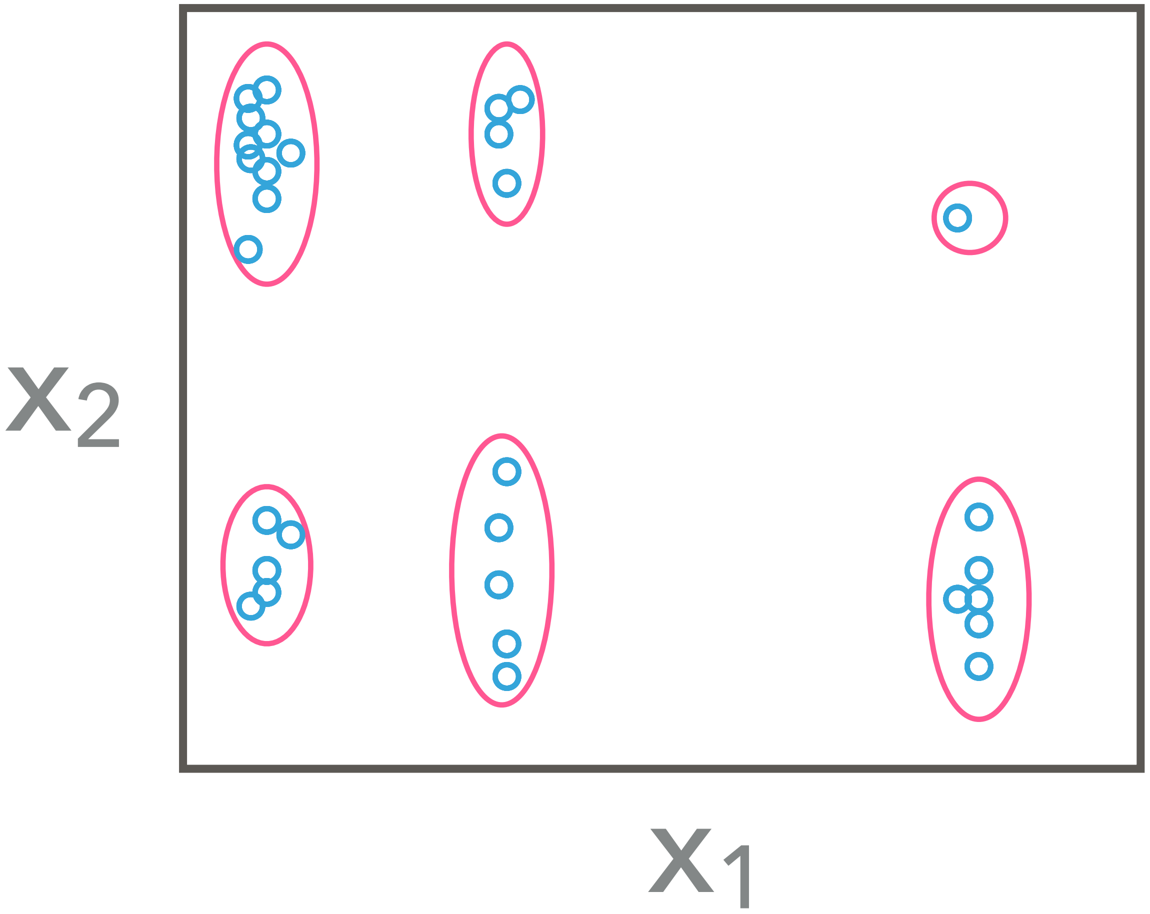


ILLUSTRATION 2: CONTINUOUS DATA

- Why we need additional information
    - First cluster on the basis of both variables
    - Now let’s consider clustering just on the basis of the x_1 variable

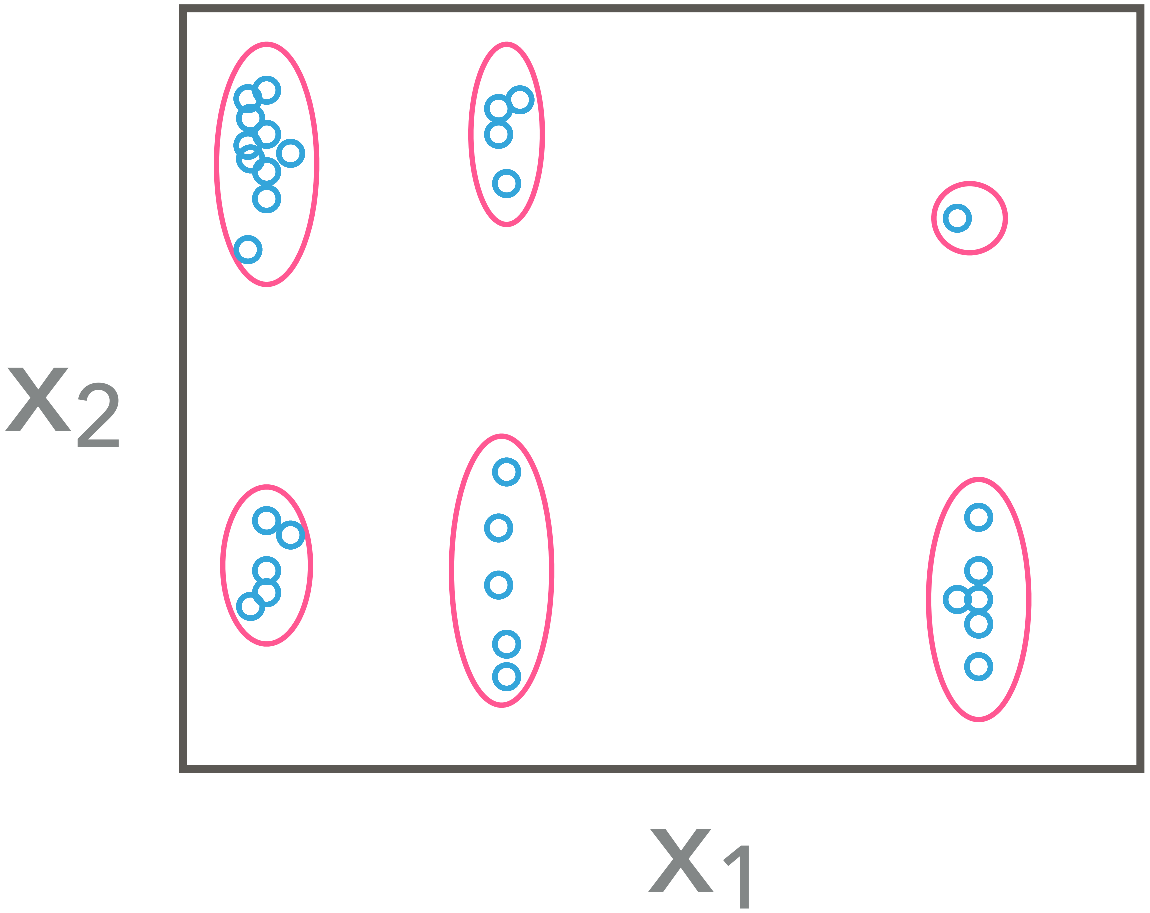


ILLUSTRATION 2: CONTINUOUS DATA

- Why we need additional information
  - First cluster on the basis of both variables
  - Now let's consider clustering just on the basis of the x_1 variable

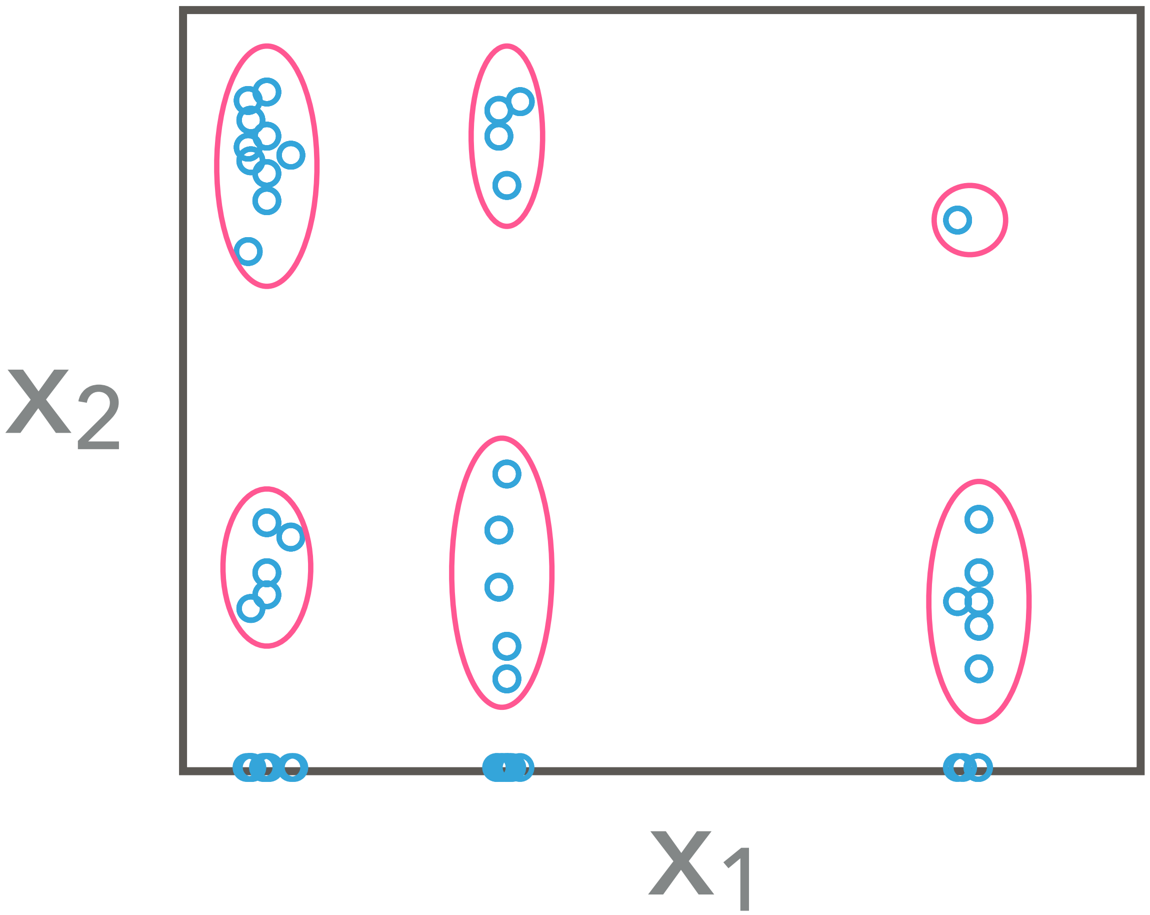


ILLUSTRATION 2: CONTINUOUS DATA

- Why we need additional information
  - First cluster on the basis of both variables
  - Now let's consider clustering just on the basis of the x_1 variable

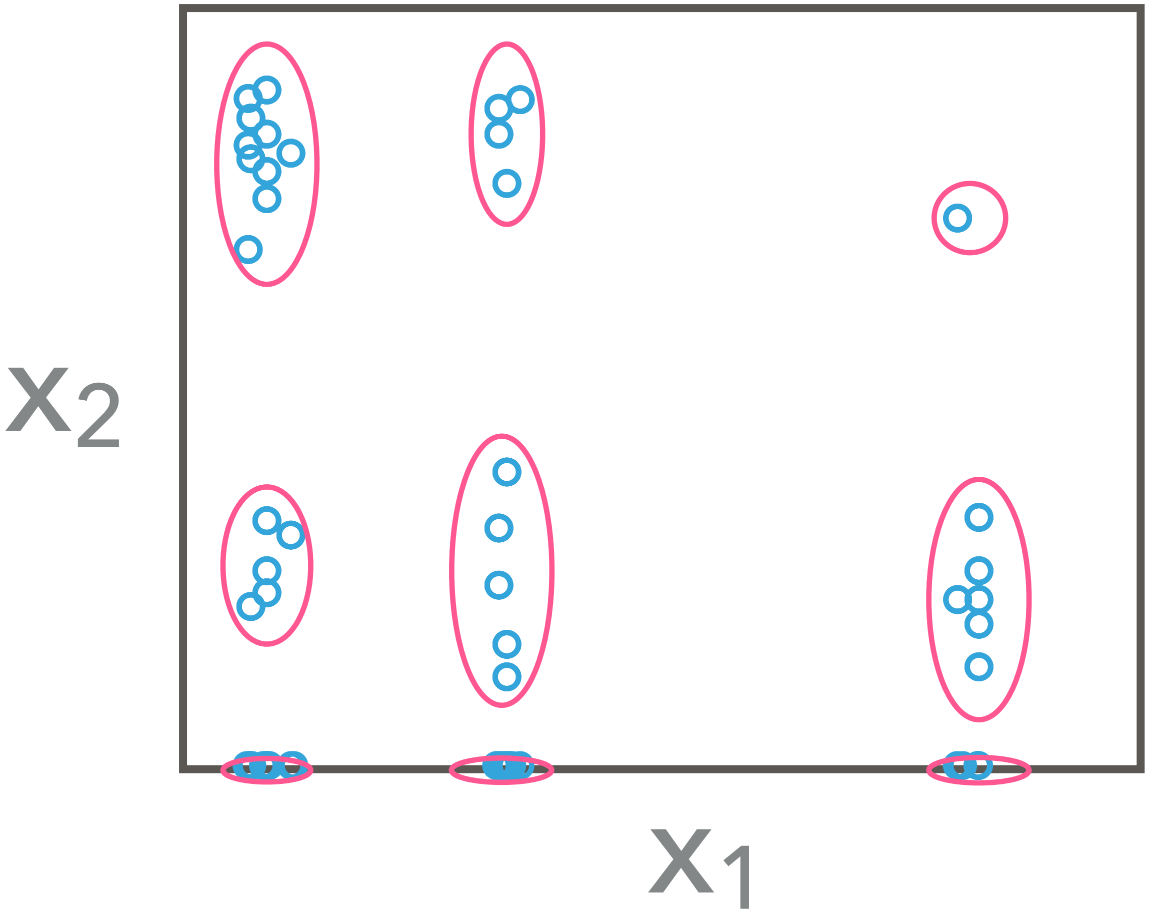
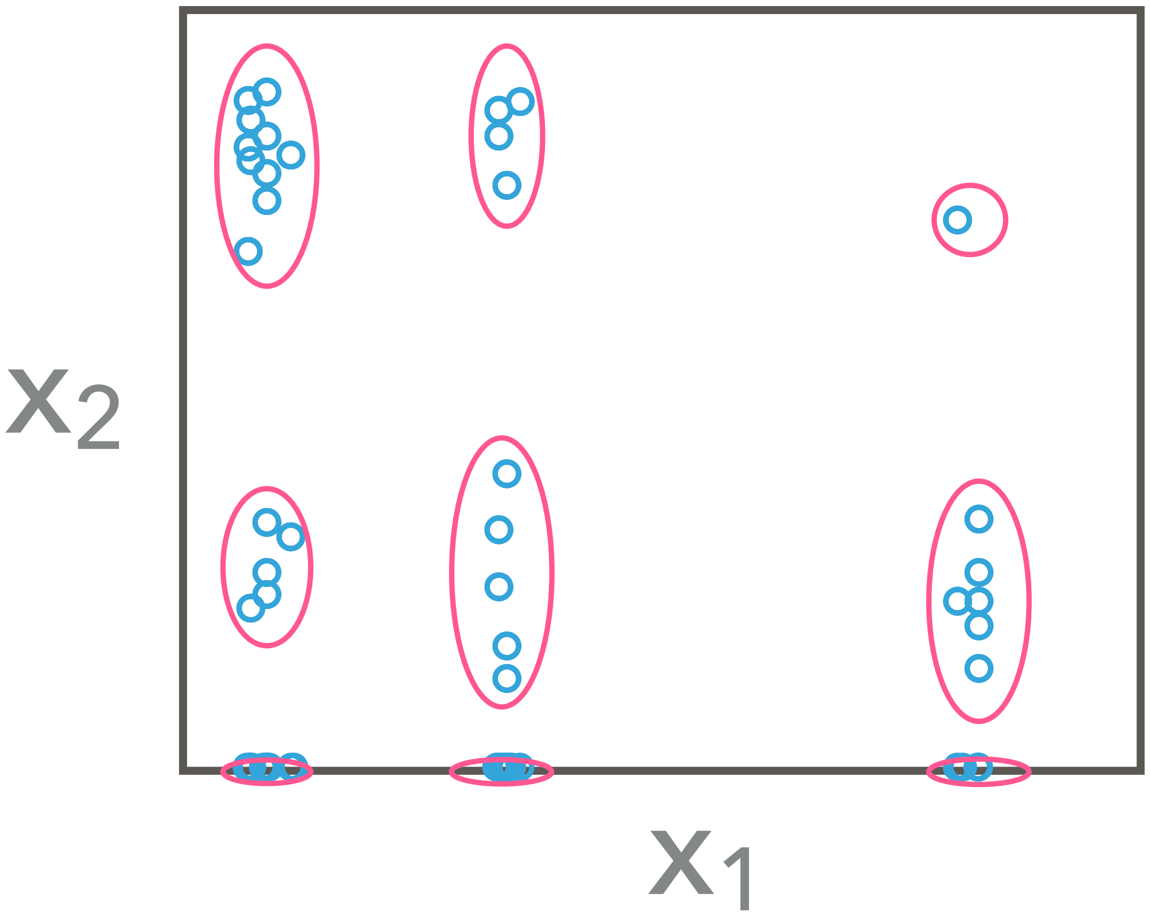


ILLUSTRATION 2: CONTINUOUS DATA

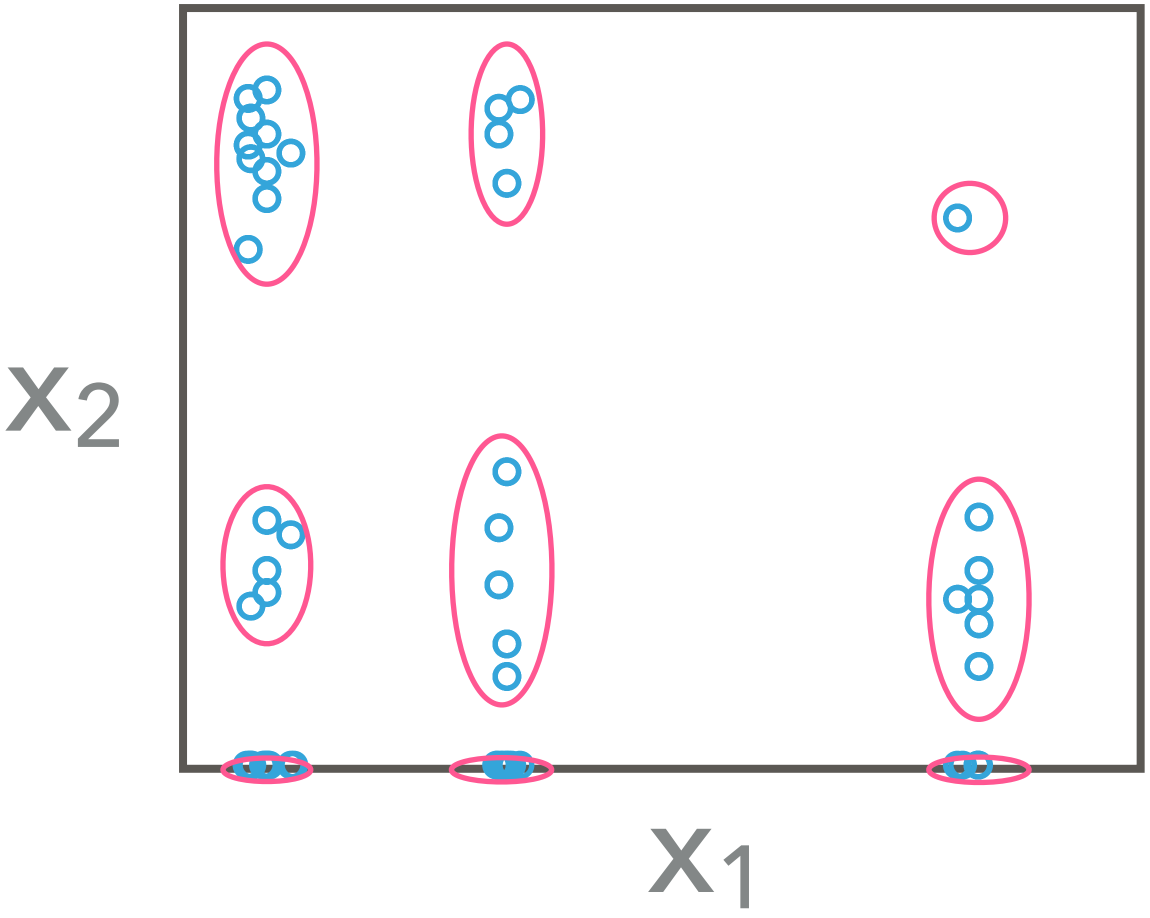
- Why we need additional information
  - First cluster on the basis of both variables
  - Now let's consider clustering just on the basis of the x_1 variable



Clustering just on the basis of x_1 results in 3 clusters

ILLUSTRATION 2: CONTINUOUS DATA

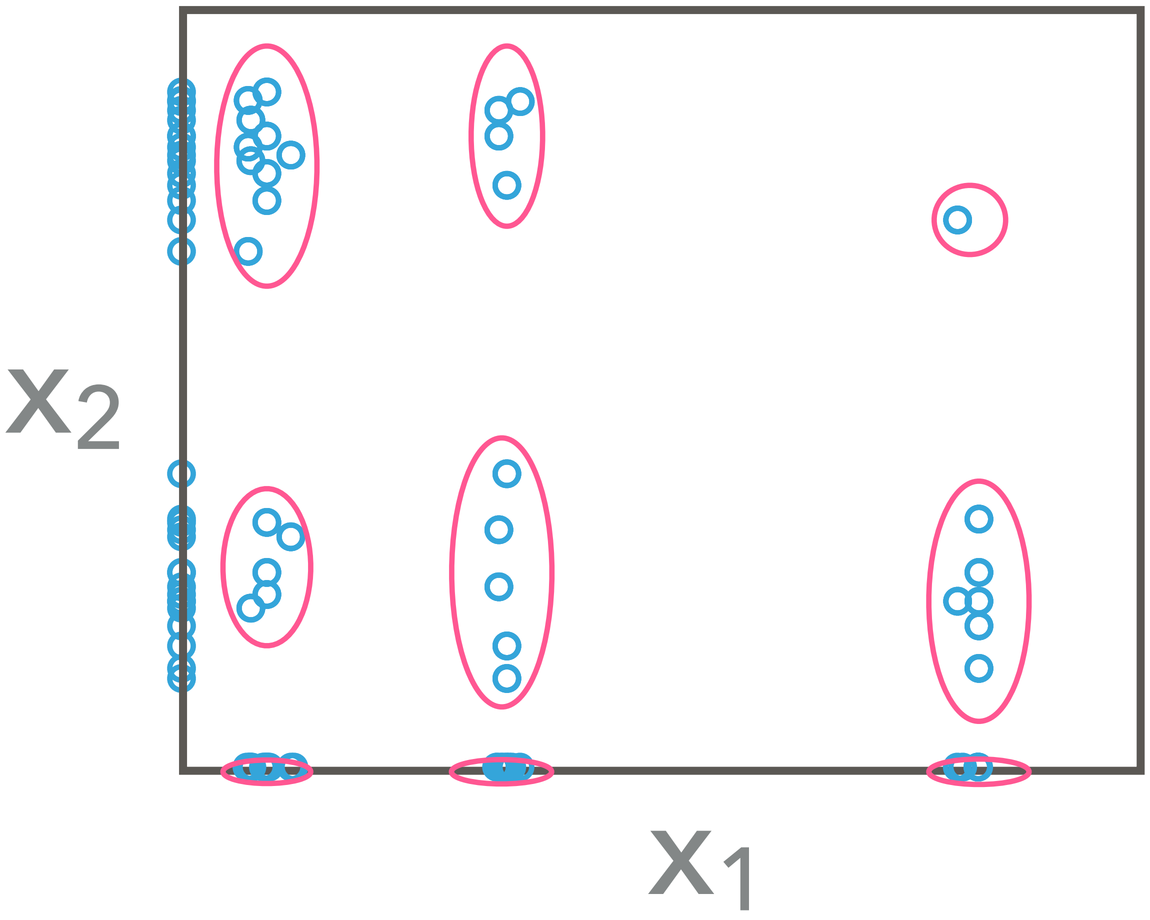
- Why we need additional information
  - First cluster on the basis of both variables
  - Now let's consider clustering just on the basis of the x_1 variable
  - And now let's do the same for the x_2 variable



Clustering just on the basis of x_1 results in 3 clusters

ILLUSTRATION 2: CONTINUOUS DATA

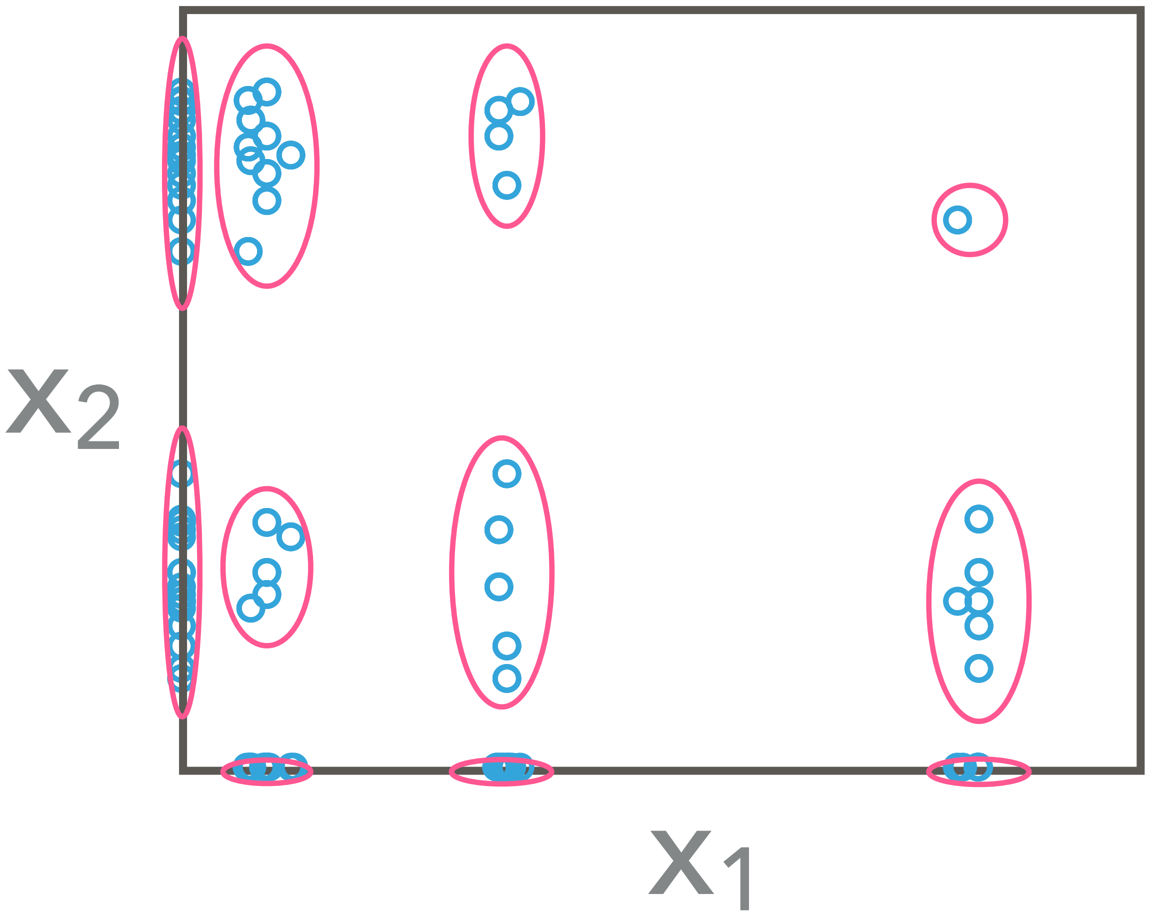
- Why we need additional information
  - First cluster on the basis of both variables
  - Now let's consider clustering just on the basis of the x_1 variable
  - And now let's do the same for the x_2 variable



Clustering just on the basis of x_1 results in 3 clusters

ILLUSTRATION 2: CONTINUOUS DATA

- Why we need additional information
  - First cluster on the basis of both variables
  - Now let's consider clustering just on the basis of the x_1 variable
  - And now let's do the same for the x_2 variable

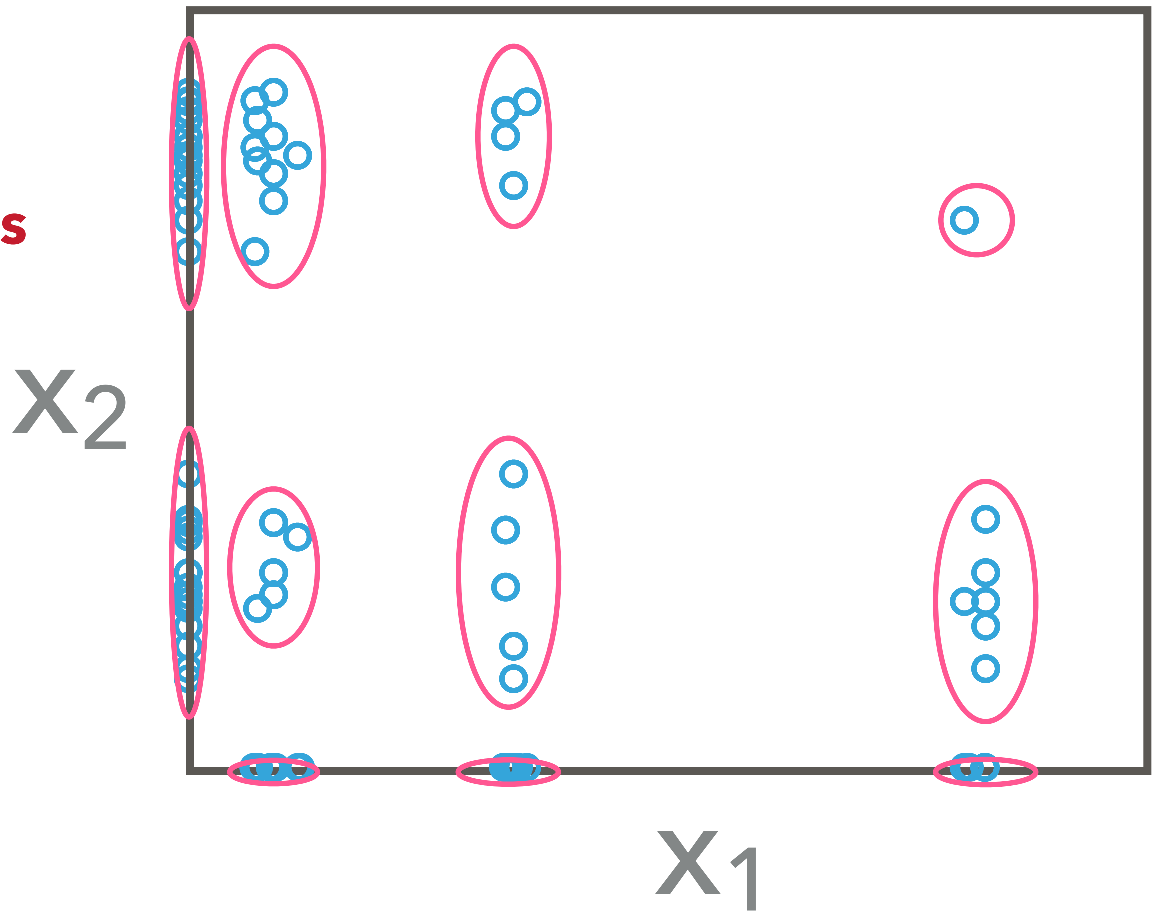


Clustering just on the basis of x_1 results in 3 clusters

ILLUSTRATION 2: CONTINUOUS DATA

- Why we need additional information
  - First cluster on the basis of both variables
  - Now let's consider clustering just on the basis of the x_1 variable
  - And now let's do the same for the x_2 variable

Clustering just on the basis of x_2 results in 2 clusters

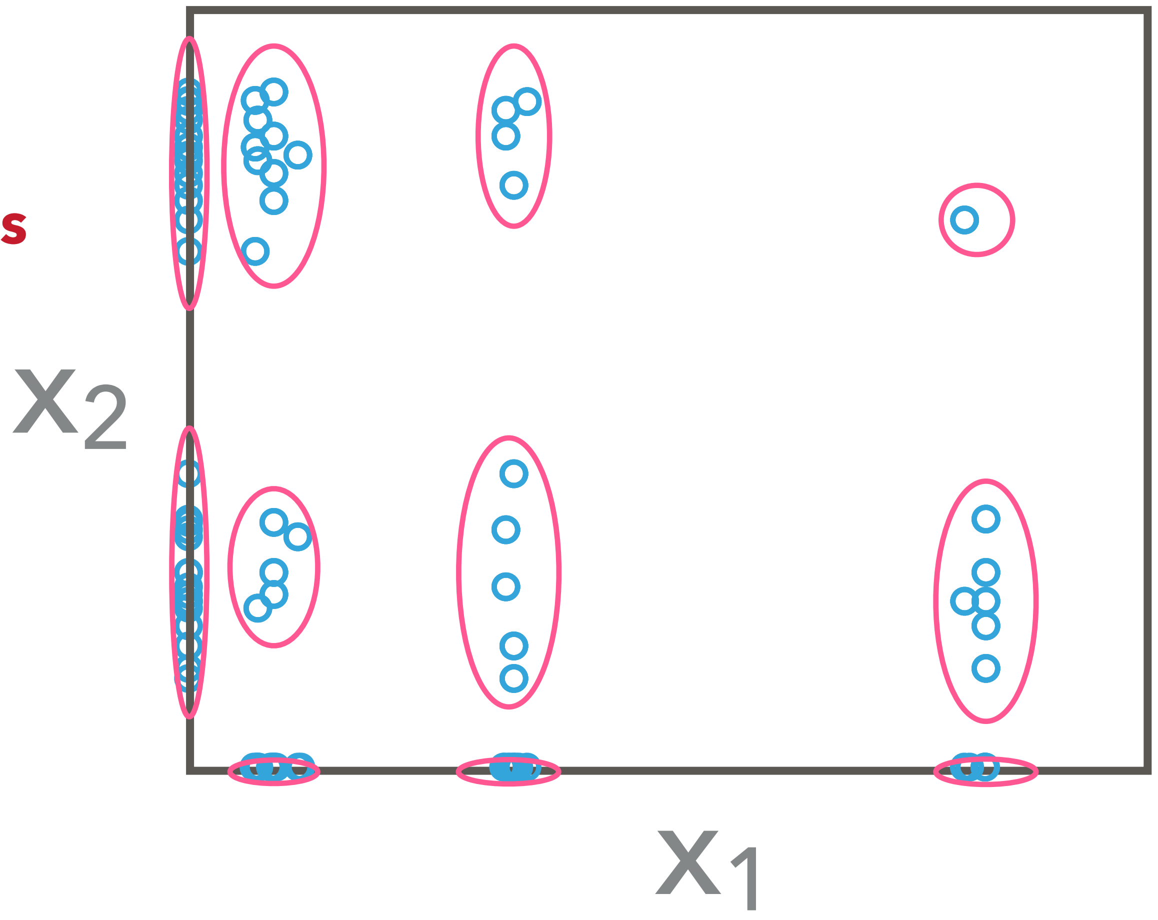


Clustering just on the basis of x_1 results in 3 clusters

ILLUSTRATION 2: CONTINUOUS DATA

- Why we need additional information
  - First cluster on the basis of both variables
  - Now let's consider clustering just on the basis of the x_1 variable
  - And now let's do the same for the x_2 variable

Clustering just on the basis of x_2 results in 2 clusters



Depending on which variable we select, we end up with different clustering results

Clustering just on the basis of x_1 results in 3 clusters

- Should we select x_1, x_2, or both?

- Should we select x_1, x_2, or both?
- Should we be targeting the 2, 3, or 6 cluster structure?

- Should we select x_1, x_2, or both?
- Should we be targeting the 2, 3, or 6 cluster structure?
- Which is the “relevant” clustering structure?

- Should we select x_1, x_2, or both?
- Should we be targeting the 2, 3, or 6 cluster structure?
- Which is the “relevant” clustering structure?

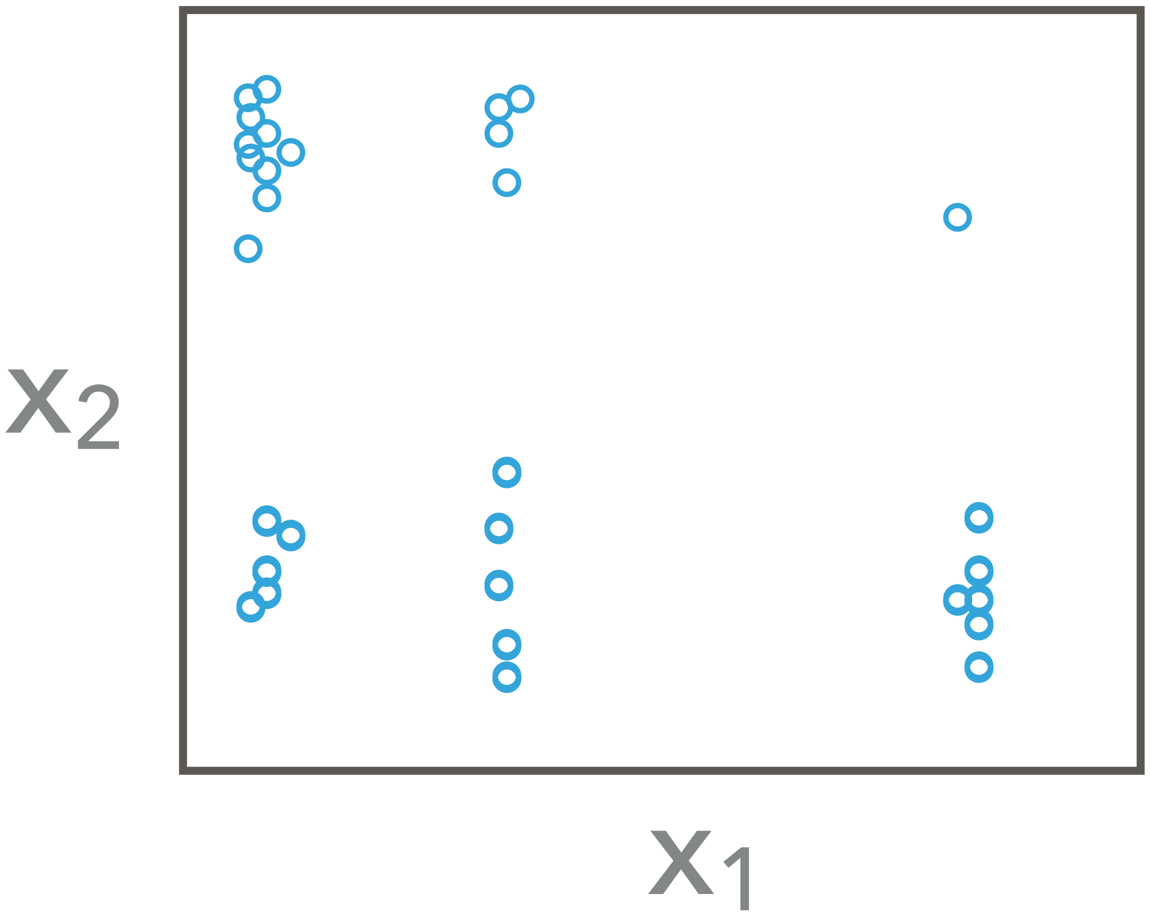
HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS “RELEVANT”?

HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

- Make use of a “response”: left-out variable(s) directly/closely linked to our true interest.
- e.g. x_1 and x_2 are blood protein abundances measured on a number of individuals, but our true interest is in y (high/low risk)

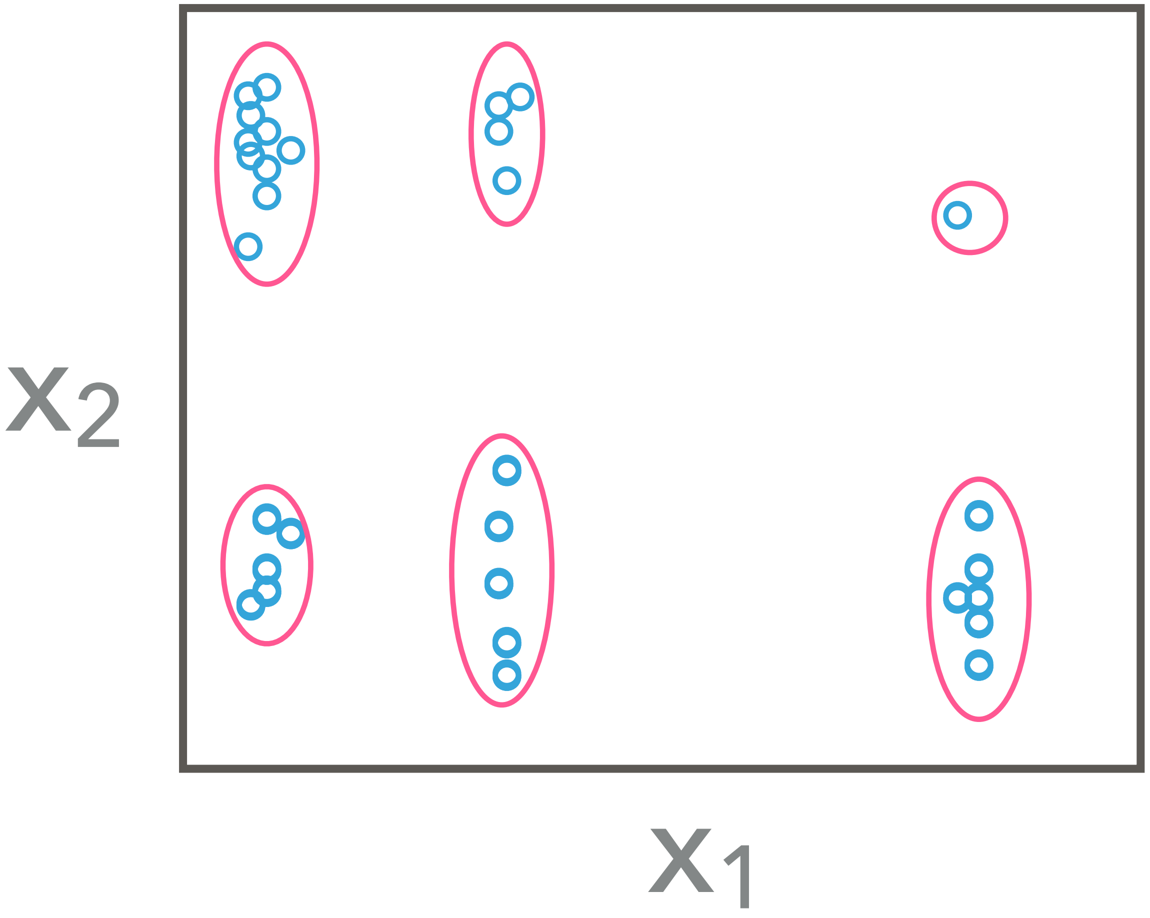
HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

- Make use of a “response”: left-out variable(s) directly/closely linked to our true interest.
- e.g. x_1 and x_2 are blood protein abundances measured on a number of individuals, but our true interest is in y (high/low risk)



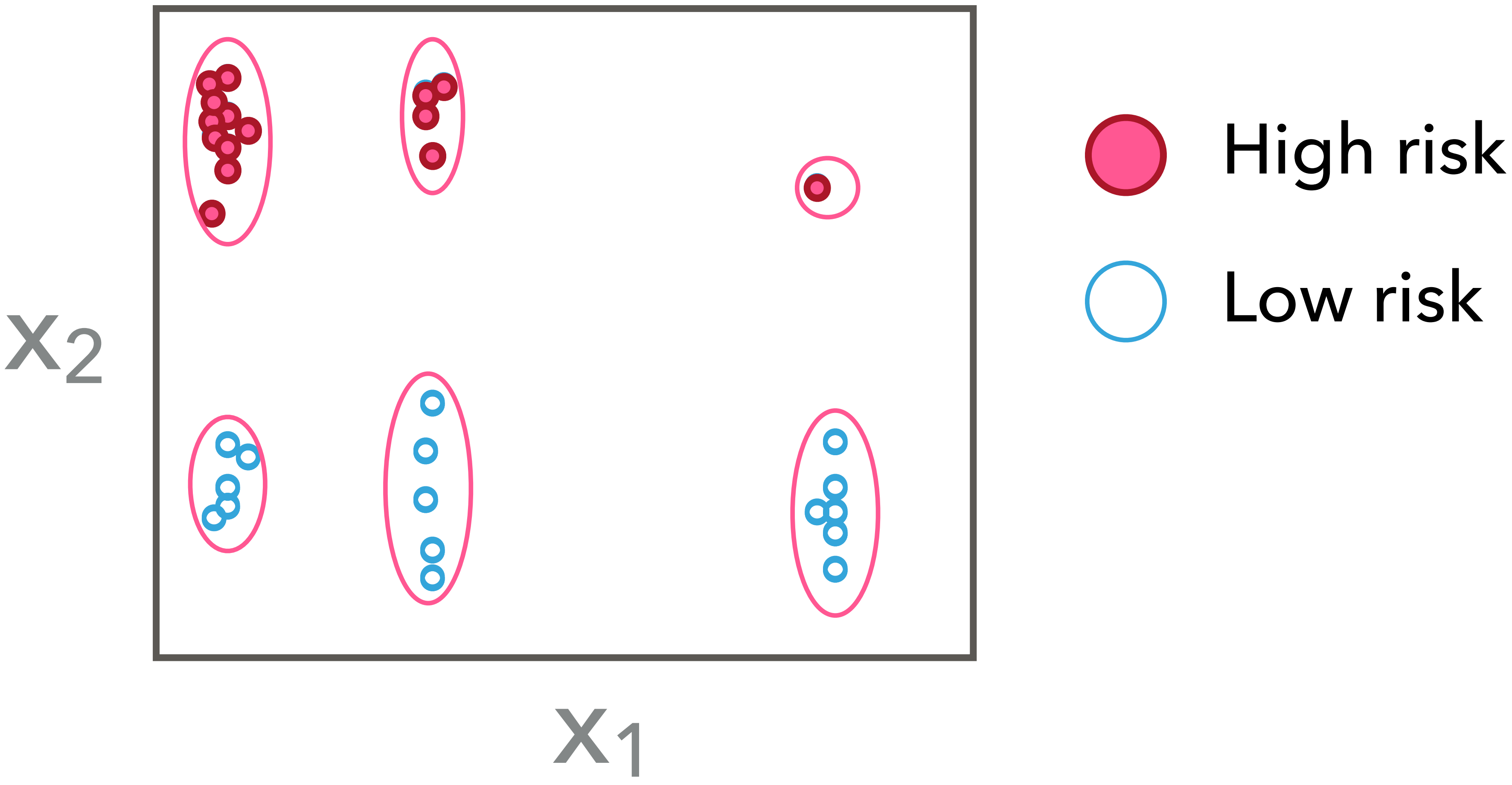
HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

- Make use of a “response”: left-out variable(s) directly/closely linked to our true interest.
- e.g. x_1 and x_2 are blood protein abundances measured on a number of individuals, but our true interest is in y (high/low risk)



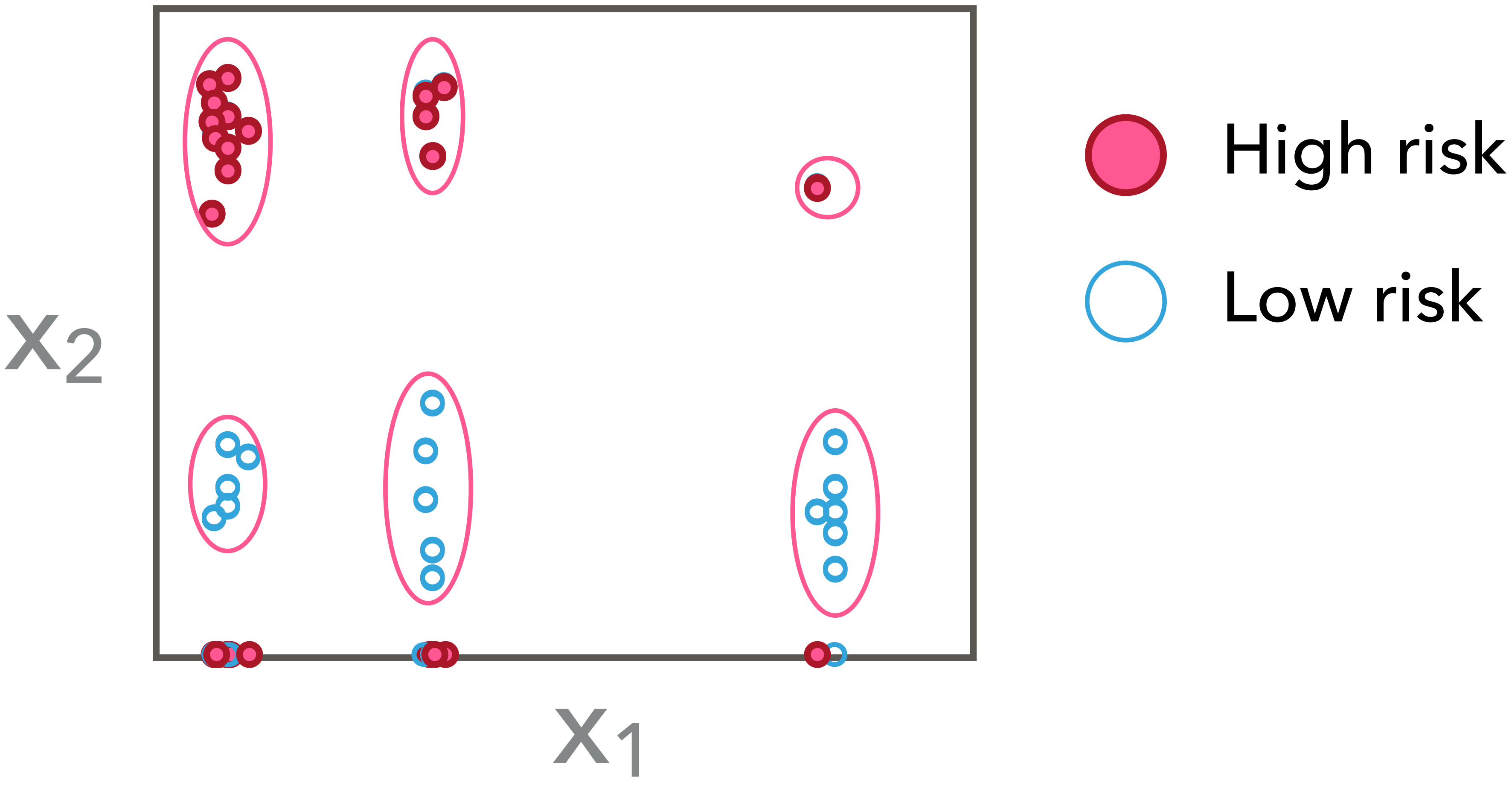
HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

- Make use of a “response”: left-out variable(s) directly/closely linked to our true interest.
- e.g. x_1 and x_2 are blood protein abundances measured on a number of individuals, but our true interest is in y (high/low risk)



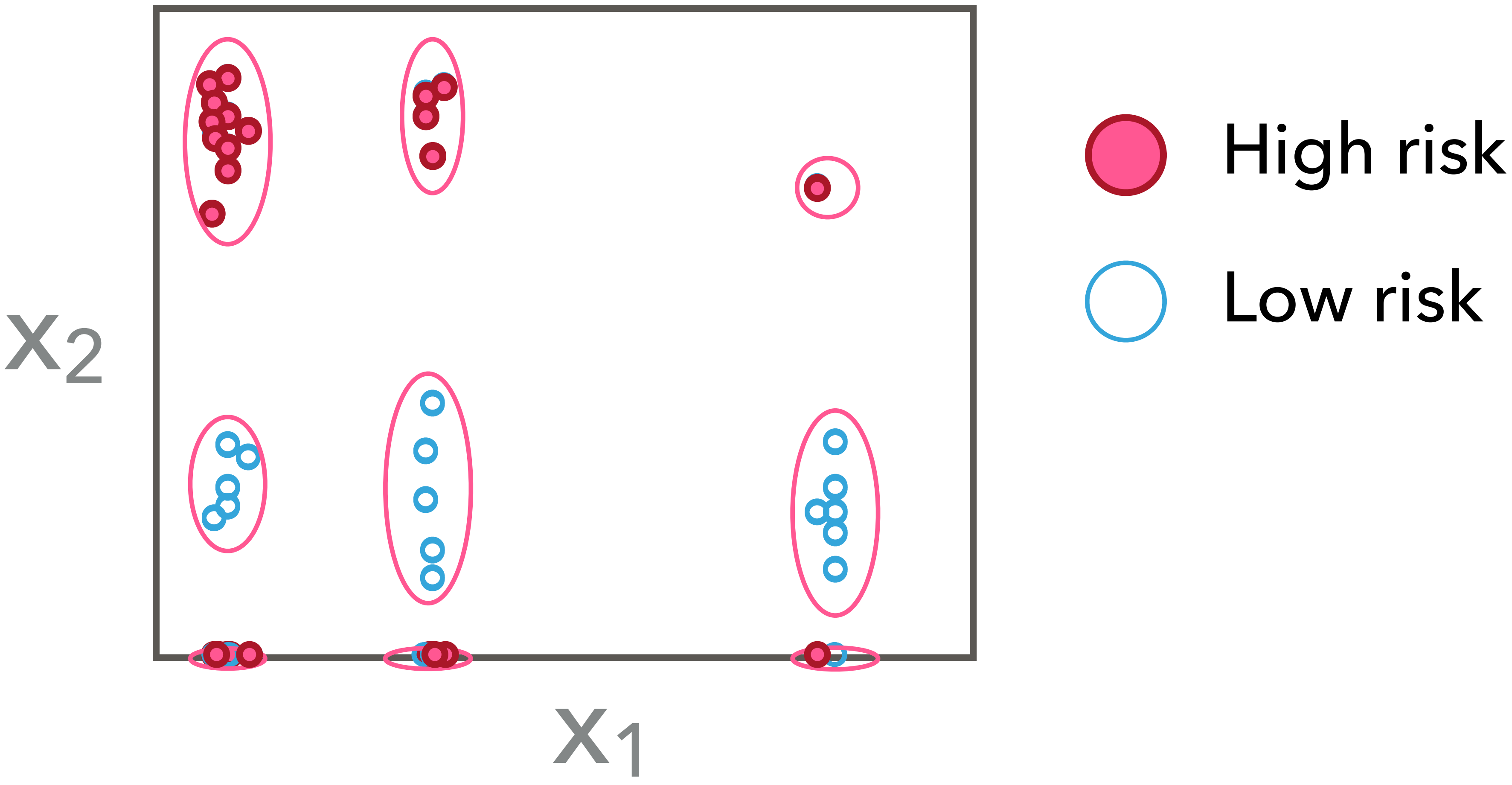
HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

- Make use of a “response”: left-out variable(s) directly/closely linked to our true interest.
- e.g. x_1 and x_2 are blood protein abundances measured on a number of individuals, but our true interest is in y (high/low risk)



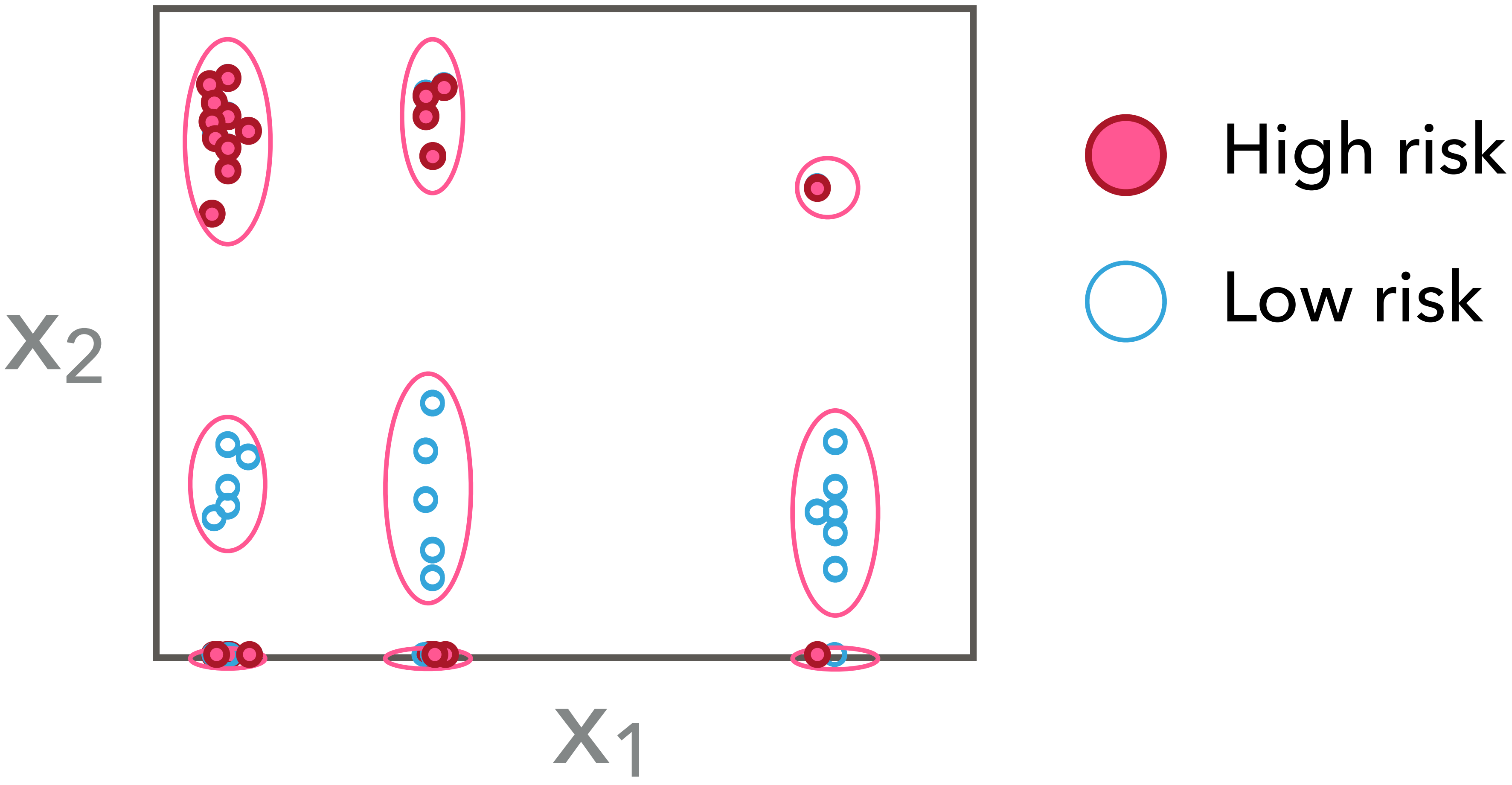
HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

- Make use of a “response”: left-out variable(s) directly/closely linked to our true interest.
- e.g. x_1 and x_2 are blood protein abundances measured on a number of individuals, but our true interest is in y (high/low risk)



HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

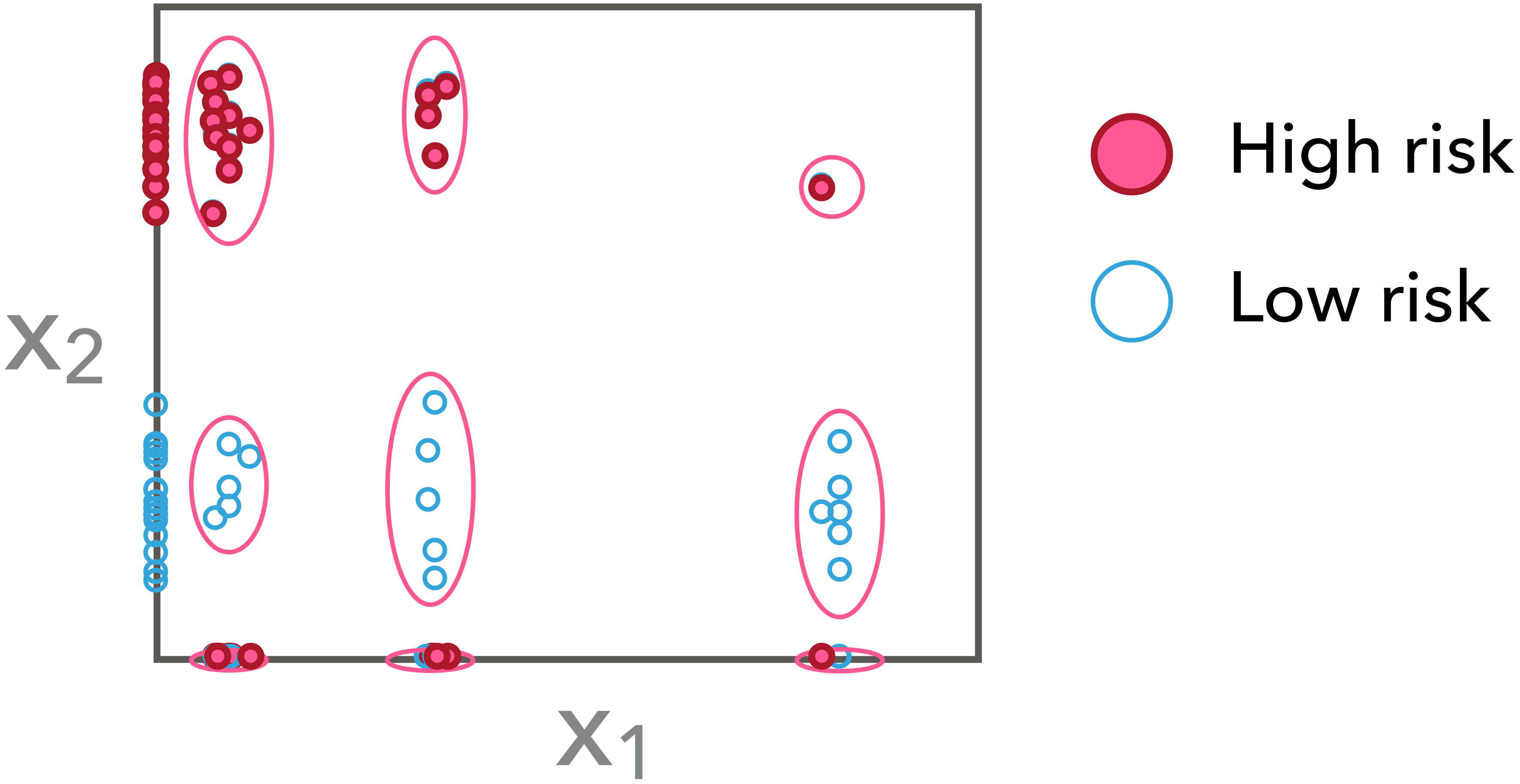
- Make use of a "response": left-out variable(s) directly/closely linked to our true interest.
- e.g. x_1 and x_2 are blood protein abundances measured on a number of individuals, but our true interest is in y (high/low risk)



x_1 possesses a clear clustering structure… but it is irrelevant for this stratification task

HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

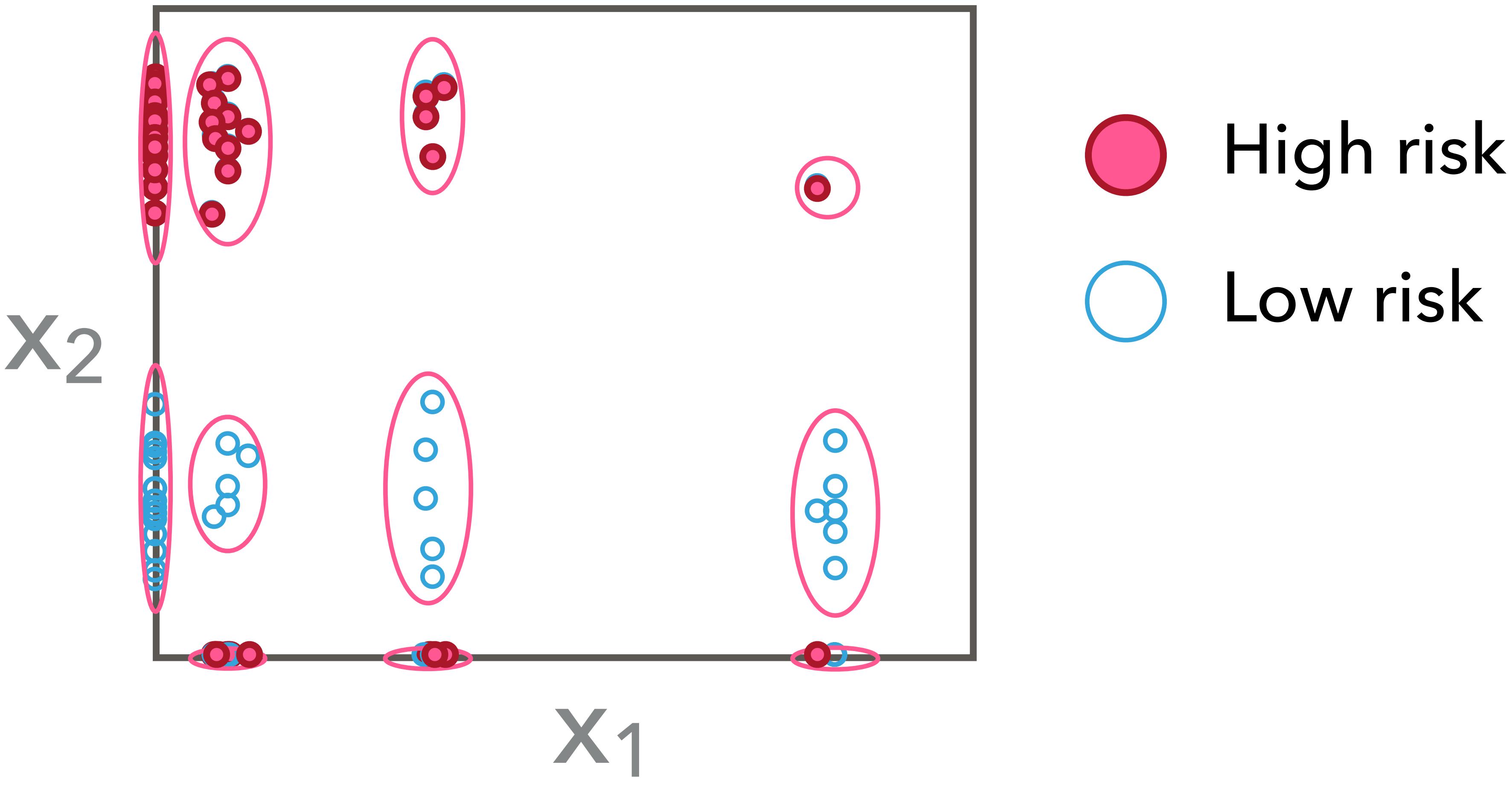
- Make use of a “response”: left-out variable(s) directly/closely linked to our true interest.
- e.g. x_1 and x_2 are blood protein abundances measured on a number of individuals, but our true interest is in y (high/low risk)



x_1 possesses a clear clustering structure… but it is irrelevant for this stratification task

HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

- Make use of a "response": left-out variable(s) directly/closely linked to our true interest.
- e.g. x_1 and x_2 are blood protein abundances measured on a number of individuals, but our true interest is in y (high/low risk)

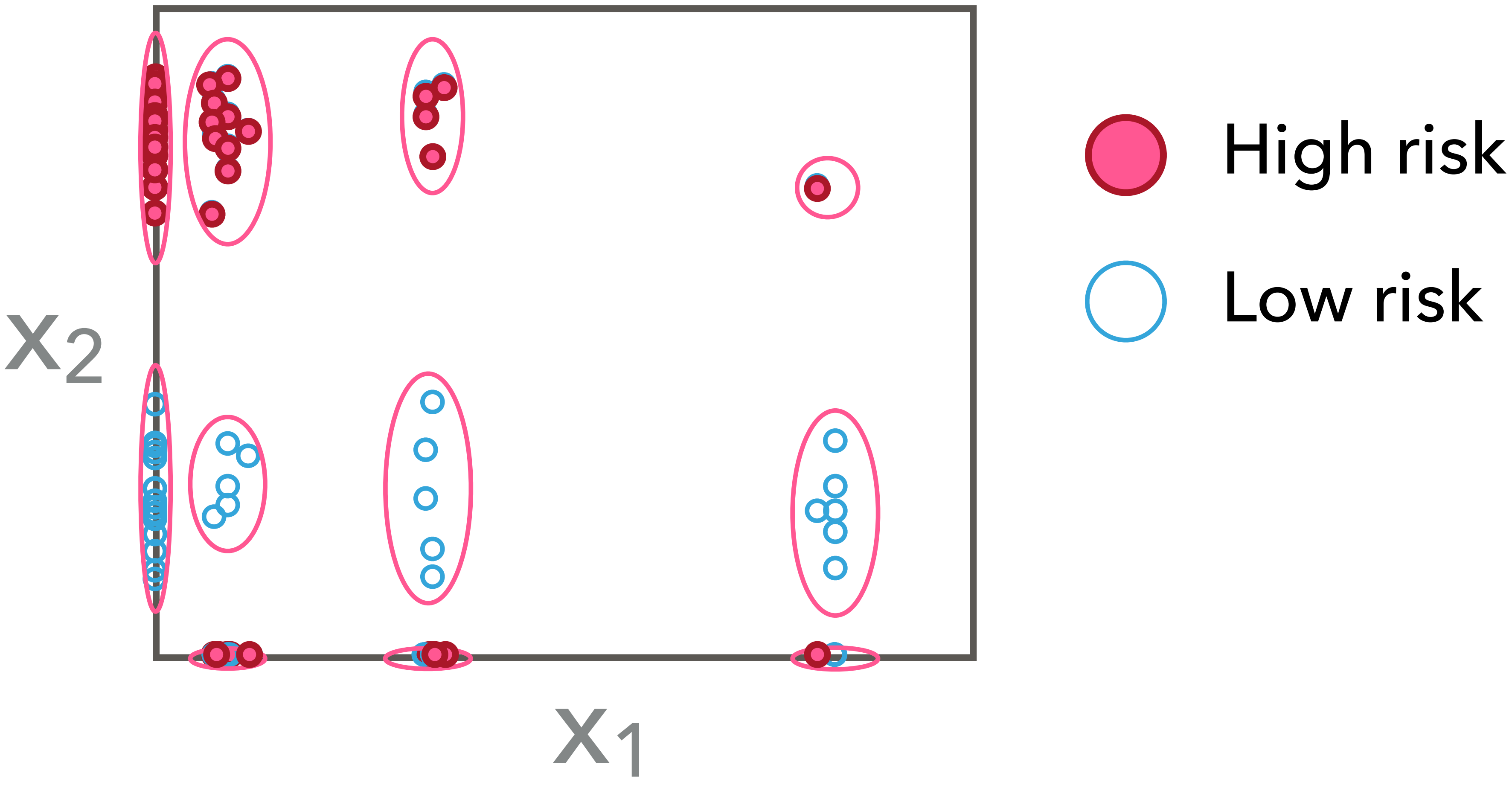


x_1 possesses a clear clustering structure… but it is irrelevant for this stratification task

HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

- Make use of a “response”: left-out variable(s) directly/closely linked to our true interest.
- e.g. x_1 and x_2 are blood protein abundances measured on a number of individuals, but our true interest is in y (high/low risk)

x_2 possesses a clustering structure that is relevant for this stratification task



x_1 possesses a clear clustering structure… but it is irrelevant for this stratification task

ILLUSTRATION 3: BREAST CANCER SUBTYPING

ILLUSTRATION 3: BREAST CANCER SUBTYPING

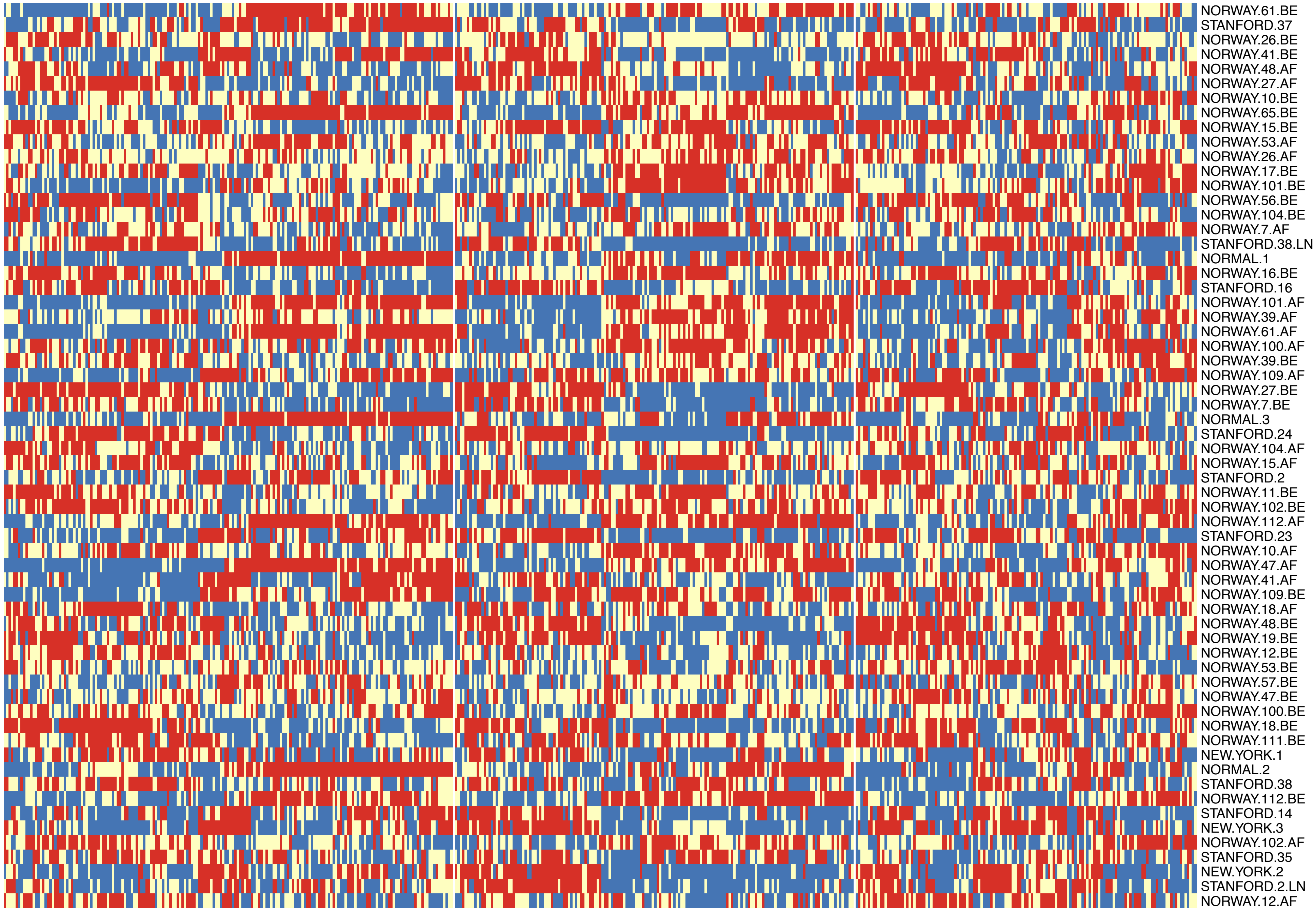


ILLUSTRATION 3: BREAST CANCER SUBTYPING

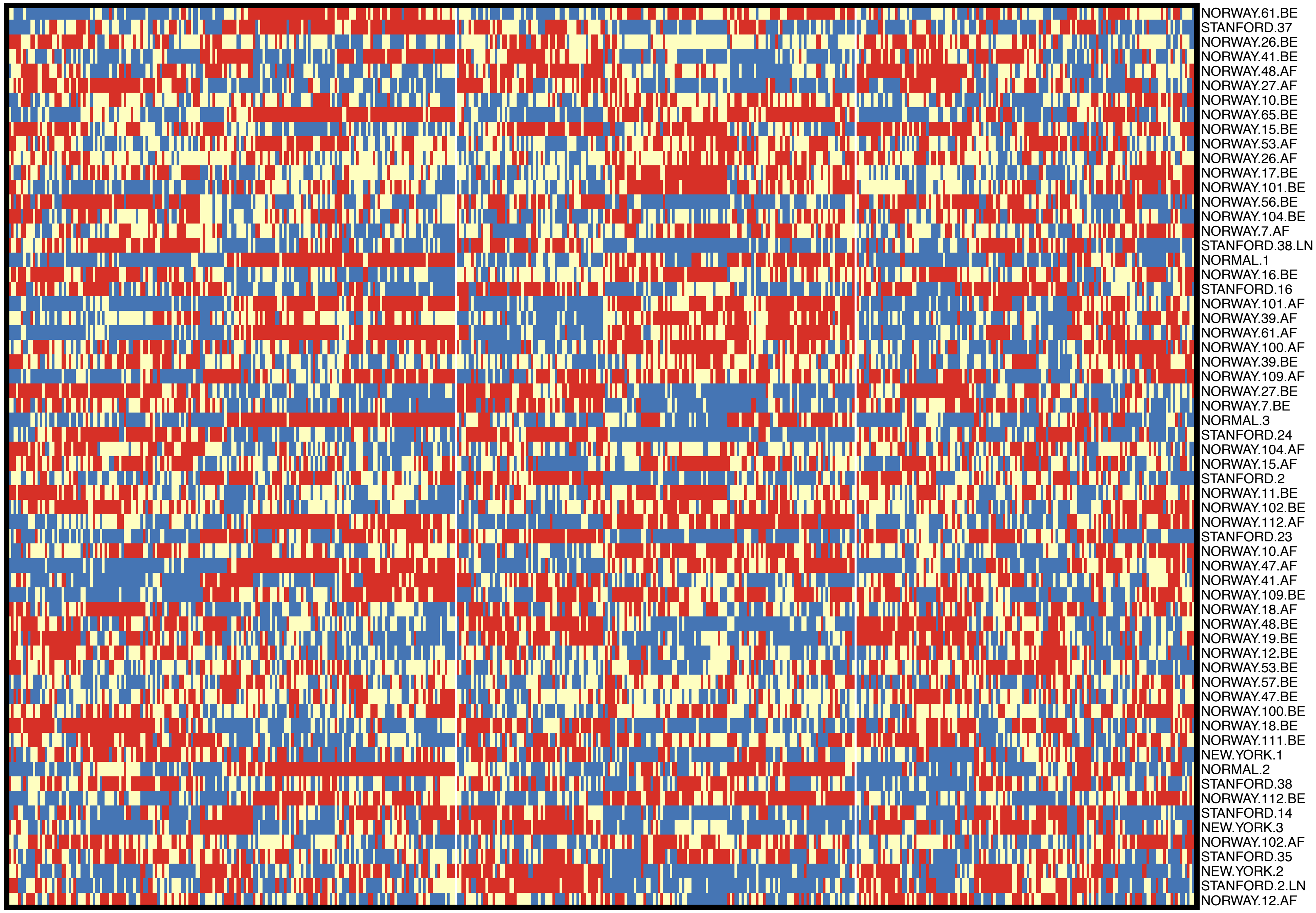


ILLUSTRATION 3: BREAST CANCER SUBTYPING

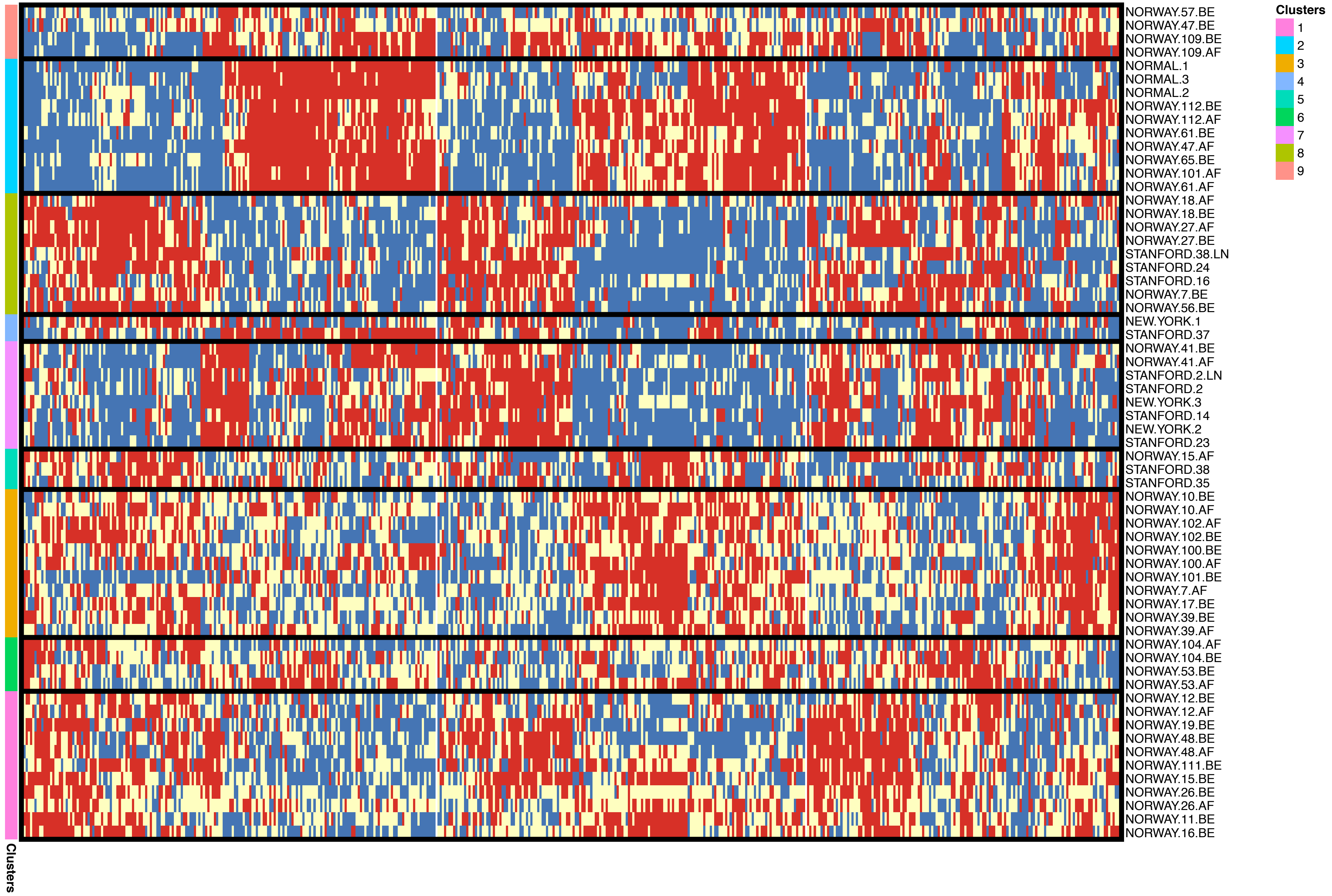


ILLUSTRATION 3: BREAST CANCER SUBTYPING

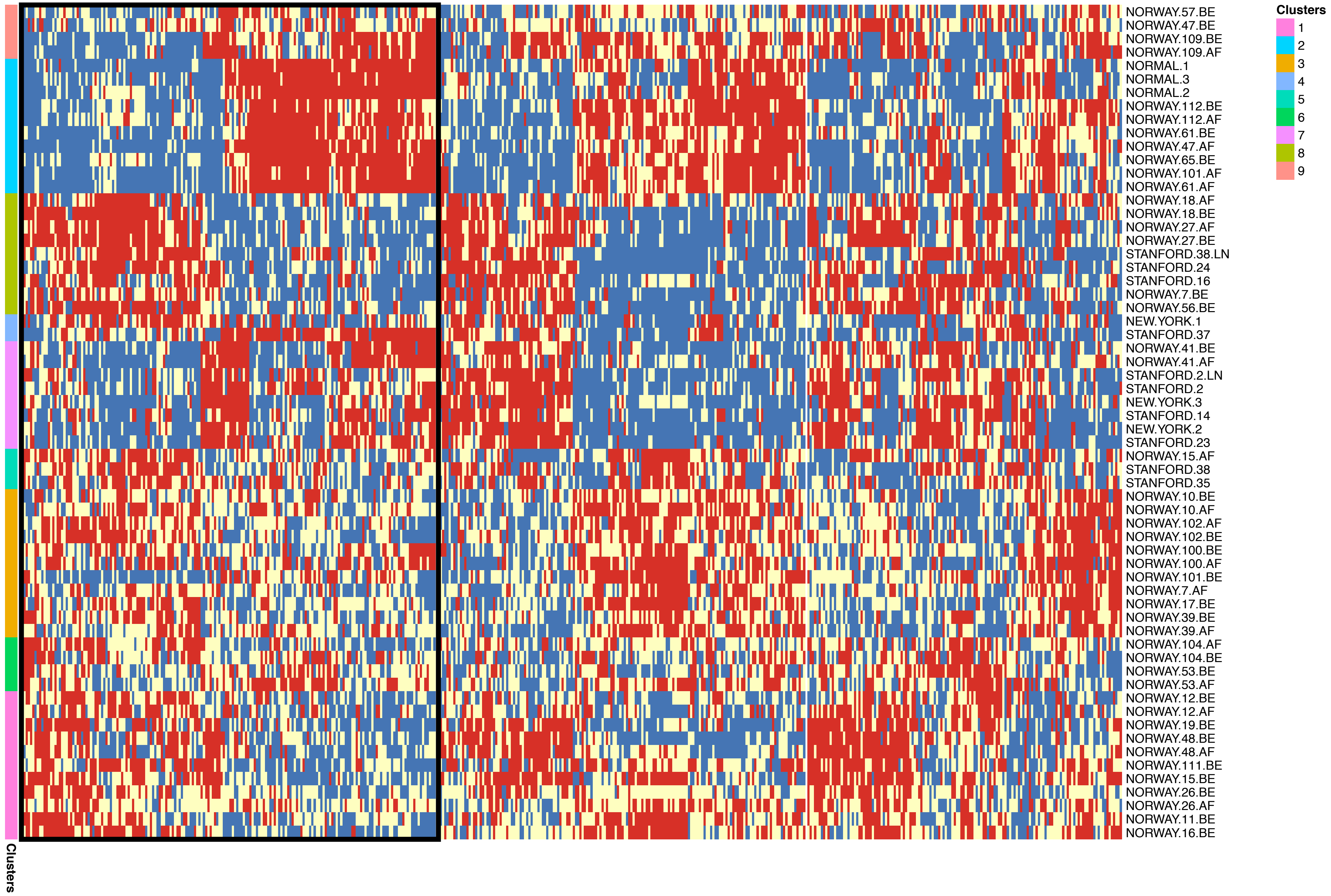


ILLUSTRATION 3: BREAST CANCER SUBTYPING

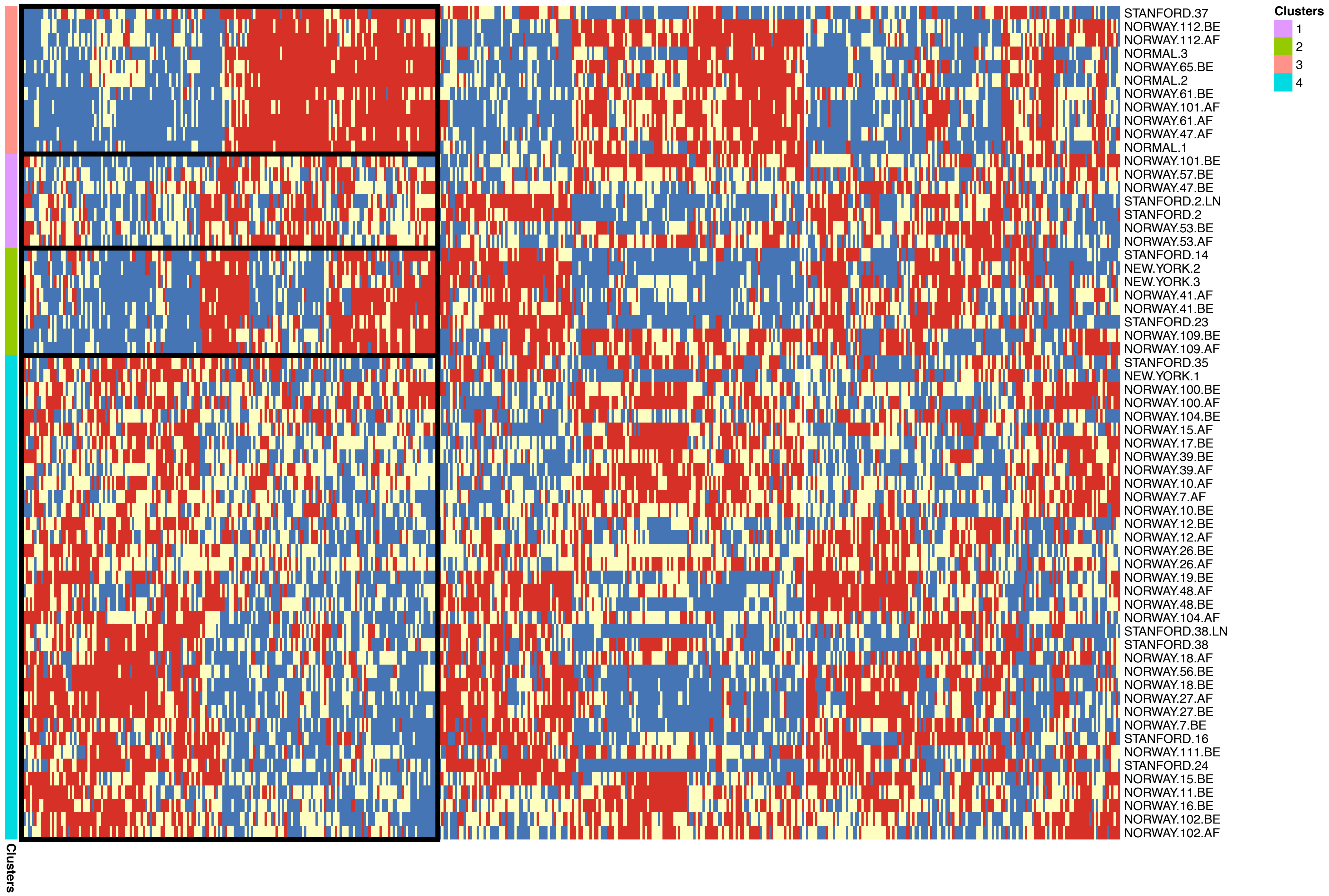


ILLUSTRATION 3: BREAST CANCER SUBTYPING

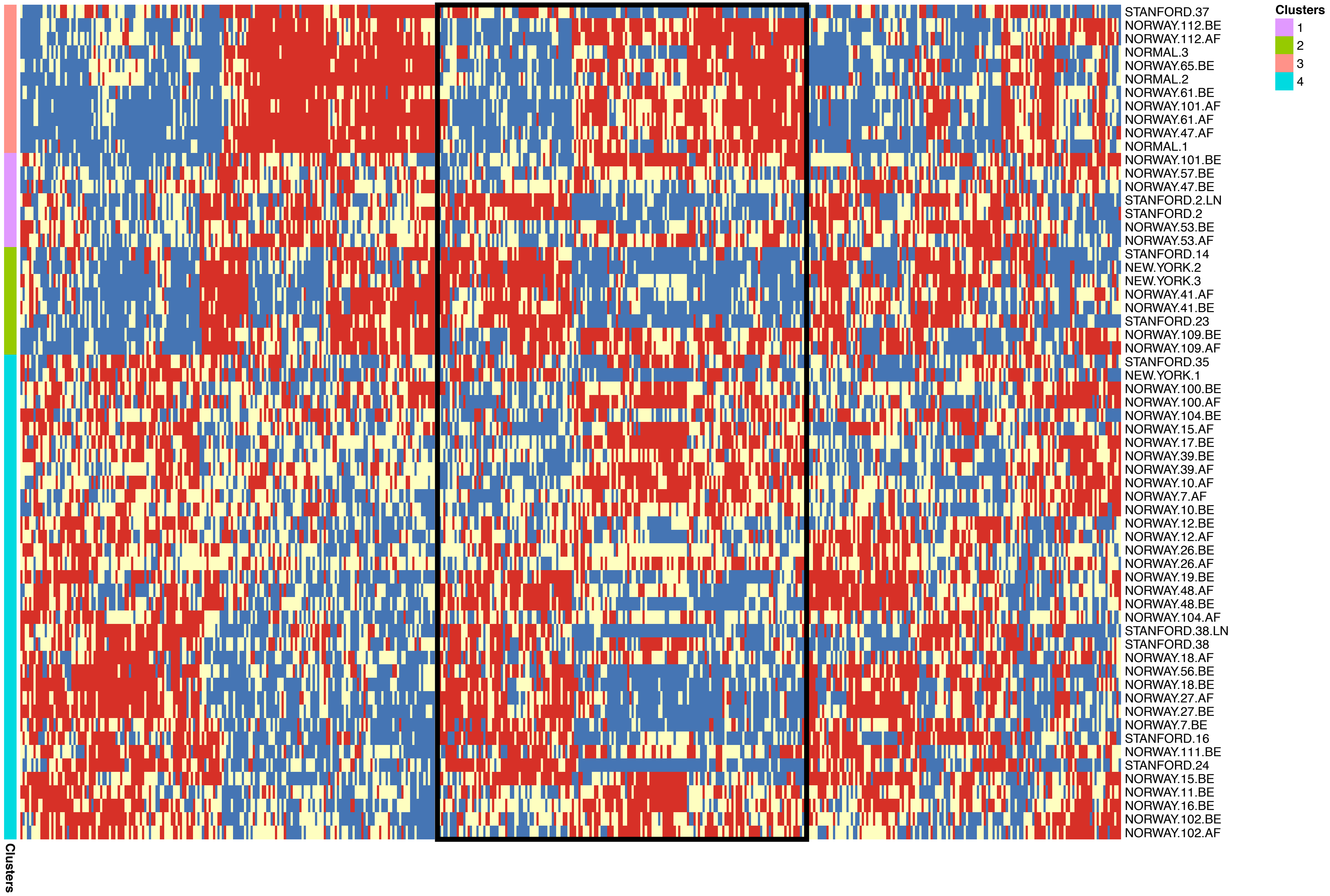


ILLUSTRATION 3: BREAST CANCER SUBTYPING

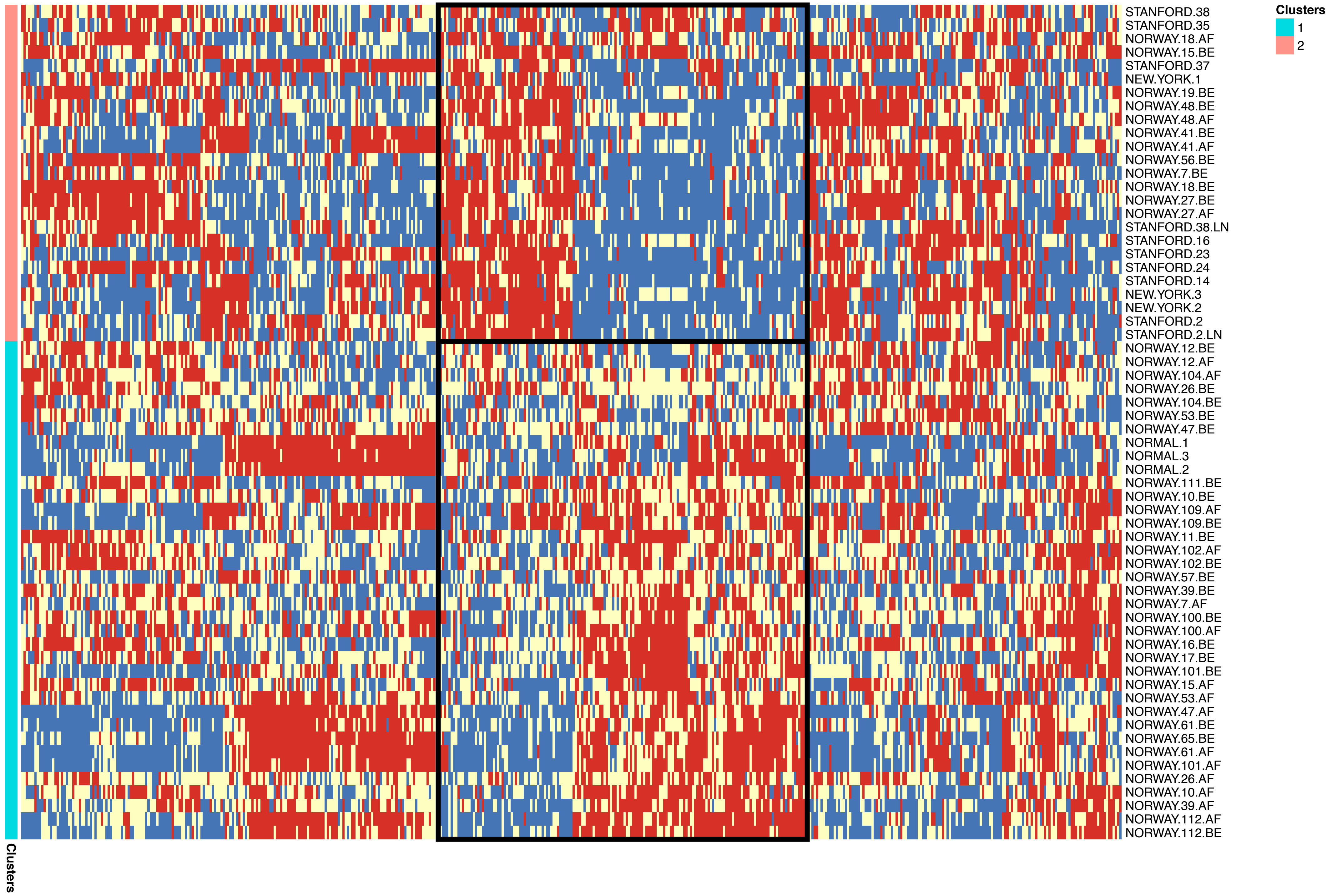


ILLUSTRATION 3: BREAST CANCER SUBTYPING

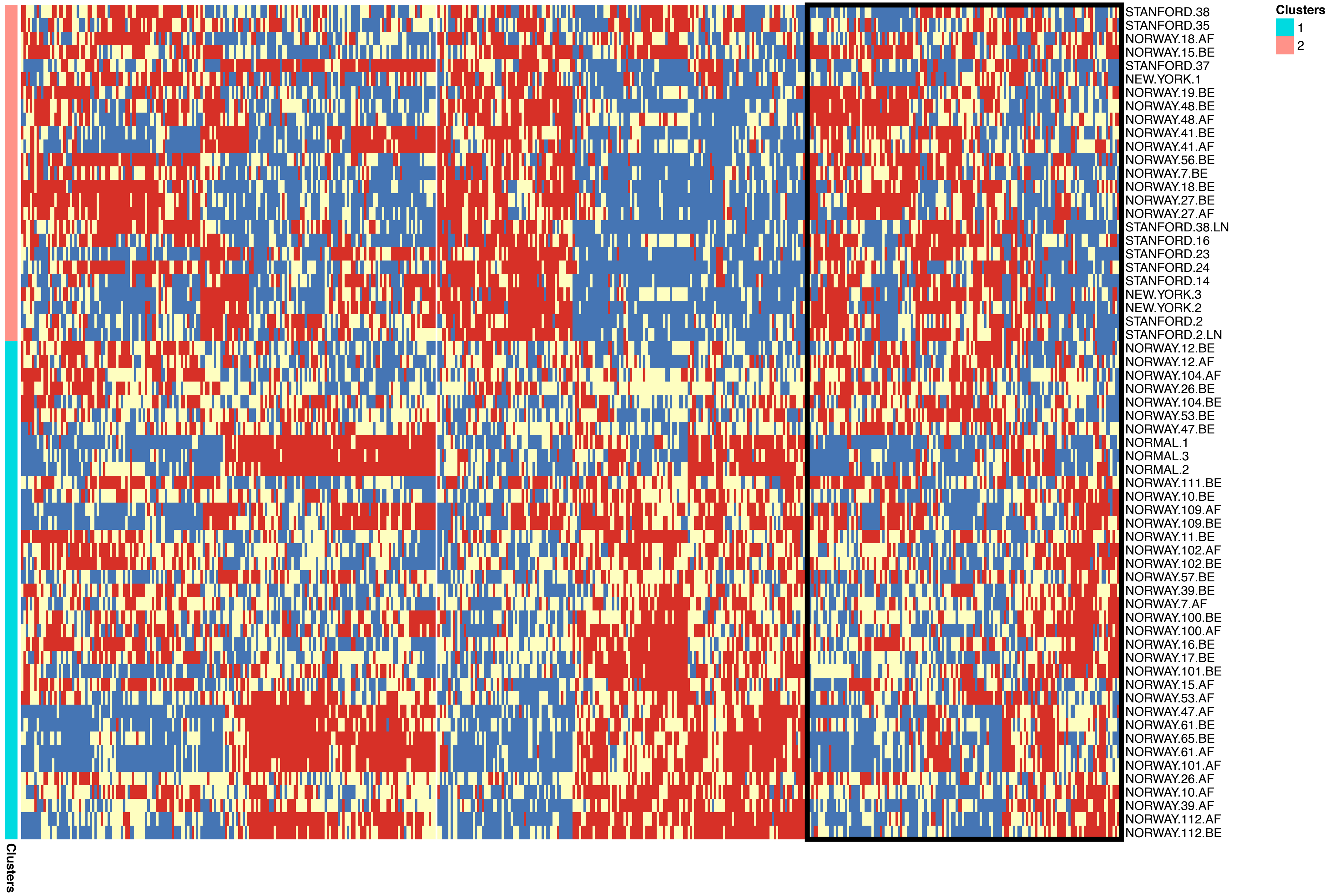


ILLUSTRATION 3: BREAST CANCER SUBTYPING

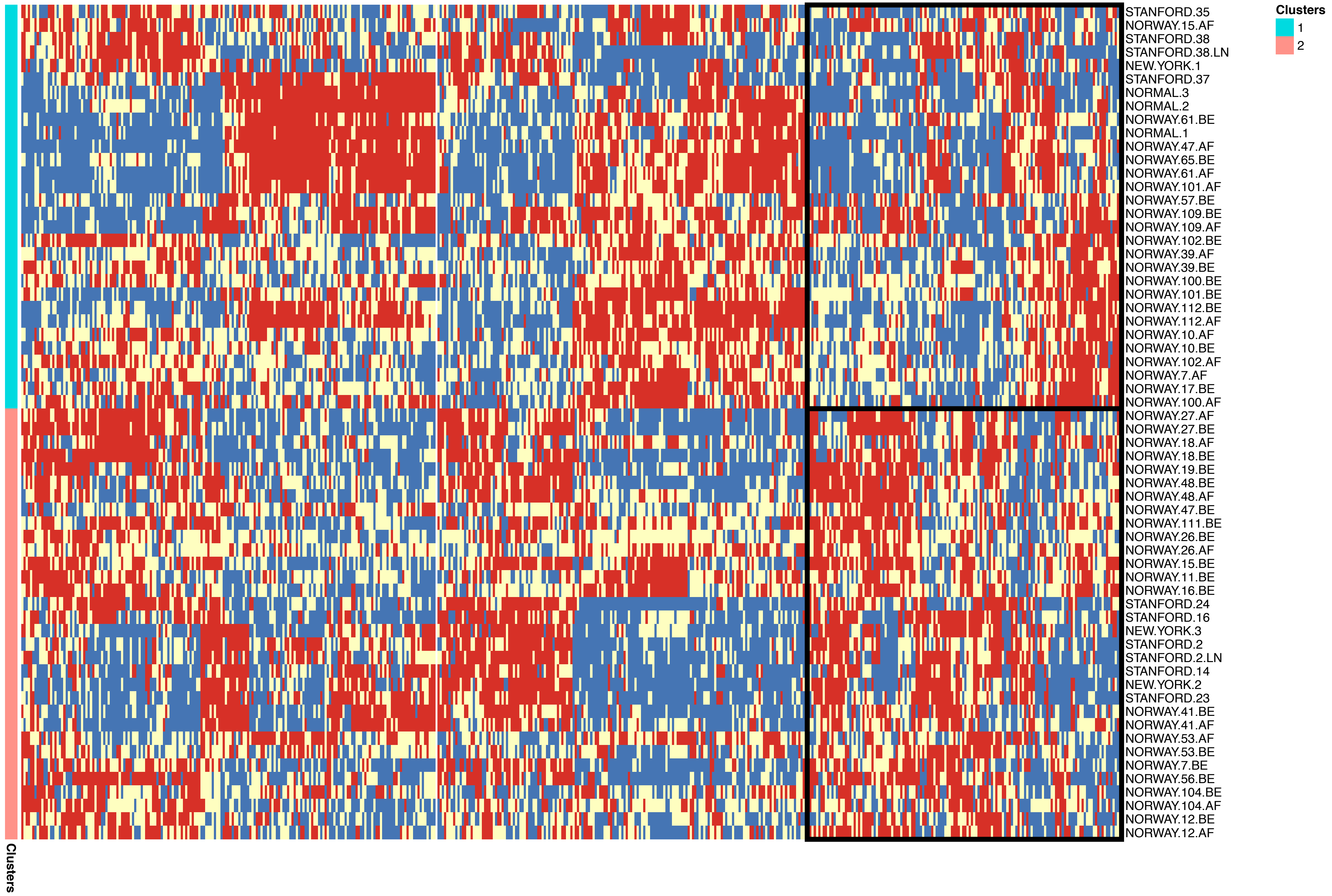
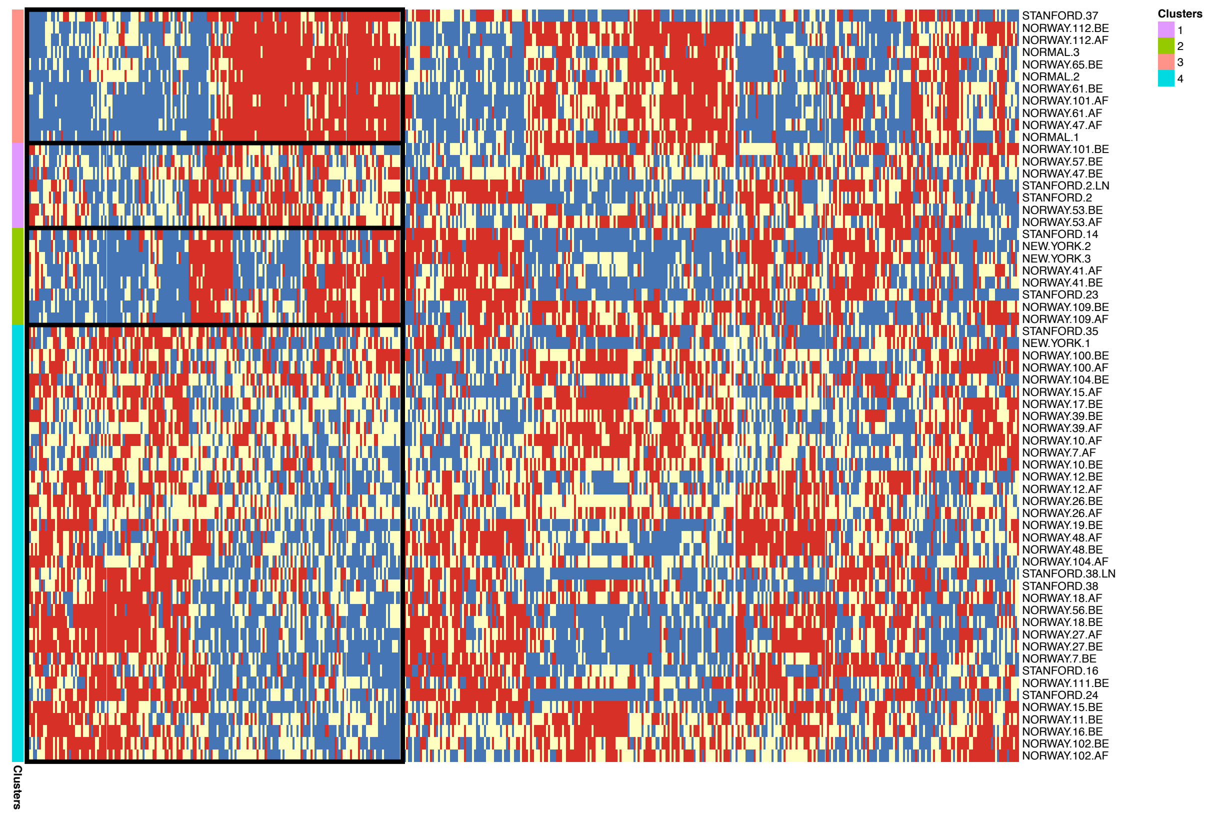


ILLUSTRATION 3: BREAST CANCER SUBTYPING

- Again, we obtain different clustering structures, depending upon which subset of variables we consider
- Which should we prefer?

ILLUSTRATION 3: BREAST CANCER SUBTYPING

- Again, we obtain different clustering structures, depending upon which subset of variables we consider
- Which should we prefer?



- This clustering defines clinically actionable subtypes.

HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

- Make use of a “response”: left-out variable(s) directly/closely linked to our true interest.

HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

- Make use of a “response”: left-out variable(s) directly/closely linked to our true interest.
- Traditional approaches only make use of this “response” **after** the clustering has been performed, to decide if the clustering has revealed useful structure.

HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

- Make use of a “response”: left-out variable(s) directly/closely linked to our true interest.

- Traditional approaches only make use of this “response” **after** the clustering has been performed, to decide if the clustering has revealed useful structure.

- Instead, we use the “response” to *supervise* the clustering, and hence pick out the most relevant structure…

HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

- Make use of a “response”: left-out variable(s) directly/closely linked to our true interest.
- Traditional approaches only make use of this “response” **after** the clustering has been performed, to decide if the clustering has revealed useful structure.
- Instead, we use the “response” to *supervise* the clustering, and hence pick out the most relevant structure…
    - … and the subset of variables that define it.

HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

- Make use of a “response”: left-out variable(s) directly/closely linked to our true interest.
- Traditional approaches only make use of this “response” **after** the clustering has been performed, to decide if the clustering has revealed useful structure.
- Instead, we use the “response” to *supervise* the clustering, and hence pick out the most relevant structure…
    - … and the subset of variables that define it.
- Increasingly helpful for high-dimensional problems.

HOW DO WE DECIDE IF A CLUSTERING STRUCTURE IS RELEVANT?

- Make use of a “response”: left-out variable(s) directly/closely linked to our true interest.
- Traditional approaches only make use of this “response” **after** the clustering has been performed, to decide if the clustering has revealed useful structure.
- Instead, we use the “response” to *supervise* the clustering, and hence pick out the most relevant structure…
  - … and the subset of variables that define it.
- Increasingly helpful for high-dimensional problems.

THIS IS THE BASIC IDEA BEHIND PROFILE REGRESSION.

PART 2:

PART 2:
PROFILE REGRESSION
(SEMI-SUPERVISED CLUSTERING)

MIXTURE MODELS: INTRODUCTION AND NOTATION

We model our data using a mixture model:

$$p(x) = \sum_{c=1}^{K} \pi_c f(x|\theta_c). \tag{1}$$

MIXTURE MODELS: INTRODUCTION AND NOTATION

We model our data using a mixture model:

$$p(x) = \sum_{c=1}^{K} \pi_c f(x|\theta_c). \tag{1}$$

- K is the number of mixture components
- π_c are the mixture proportions
- f is a parametric density (such as a Gaussian)
- θ_c are the parameters associated with the c-th component

MIXTURE MODELS: INTRODUCTION AND NOTATION

We model our data using a mixture model:

$$p(x) = \sum_{c=1}^{K} \pi_c f(x|\theta_c). \tag{1}$$

- K is the number of mixture components
- π_c are the mixture proportions
- f is a parametric density (such as a Gaussian)
- θ_c are the parameters associated with the c-th component

Different choices for the density f allow us to model different types of data

- a normal distribution might be appropriate for continuous data, while a multinomial might be appropriate for categorical data

MIXTURE MODELS: INTRODUCTION AND NOTATION

We model our data using a mixture model:

$$p(x) = \sum_{c=1}^{K} \pi_c f(x|\theta_c). \tag{1}$$

- K is the number of mixture components
- π_c are the mixture proportions
- f is a parametric density (such as a Gaussian)
- θ_c are the parameters associated with the c-th component

Different choices for the density f allow us to model different types of data

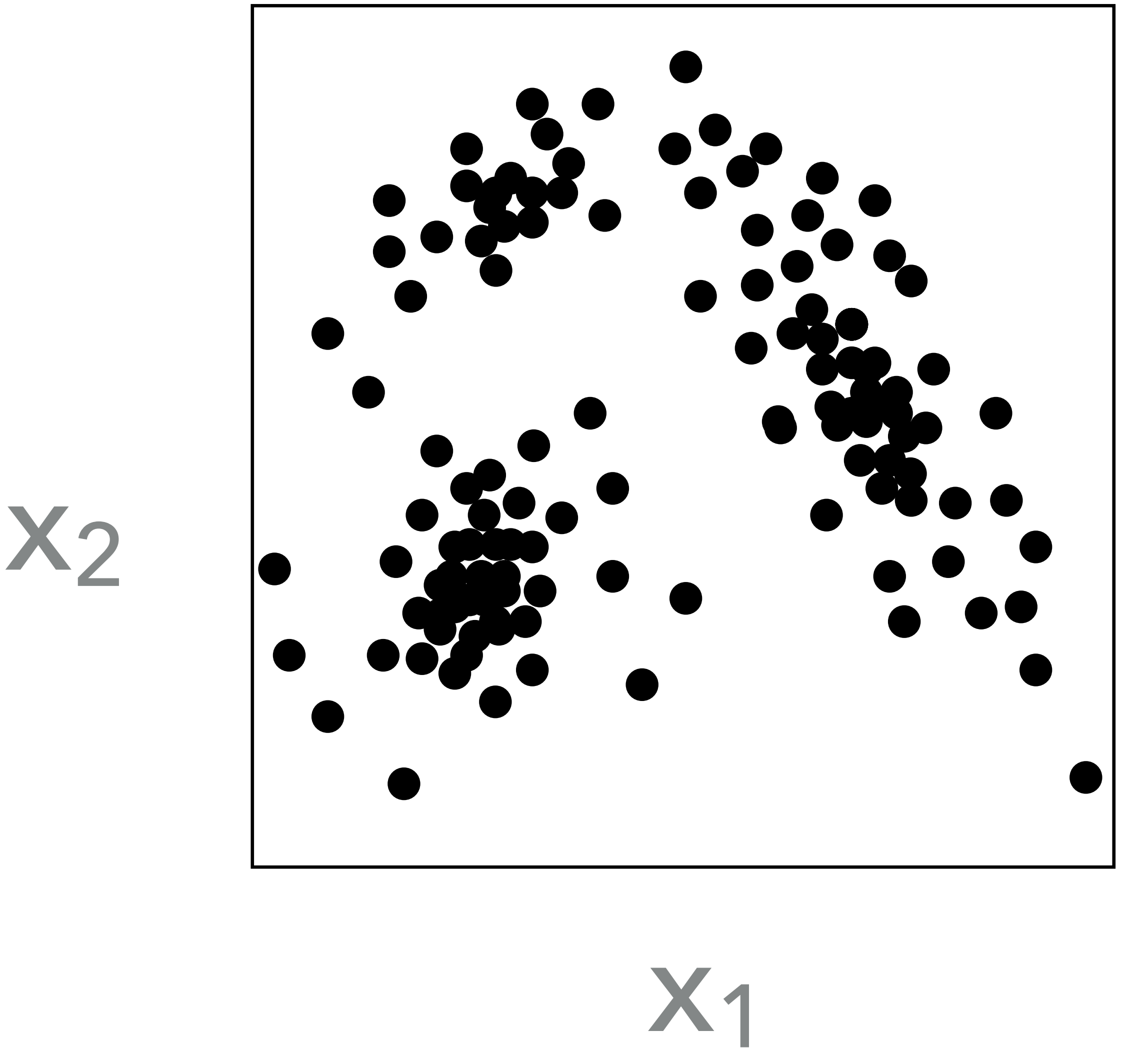
- a normal distribution might be appropriate for continuous data, while a multinomial might be appropriate for categorical data
  - **Kirk**, Huvet, Melamed, Maertens, & Bangham (2016). Retroviruses integrate into a shared, non-palindromic DNA motif. *Nature Microbiology.*

MIXTURE MODELS: INITIAL BASIC CASE

$$p(x) = \pi_1 f(x|\mu_1, \Sigma_1) + \pi_2 f(x|\mu_2, \Sigma_2) + \pi_3 f(x|\mu_3, \Sigma_3).$$

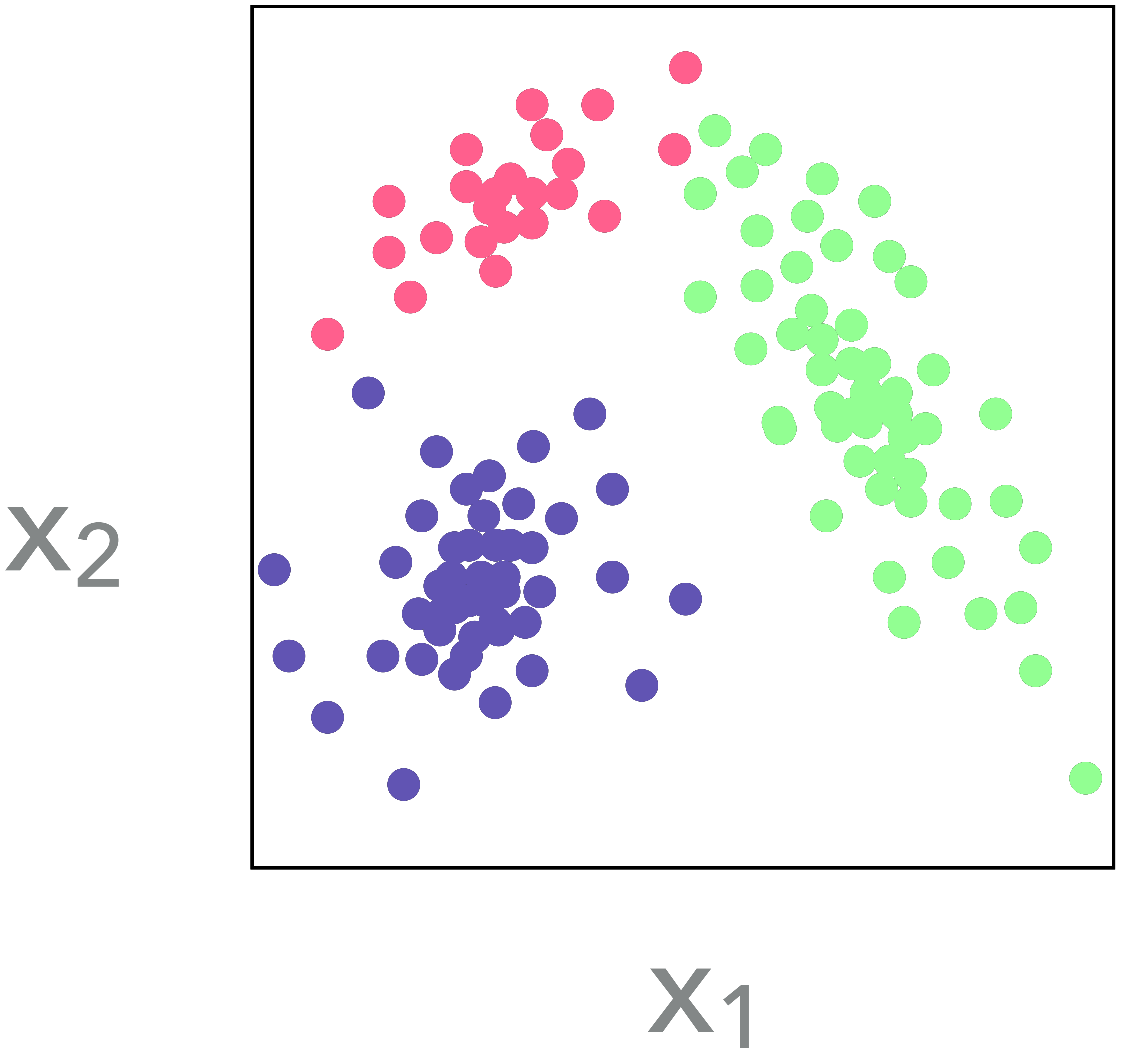
MIXTURE MODELS: INITIAL BASIC CASE

$$p(x) = \pi_1 f(x|\mu_1, \Sigma_1) + \pi_2 f(x|\mu_2, \Sigma_2) + \pi_3 f(x|\mu_3, \Sigma_3).$$



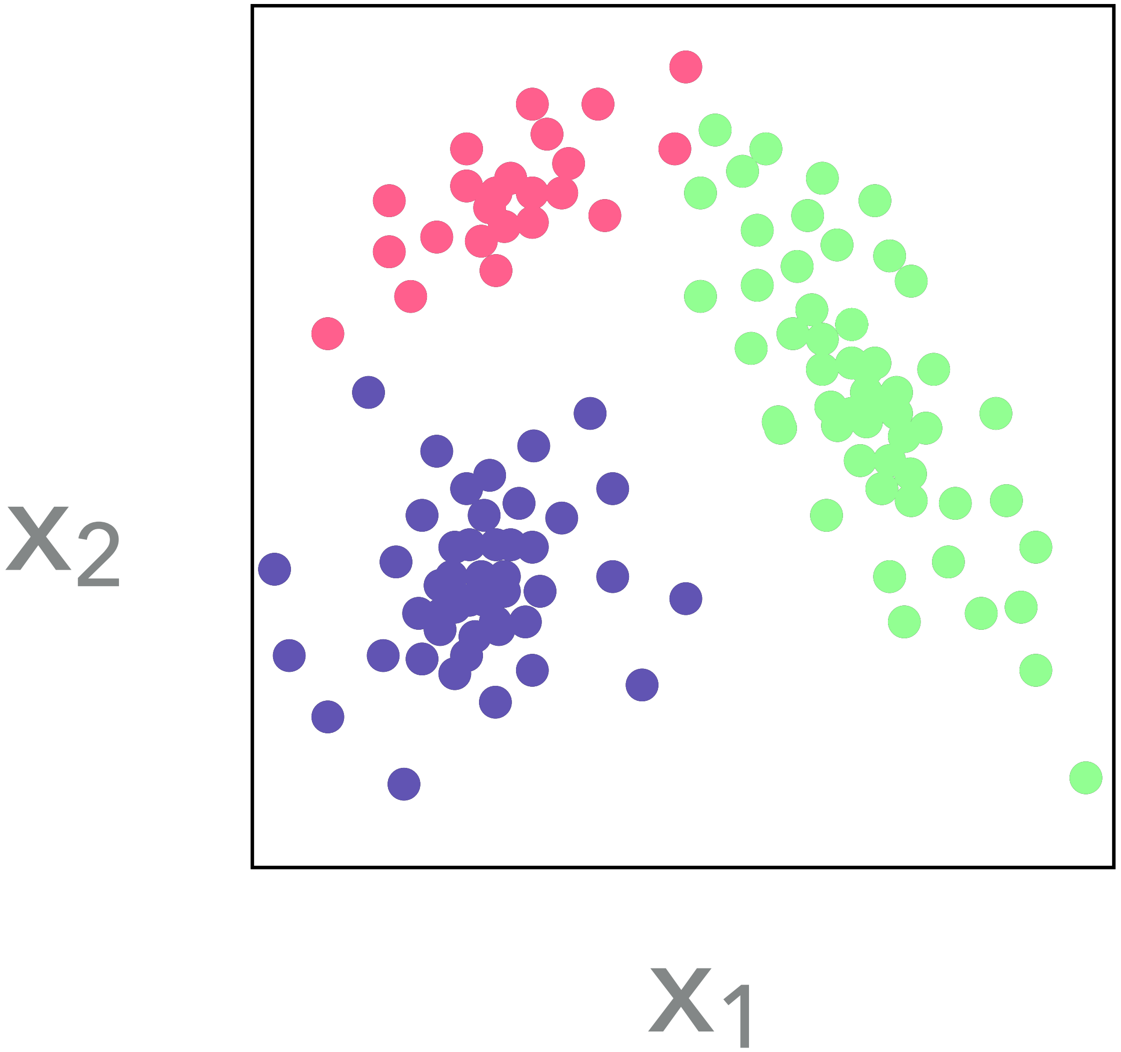
MIXTURE MODELS: INITIAL BASIC CASE

$$p(x) = \pi_1 f(x|\mu_1, \Sigma_1) + \pi_2 f(x|\mu_2, \Sigma_2) + \pi_3 f(x|\mu_3, \Sigma_3).$$



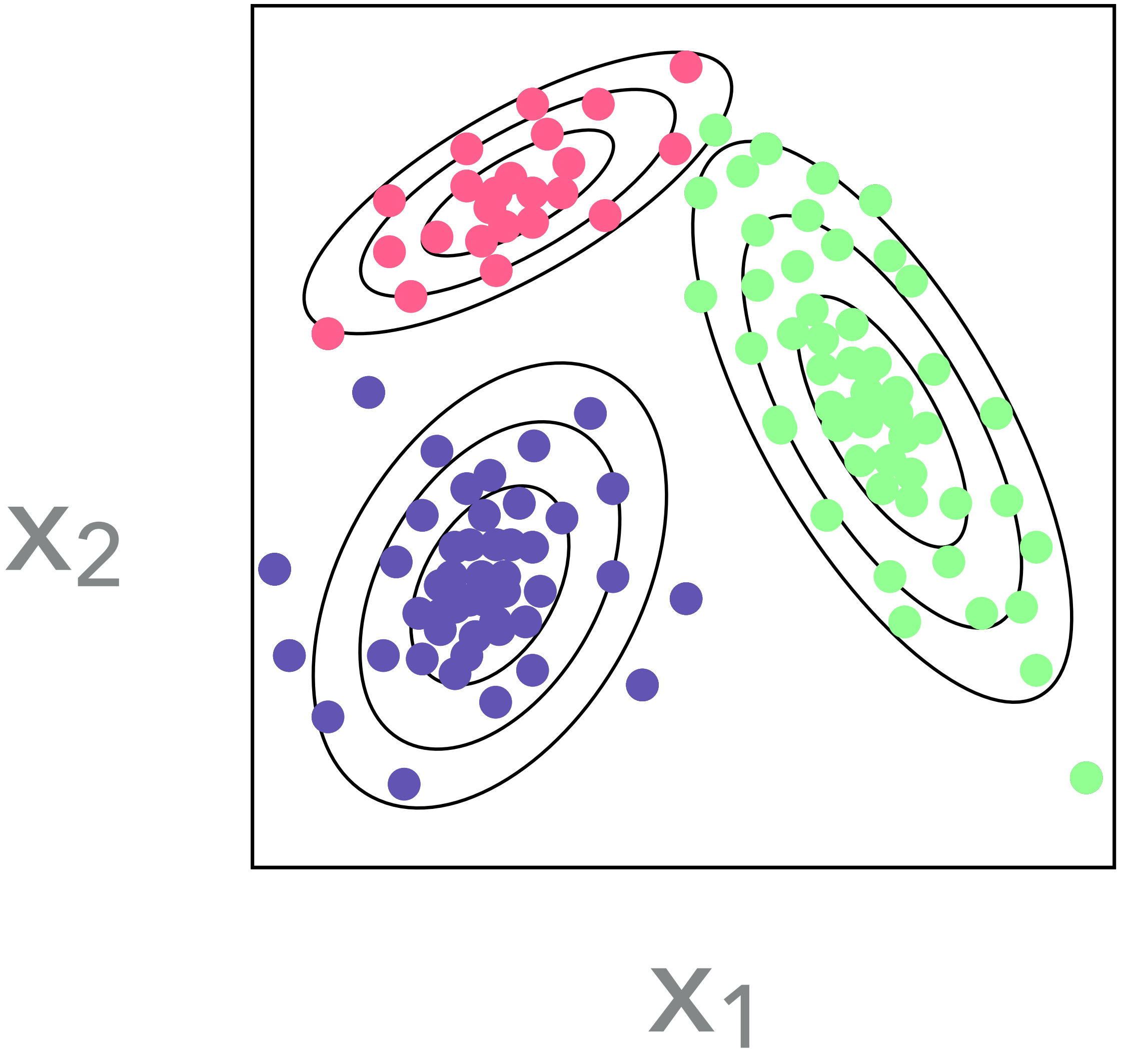
MIXTURE MODELS: INITIAL BASIC CASE

$$p(x) = \pi_1 f(x|\mu_1, \Sigma_1) + \pi_2 f(x|\mu_2, \Sigma_2) + \pi_3 f(x|\mu_3, \Sigma_3).$$



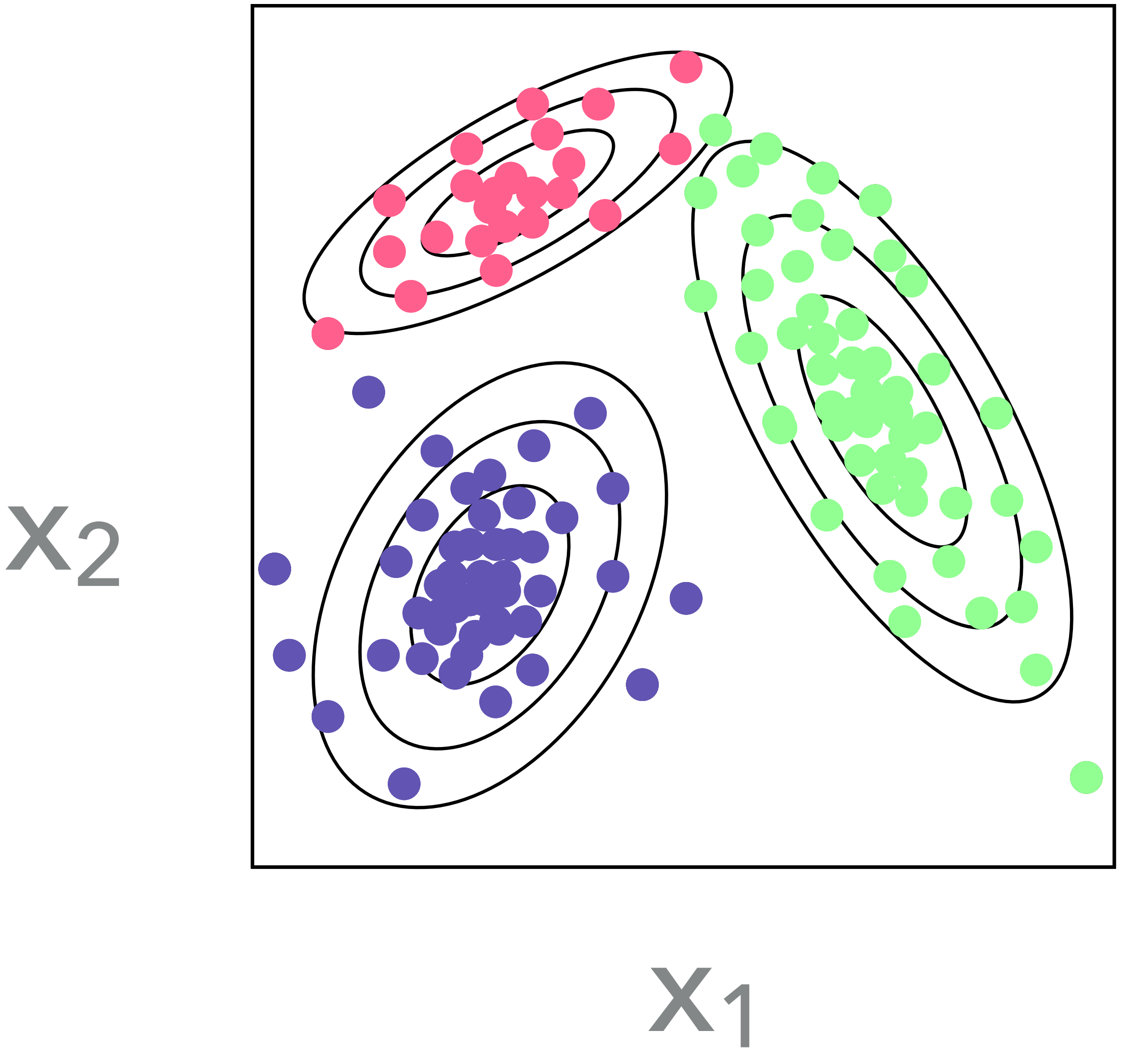
MIXTURE MODELS: INITIAL BASIC CASE

$$p(x) = \pi_1 f(x|\mu_1, \Sigma_1) + \pi_2 f(x|\mu_2, \Sigma_2) + \pi_3 f(x|\mu_3, \Sigma_3).$$



MIXTURE MODELS: INITIAL BASIC CASE

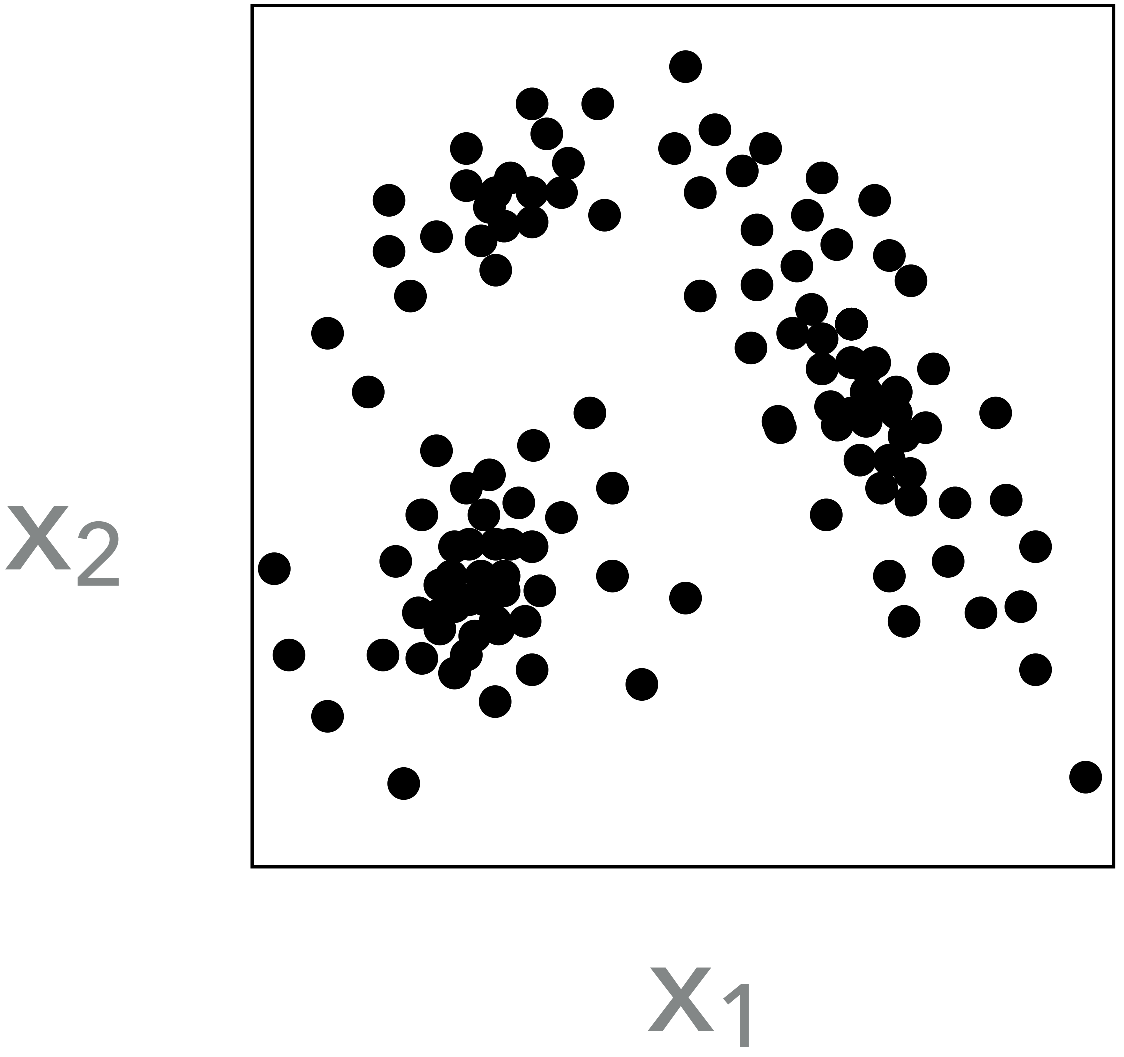
$$p(x) = \pi_1 f(x|\mu_1, \Sigma_1) + \pi_2 f(x|\mu_2, \Sigma_2) + \pi_3 f(x|\mu_3, \Sigma_3).$$



MIXTURE MODELS: INITIAL BASIC CASE

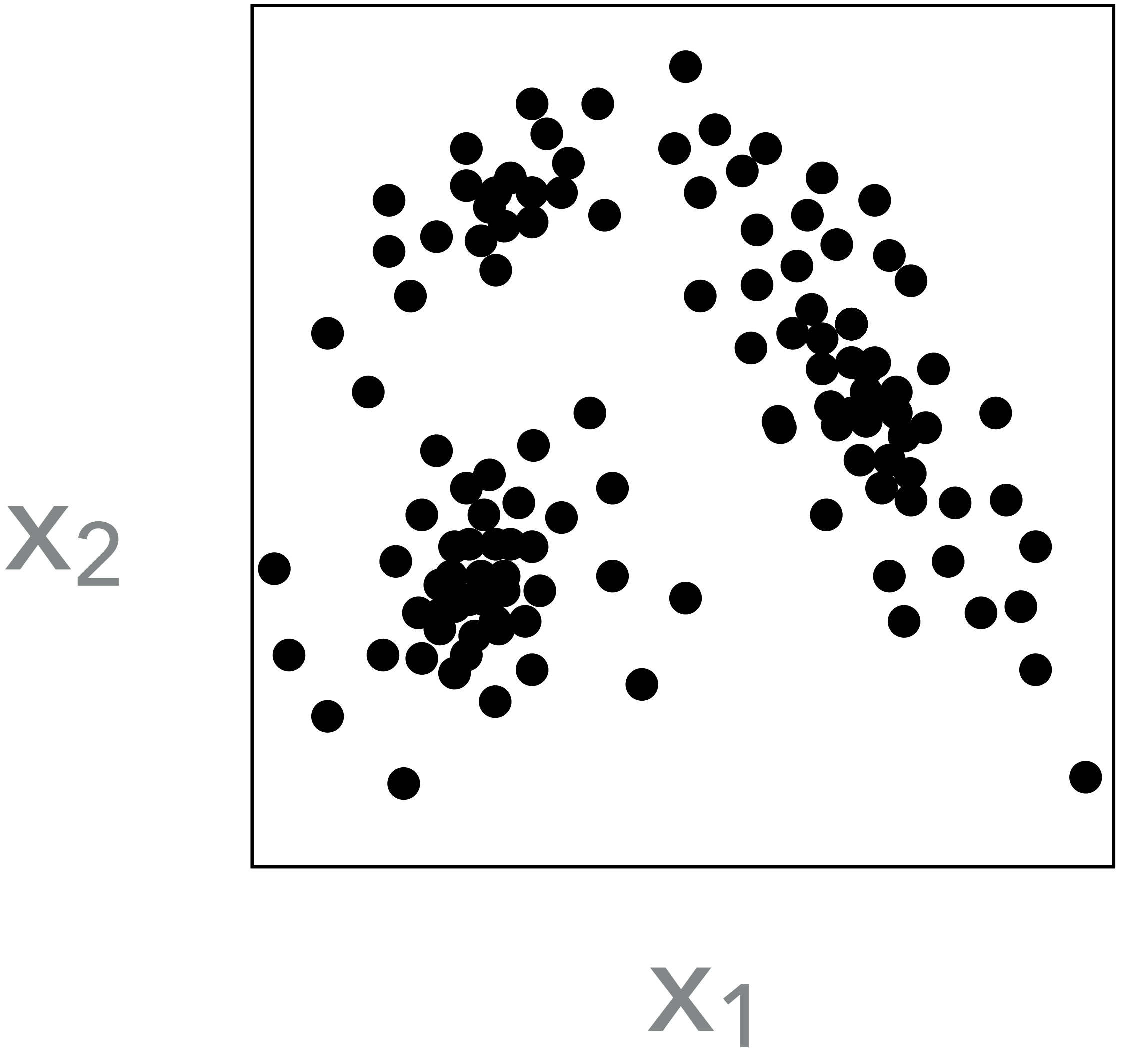
MIXTURE MODELS: INITIAL BASIC CASE

$$p(x) = \pi_1 f(x|\mu_1, \Sigma_1) + \pi_2 f(x|\mu_2, \Sigma_2) + \pi_3 f(x|\mu_3, \Sigma_3).$$



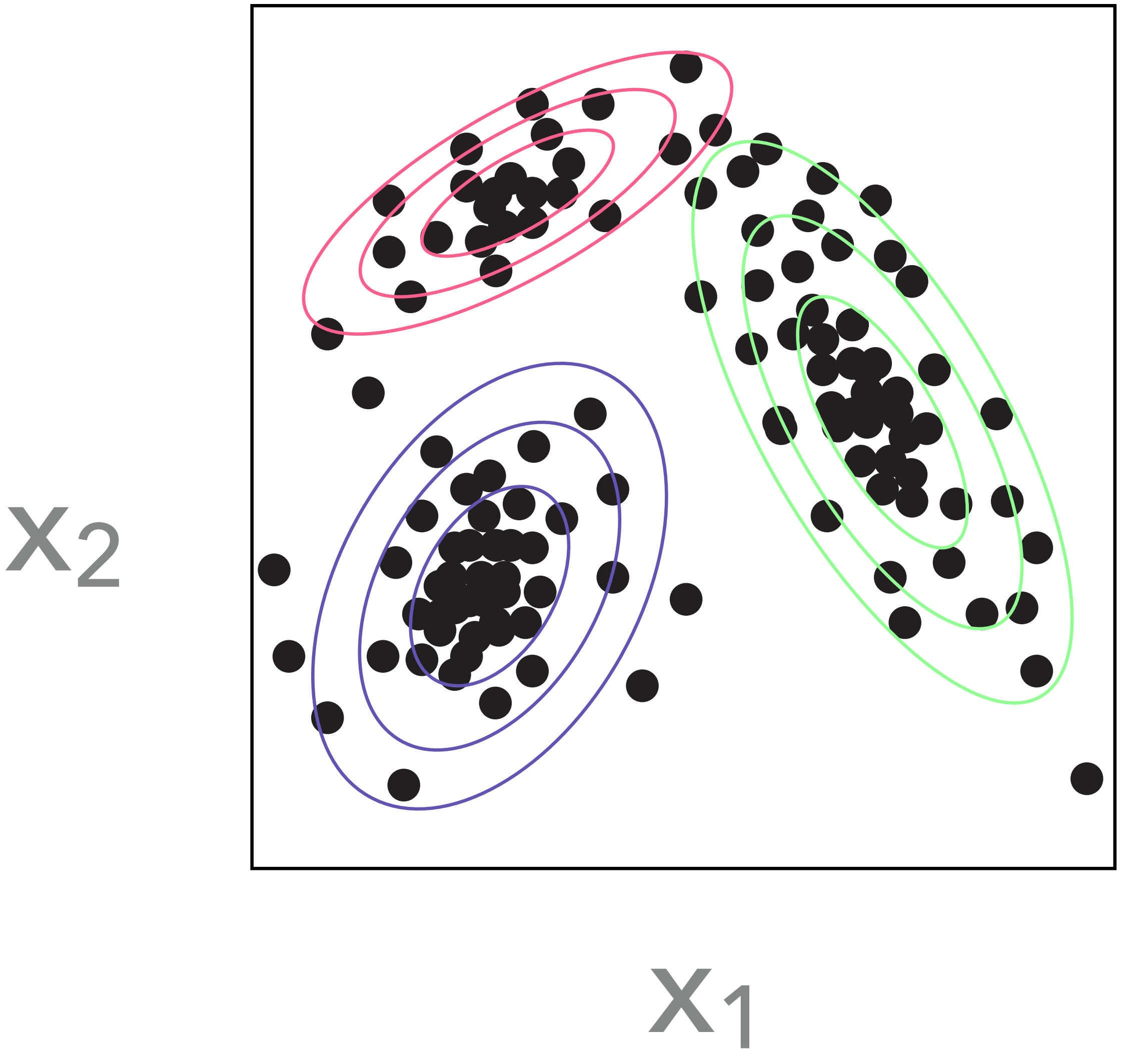
MIXTURE MODELS: INITIAL BASIC CASE

$$p(x) = \pi_1 f(x|\mu_1, \Sigma_1) + \pi_2 f(x|\mu_2, \Sigma_2) + \pi_3 f(x|\mu_3, \Sigma_3).$$



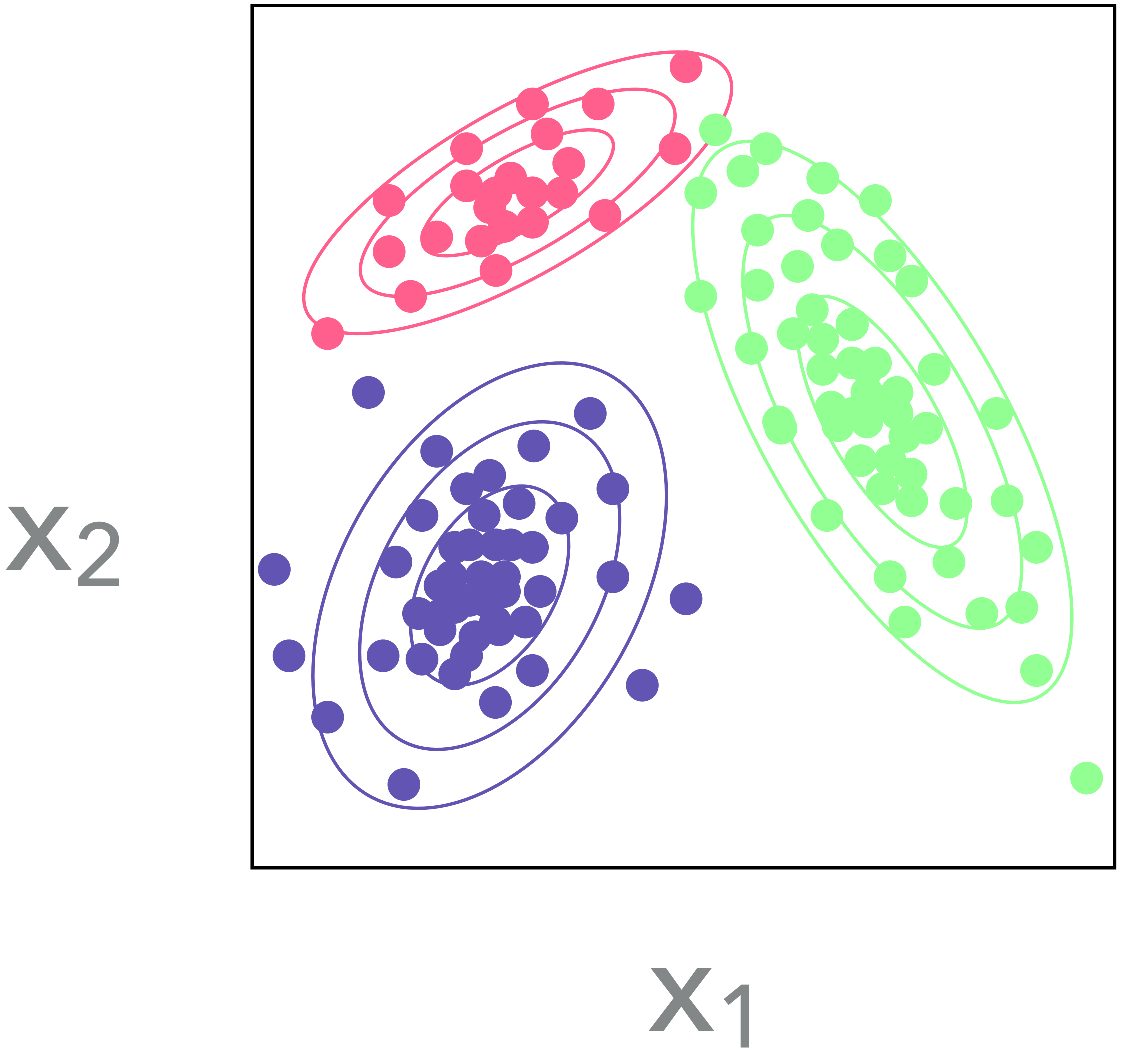
MIXTURE MODELS: INITIAL BASIC CASE

$$p(x) = \pi_1 f(x|\mu_1, \Sigma_1) + \pi_2 f(x|\mu_2, \Sigma_2) + \pi_3 f(x|\mu_3, \Sigma_3).$$



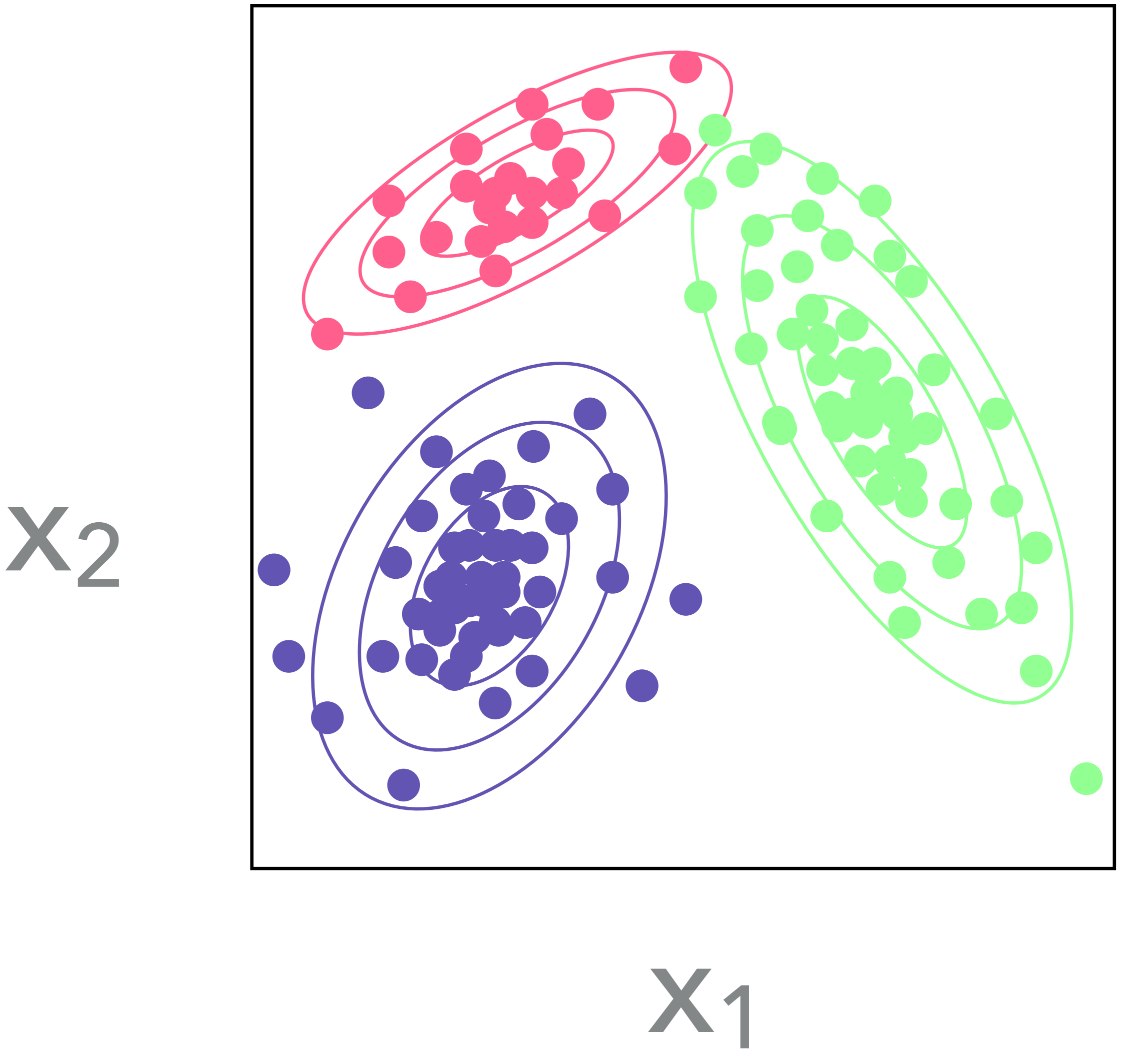
MIXTURE MODELS: INITIAL BASIC CASE

$$p(x) = \pi_1 f(x|\mu_1, \Sigma_1) + \pi_2 f(x|\mu_2, \Sigma_2) + \pi_3 f(x|\mu_3, \Sigma_3).$$



MIXTURE MODELS: INITIAL BASIC CASE

$$p(x) = \pi_1 f(x|\mu_1, \Sigma_1) + \pi_2 f(x|\mu_2, \Sigma_2) + \pi_3 f(x|\mu_3, \Sigma_3).$$



MIXTURE MODELS: INTRODUCTION AND NOTATION

- Given observed data $x_1, \ldots, x_n$, we wish to perform Bayesian inference for the unknown parameters.

MIXTURE MODELS: INTRODUCTION AND NOTATION

- Given observed data $x_1, \ldots, x_n$, we wish to perform Bayesian inference for the unknown parameters.
- We introduce latent *component allocation* variables $c_j \in \{1, \ldots, K\}$, such that c_i is the component responsible for observation x_i.

MIXTURE MODELS: INTRODUCTION AND NOTATION

- Given observed data $x_1, \ldots, x_n$, we wish to perform Bayesian inference for the unknown parameters.
- We introduce latent *component allocation* variables $c_j \in \{1, \ldots, K\}$, such that c_i is the component responsible for observation x_i.
- We specify the complete model as follows:

$$\begin{aligned} x_i | c_i, \boldsymbol{\theta} &\sim F(\theta_{c_i}), \\ c_i | \boldsymbol{\pi} &\sim \text{Categorical}(\pi_1, \ldots, \pi_K), \\ \pi_1, \ldots, \pi_K &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K), \\ \theta_c &\sim G^{(0)}, \end{aligned} \tag{3}$$

  - F is the distribution corresponding to density f
  - $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ is the collection of K mixture proportions
  - α is a mass/concentration parameter (which may also be inferred)
  - $G^{(0)}$ is the prior for the component parameters

MIXTURE MODELS: INTRODUCTION AND NOTATION

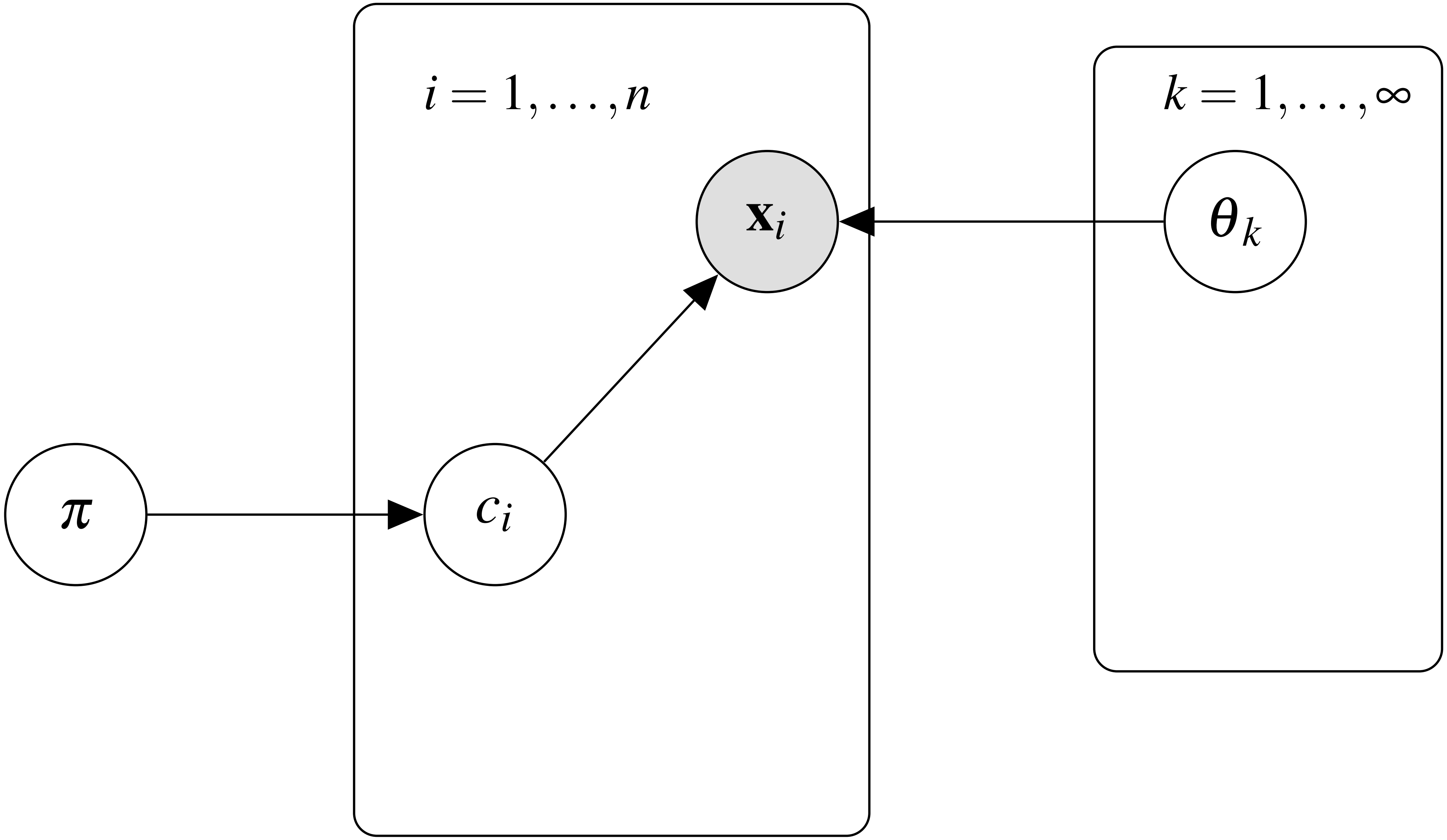
CLUSTERING THE DATA

- A realisation of the collection of component allocation variables, $(c_1, \dots, c_n)$, defines a *clustering* of the data
  - If $c_i = c_j$, then x_i and x_j are clustered together
- Each c_j is a member of the set $\{1, \dots, K\}$, so K places an upper bound on the number of clusters

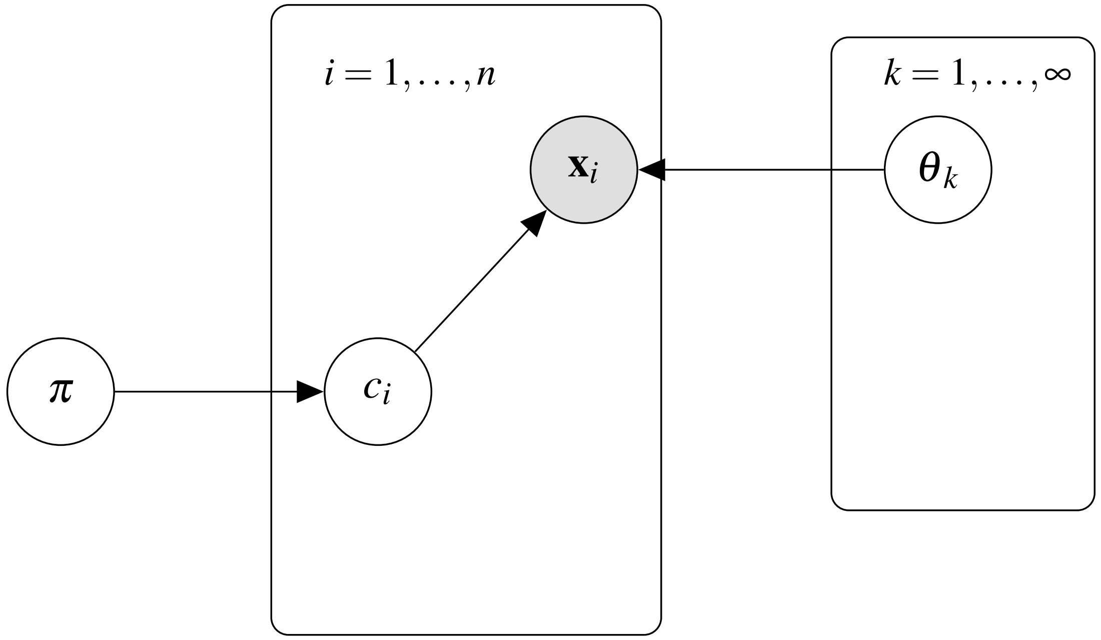
- The Dirichlet process mixture model may be derived by considering the limit $K \to \infty$

MIXTURE MODELS: INTRODUCTION AND NOTATION

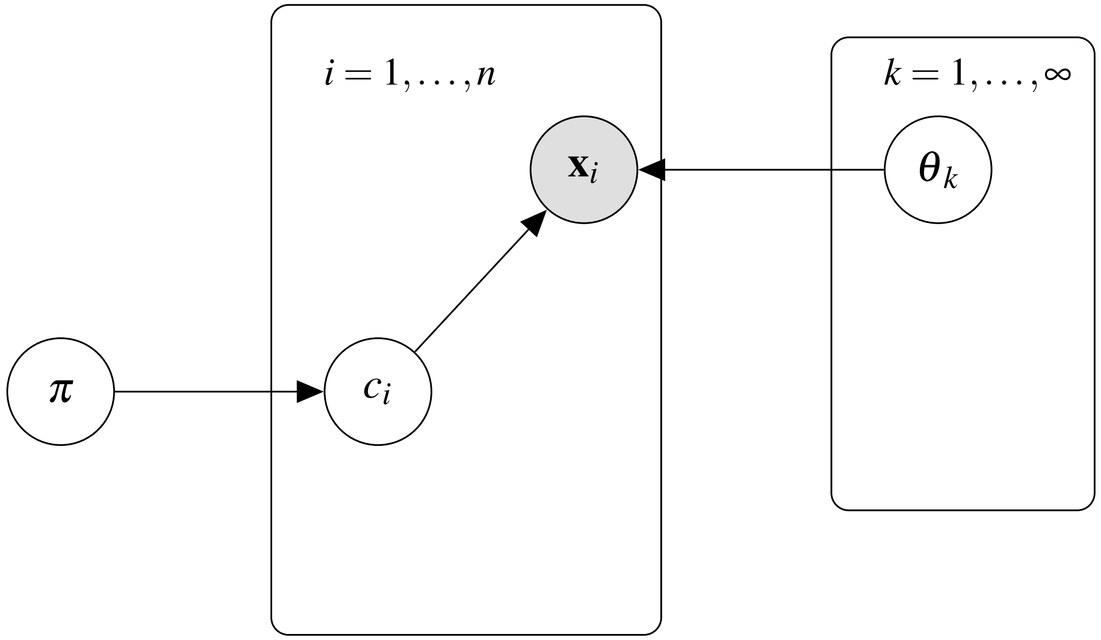
PLATE DIAGRAM:



MIXTURE MODELS: INTRODUCTION AND NOTATION

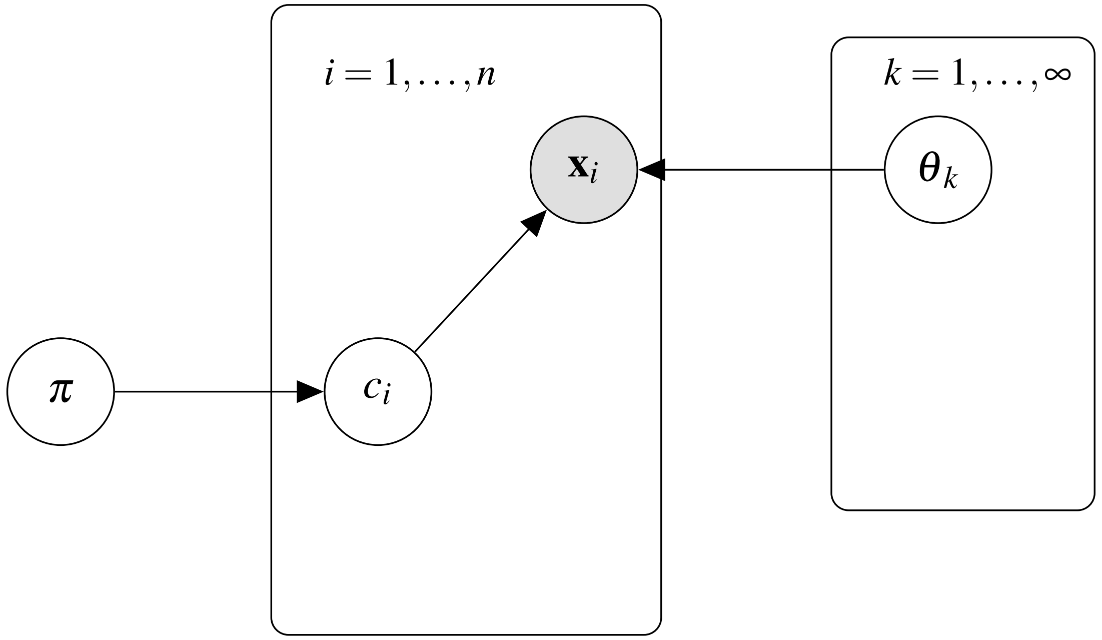


MIXTURE MODELS: INTRODUCTION AND NOTATION



$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n, c_1, \ldots, c_n, \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K)$$

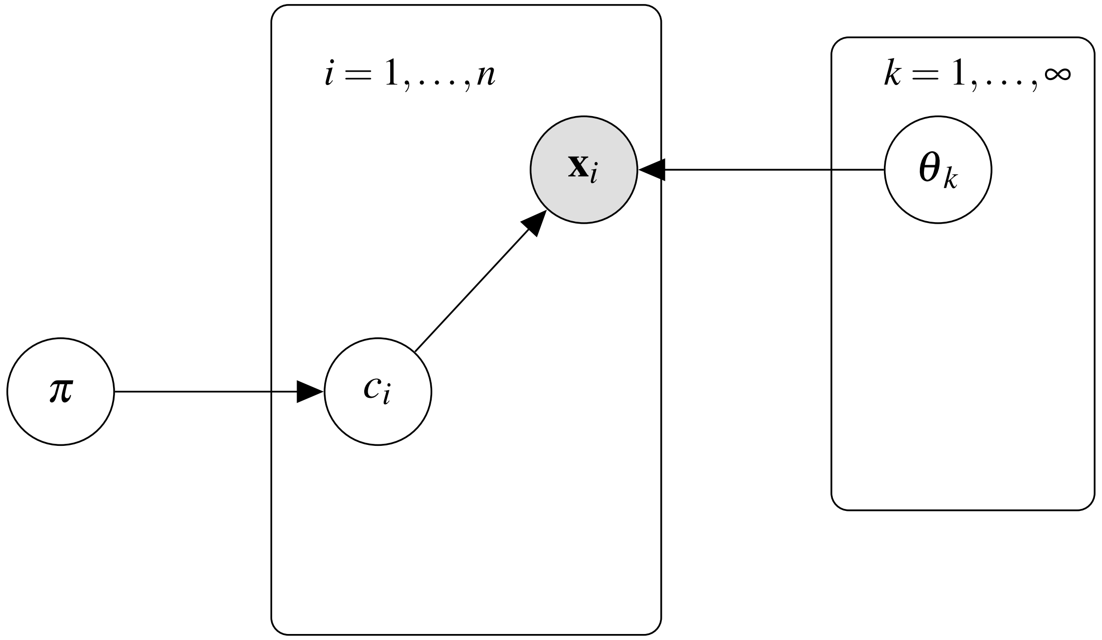
MIXTURE MODELS: INTRODUCTION AND NOTATION



$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n, c_1, \ldots, c_n, \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K)$$

$$= p(\mathbf{x}_1, \ldots, \mathbf{x}_n | c_1, \ldots, c_n, \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K) \times p(c_1, \ldots, c_n | \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K)$$
$$\times\, p(\theta_1, \ldots, \theta_K | \pi_1, \ldots, \pi_K) \times p(\pi_1, \ldots, \pi_K)$$

MIXTURE MODELS: INTRODUCTION AND NOTATION

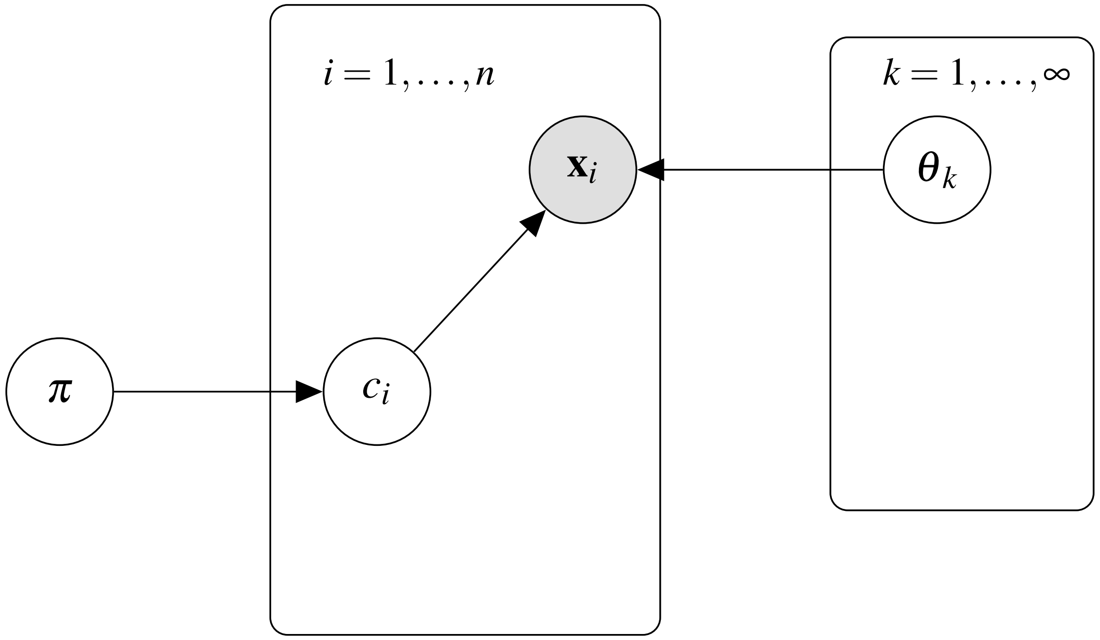


$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n, c_1, \ldots, c_n, \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K)$$

$$= p(\mathbf{x}_1, \ldots, \mathbf{x}_n | c_1, \ldots, c_n, \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K) \times p(c_1, \ldots, c_n | \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K)$$
$$\times p(\theta_1, \ldots, \theta_K | \pi_1, \ldots, \pi_K) \times p(\pi_1, \ldots, \pi_K)$$

$$= p(\mathbf{x}_1, \ldots, \mathbf{x}_n | c_1, \ldots, c_n, \theta_1, \ldots, \theta_K) p(c_1, \ldots, c_n | \pi_1, \ldots, \pi_K) p(\pi_1, \ldots, \pi_K) p(\theta_1, \ldots, \theta_K)$$

MIXTURE MODELS: INTRODUCTION AND NOTATION



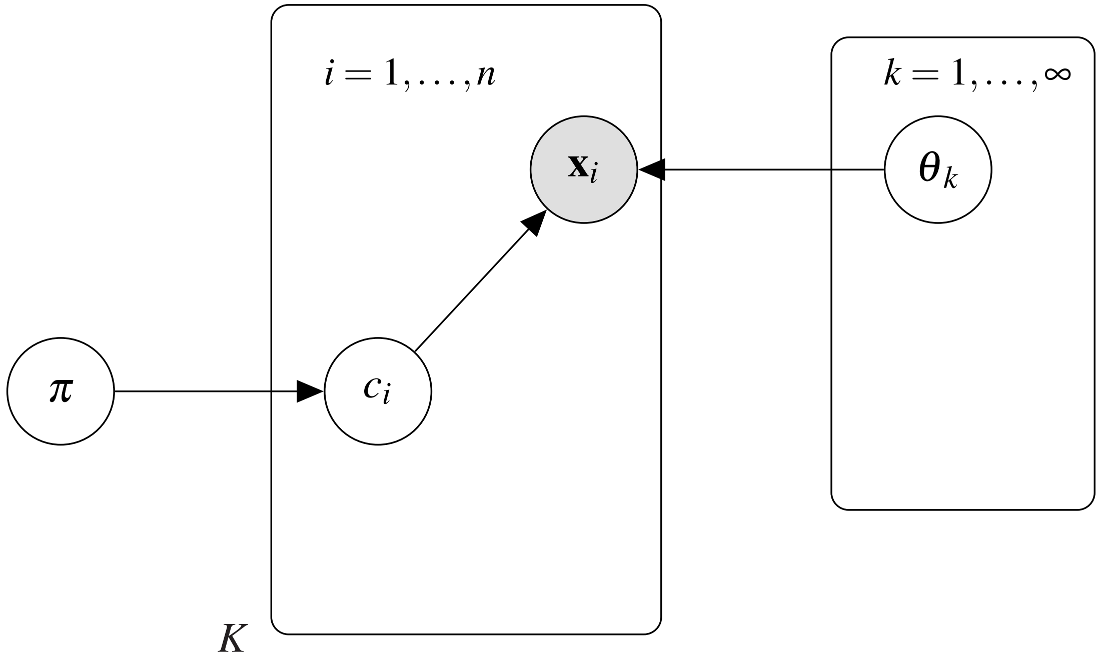
$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n, c_1, \ldots, c_n, \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K)$$

$$= p(\mathbf{x}_1, \ldots, \mathbf{x}_n | c_1, \ldots, c_n, \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K) \times p(c_1, \ldots, c_n | \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K) \\ \times p(\theta_1, \ldots, \theta_K | \pi_1, \ldots, \pi_K) \times p(\pi_1, \ldots, \pi_K)$$

$$= p(\mathbf{x}_1, \ldots, \mathbf{x}_n | c_1, \ldots, c_n, \theta_1, \ldots, \theta_K) p(c_1, \ldots, c_n | \pi_1, \ldots, \pi_K) p(\pi_1, \ldots, \pi_K) p(\theta_1, \ldots, \theta_K)$$

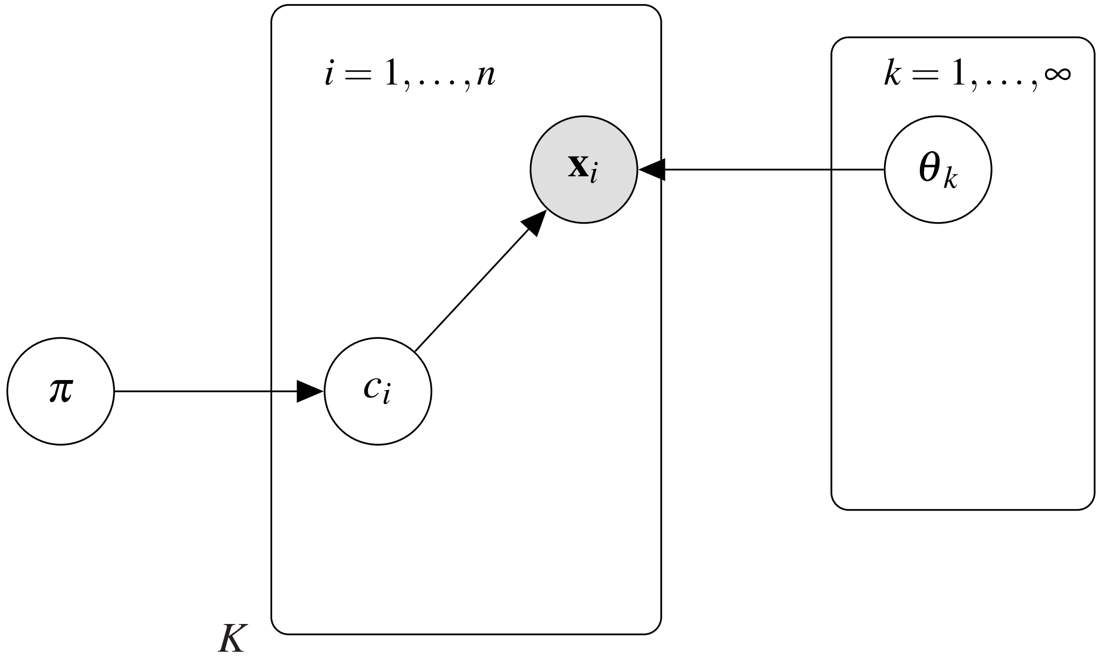
$$= \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) p(c_i | \pi_1, \ldots, \pi_K) \right) p(\pi_1, \ldots, \pi_K) \prod_{k=1}^{K} p(\theta_k)$$

MIXTURE MODELS: INTRODUCTION AND NOTATION



$$p(\mathbf{x}_1, \dots, \mathbf{x}_n, c_1, \dots, c_n, \theta_1, \dots, \theta_K, \pi_1, \dots, \pi_K)$$
$$= \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) p(c_i | \pi_1, \dots, \pi_K) \right) p(\pi_1, \dots, \pi_K) \prod_{k=1}^{K} p(\theta_k)$$

MIXTURE MODELS: INTRODUCTION AND NOTATION



$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n, c_1, \ldots, c_n, \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K) = \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) p(c_i | \pi_1, \ldots, \pi_K) \right) p(\pi_1, \ldots, \pi_K) \prod_{k=1}^{K} p(\theta_k)$$

CAN INFER THE UNKNOWNS VIA GIBBS SAMPLING

MIXTURE MODELS: INTRODUCTION AND NOTATION

$$\begin{aligned}
&p(\mathbf{x}_1, \ldots, \mathbf{x}_n, c_1, \ldots, c_n, \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K) \\
&\qquad = \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) p(c_i | \pi_1, \ldots, \pi_K) \right) p(\pi_1, \ldots, \pi_K) \prod_{k=1}^{K} p(\theta_k)
\end{aligned}$$

MIXTURE MODELS: INTRODUCTION AND NOTATION

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_n,c_1,\ldots,c_n,\theta_1,\ldots,\theta_K,\pi_1,\ldots,\pi_K)$$
$$= \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i|\theta_{c_i})p(c_i|\pi_1,\ldots,\pi_K)\right) p(\pi_1,\ldots,\pi_K)\prod_{k=1}^{K} p(\theta_k)$$

- If we take a conjugate prior for the θs, we can marginalise them.

MIXTURE MODELS: INTRODUCTION AND NOTATION

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_n,c_1,\ldots,c_n,\theta_1,\ldots,\theta_K,\pi_1,\ldots,\pi_K)$$
$$= \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i|\theta_{c_i})p(c_i|\pi_1,\ldots,\pi_K)\right) p(\pi_1,\ldots,\pi_K)\prod_{k=1}^{K} p(\theta_k)$$

- If we take a conjugate prior for the θs, we can marginalise them.
- We can also integrate out the πs.

MIXTURE MODELS: INTRODUCTION AND NOTATION

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n, c_1, \ldots, c_n, \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K)$$
$$= \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) p(c_i | \pi_1, \ldots, \pi_K) \right) p(\pi_1, \ldots, \pi_K) \prod_{k=1}^{K} p(\theta_k)$$

- If we take a conjugate prior for the θs, we can marginalise them.
- We can also integrate out the πs.
- We then just have to sample the component allocations
    - (can also do inference for precision parameter, α)

MIXTURE MODELS: INTRODUCTION AND NOTATION

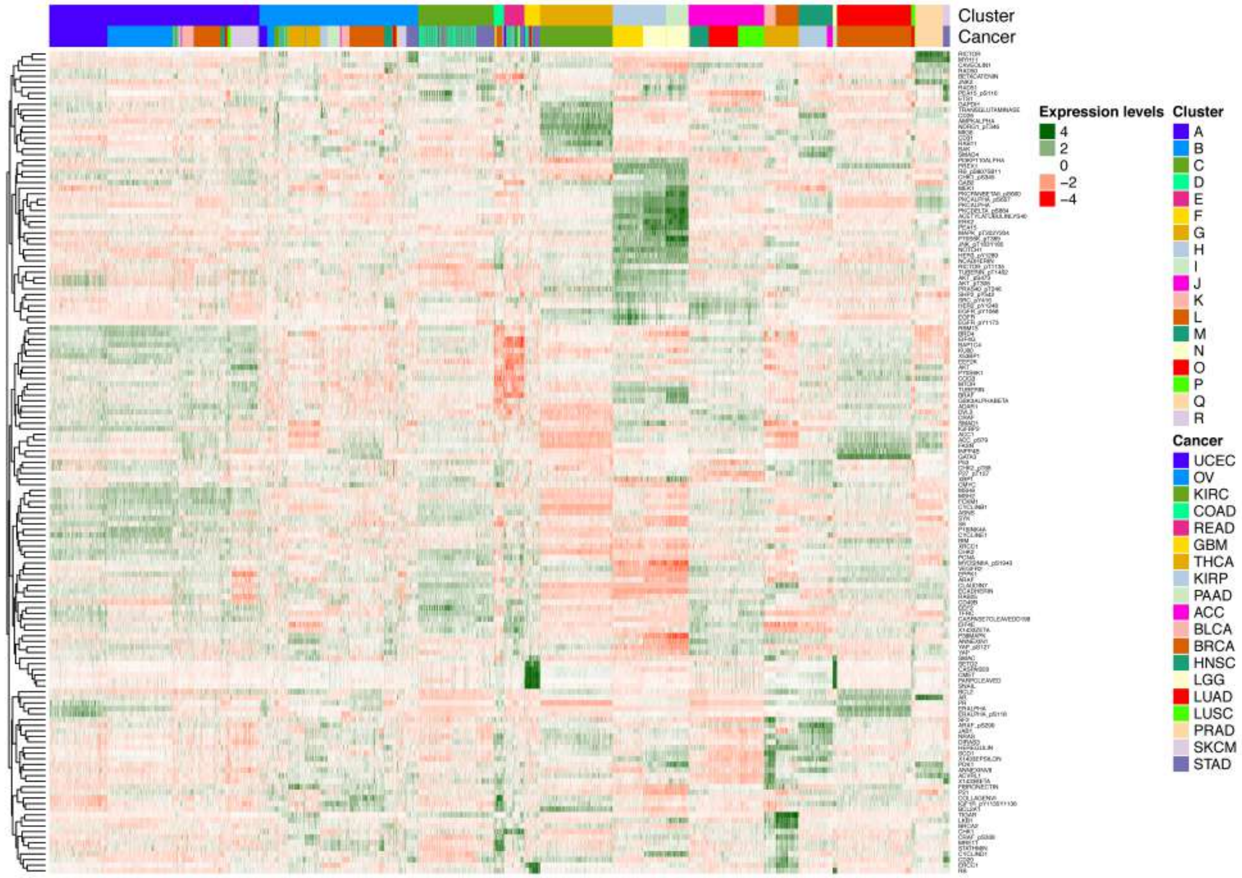
$$p(\mathbf{x}_1,\ldots,\mathbf{x}_n,c_1,\ldots,c_n,\theta_1,\ldots,\theta_K,\pi_1,\ldots,\pi_K) \\ = \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i|\theta_{c_i})p(c_i|\pi_1,\ldots,\pi_K)\right)p(\pi_1,\ldots,\pi_K)\prod_{k=1}^{K}p(\theta_k)$$

- If we take a conjugate prior for the θs, we can marginalise them.
- We can also integrate out the πs.
- We then just have to sample the component allocations
    - (can also do inference for precision parameter, α)
- For details of how to perform inference for DP mixtures, see:
    - Neal, R. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265.

AD BREAK

SUGSVARSEL: FAST APPROXIMATE INFERENCE FOR BAYESIAN CLUSTERING, WITH VARIABLE SELECTION

Oliver Crook



- Crook, Gatto & **Kirk** (2018), Fast approximate inference for variable selection in Dirichlet process mixtures, with an application to pan-cancer proteomics.
    - https://github.com/ococrook/sugsvarsel
- See also: Crook, Mulvey, **Kirk**, Lilley & Gatto (2018). A Bayesian Mixture Modelling Approach For Spatial Proteomics. bioRxiv.org. *(Accepted, PLOS Comput. Biol.)*

MIXTURE MODELS: PROFILE REGRESSION

MIXTURE MODELS: PROFILE REGRESSION

MIXTURE MODELS: PROFILE REGRESSION

Recall the basic mixture model:

$$p(\mathbf{x}) = \sum_{c=1}^{K} \pi_c f_{\mathbf{x}}(\mathbf{x}|\theta_c).$$

MIXTURE MODELS: PROFILE REGRESSION

Recall the basic mixture model:

$$p(\mathbf{x}) = \sum_{c=1}^{K} \pi_c f_{\mathbf{x}}(\mathbf{x}|\theta_c).$$

- We now suppose we have a response, y, associated with every individual.
  - This could be case/control status, survival information, . . . ,...

MIXTURE MODELS: PROFILE REGRESSION

Recall the basic mixture model:

$$p(\mathbf{x}) = \sum_{c=1}^{K} \pi_c f_{\mathbf{x}}(\mathbf{x}|\theta_c).$$

- We now suppose we have a response, y, associated with every individual.
  - This could be case/control status, survival information, . . . ,...
- In the simplest case, we just treat y like an extra covariate (or set of covariates), so that our mixture model becomes:

$$p([\mathbf{x}, y]) = \sum_{c=1}^{K} \pi_c f_{\mathbf{x},y}([\mathbf{x}, y]|[\theta_c, \phi_c]),$$

where ϕ_c denote any additional parameters required to model y.

MIXTURE MODELS: PROFILE REGRESSION

More usually, it makes sense to factorise the likelihood:

$$p([\mathbf{x}, y]) = \sum_{c=1}^{K} \pi_c f_{y|\mathbf{x}}(y|\phi_c, \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}|\theta_c),$$

MIXTURE MODELS: PROFILE REGRESSION

More usually, it makes sense to factorise the likelihood:

$$p([\mathbf{x}, y]) = \sum_{c=1}^{K} \pi_c f_{y|\mathbf{x}}(y|\phi_c, \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}|\theta_c),$$

- Note that $f_{y|\mathbf{x}}(y|\phi_c, \mathbf{x})$ can be basically any predictive model for y.

MIXTURE MODELS: PROFILE REGRESSION

More usually, it makes sense to factorise the likelihood:

$$p([\mathbf{x}, y]) = \sum_{c=1}^{K} \pi_c f_{y|\mathbf{x}}(y|\phi_c, \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}|\theta_c),$$

- Note that $f_{y|\mathbf{x}}(y|\phi_c, \mathbf{x})$ can be basically any predictive model for y.
- Key point: this predictive model has cluster specific parameters.

MIXTURE MODELS: PROFILE REGRESSION

More usually, it makes sense to factorise the likelihood:

$$p([\mathbf{x}, y]) = \sum_{c=1}^{K} \pi_c f_{y|\mathbf{x}}(y|\phi_c, \mathbf{x}) f_{\mathbf{x}}(\mathbf{x}|\theta_c),$$

- Note that $f_{y|\mathbf{x}}(y|\phi_c, \mathbf{x})$ can be basically any predictive model for y.
- Key point: this predictive model has cluster specific parameters.
- This is precisely what we want: it enables stratified predictions.

MIXTURE MODELS: PROFILE REGRESSION

The response y_i for the i-th individual might also depend upon:

- some other covariates, $\mathbf{w}_i$, that we are not clustering; and
- some “global” (i.e. not cluster-specific) parameters, β.

MIXTURE MODELS: PROFILE REGRESSION

The response y_i for the i-th individual might also depend upon:

- some other covariates, $\mathbf{w}_i$, that we are not clustering; and
- some "global" (i.e. not cluster-specific) parameters, β.

The most general profile regression model is then:

$$p([\mathbf{x}, y]) = \sum_{c=1}^{K} \pi_c f_{y|\mathbf{x}}(y|\phi_c, \mathbf{x}, \mathbf{w}, \beta) f_{\mathbf{x}}(\mathbf{x}|\theta_c),$$

MIXTURE MODELS: PROFILE REGRESSION

The response y_i for the i-th individual might also depend upon:

- some other covariates, $\mathbf{w}_i$, that we are not clustering; and
- some "global" (i.e. not cluster-specific) parameters, β.

The most general profile regression model is then:

$$p([\mathbf{x}, y]) = \sum_{c=1}^{K} \pi_c f_{y|\mathbf{x}}(y|\phi_c, \mathbf{x}, \mathbf{w}, \beta) f_{\mathbf{x}}(\mathbf{x}|\theta_c),$$

WHY "PROFILE" REGRESSION?

We have tended to assume that y depends on $\mathbf{x}$ only through the cluster assignment (i.e. y is conditionally independent of $\mathbf{x}$ given c), so that

$$p([\mathbf{x}, y]) = \sum_{c=1}^{K} \pi_c f_y(y|\phi_c, \mathbf{w}, \beta) f_{\mathbf{x}}(\mathbf{x}|\theta_c),$$

PLATE DIAGRAM: INITIAL BASIC CASE

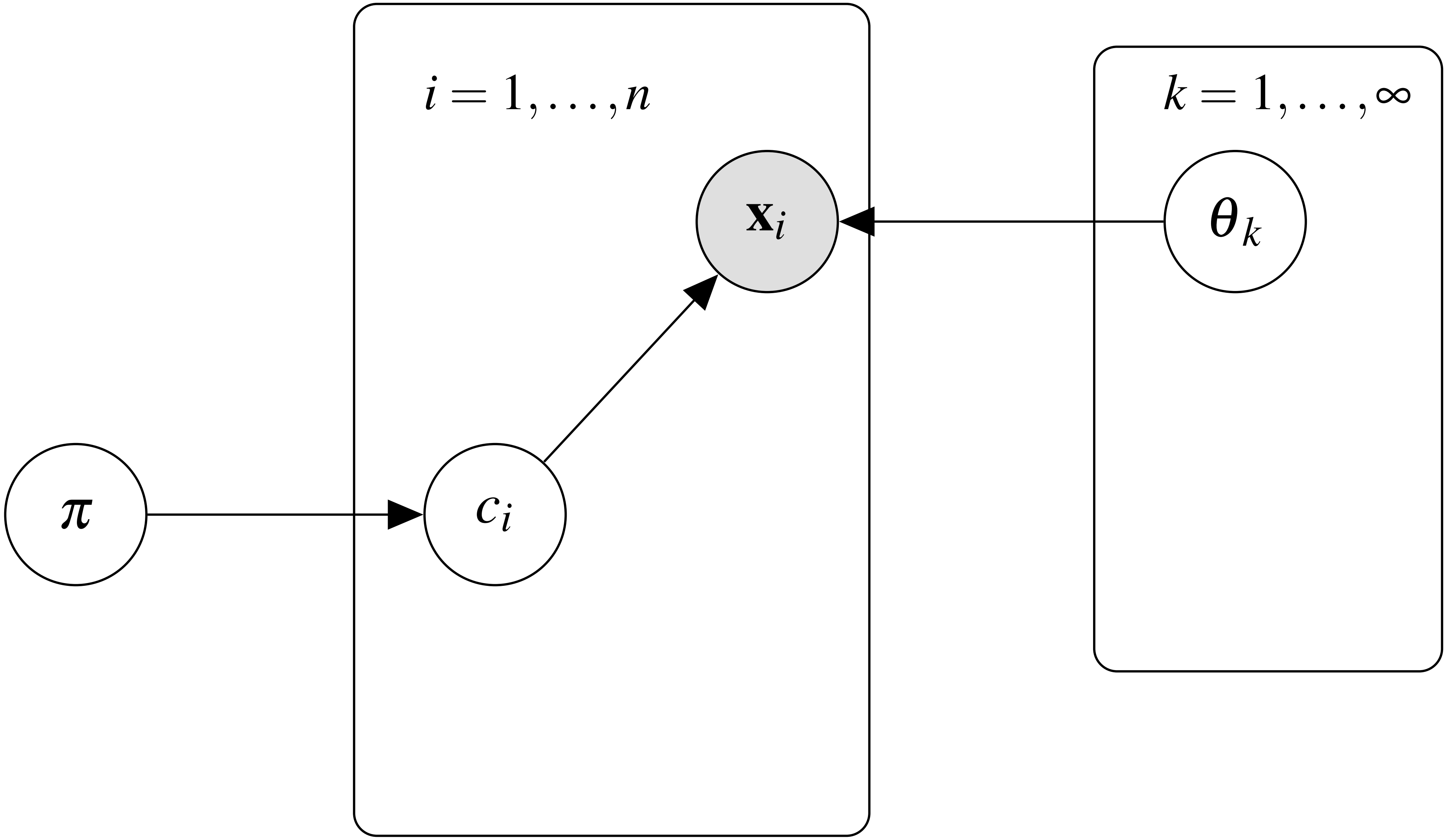
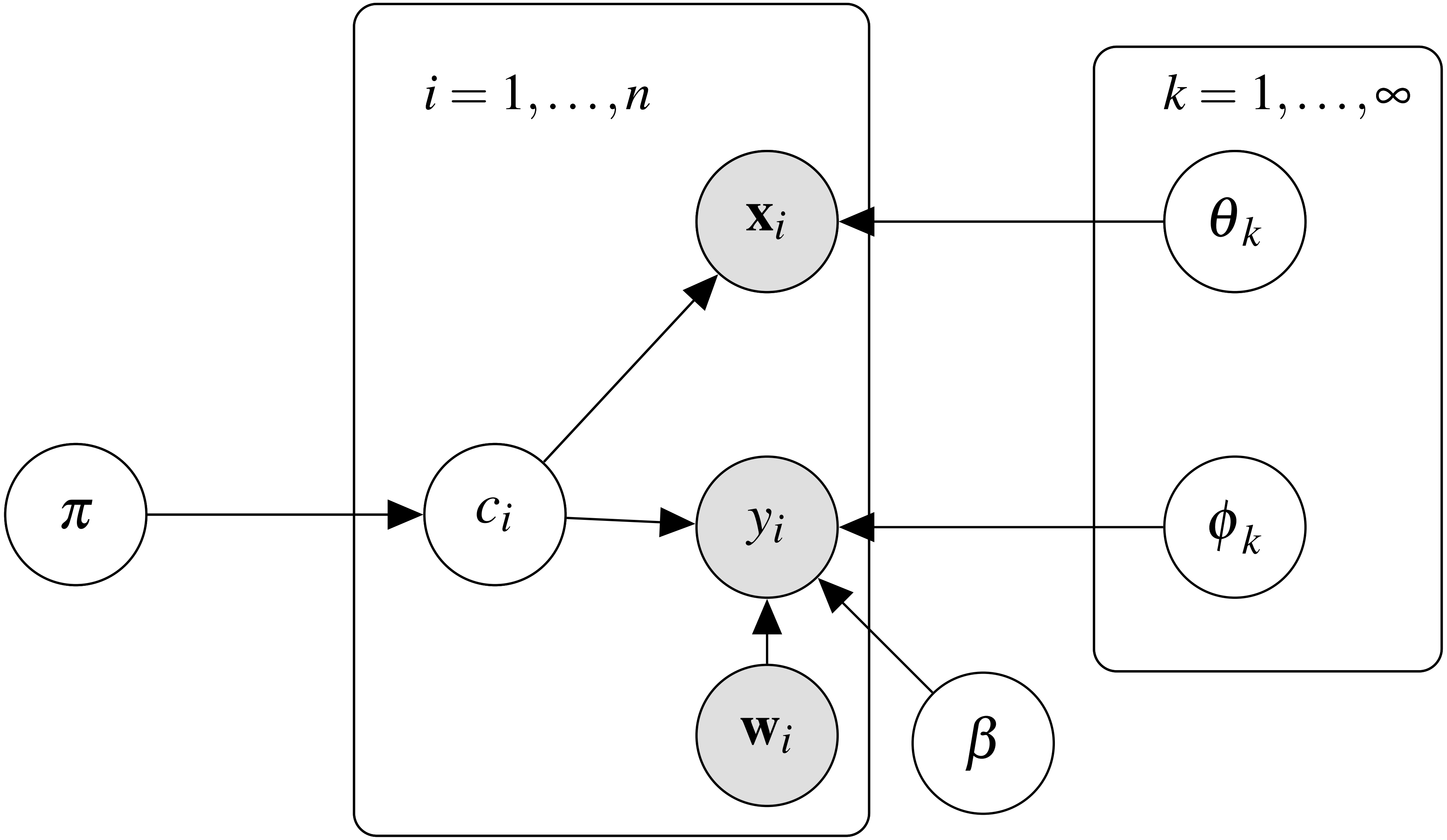
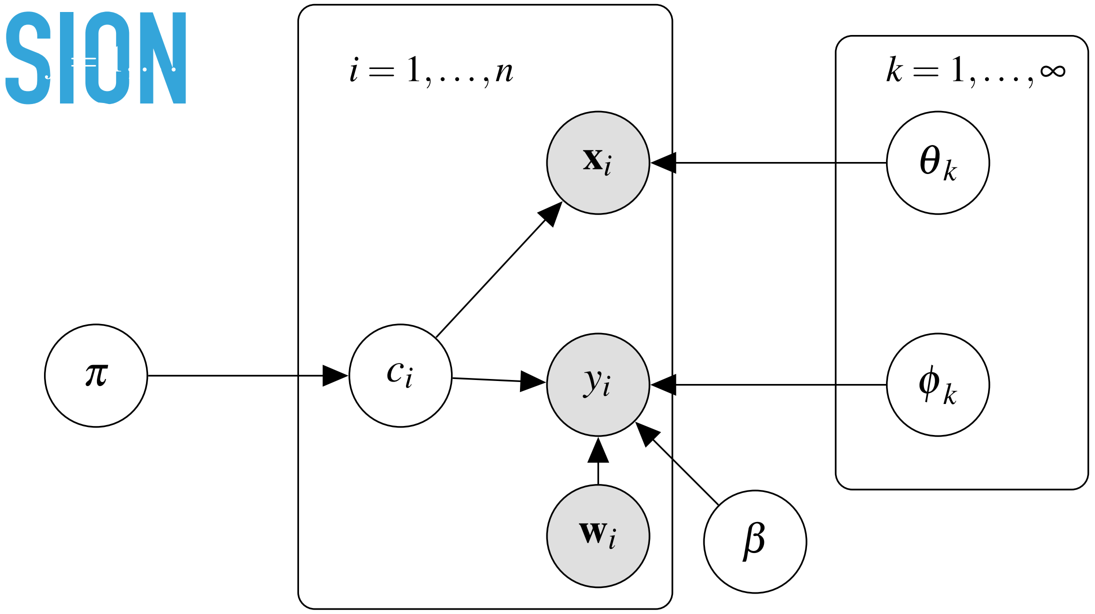


PLATE DIAGRAM: PROFILE REGRESSION



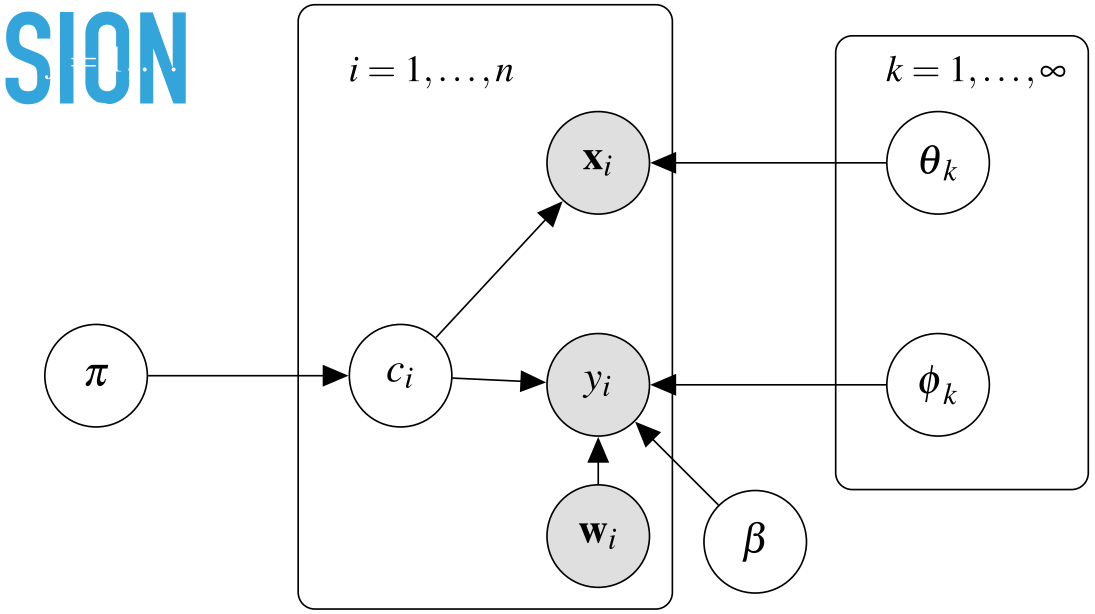
MIXTURE MODELS: PROFILE REGRESSION

THE JOINT MODEL



MIXTURE MODELS: PROFILE REGRESSION

THE JOINT MODEL

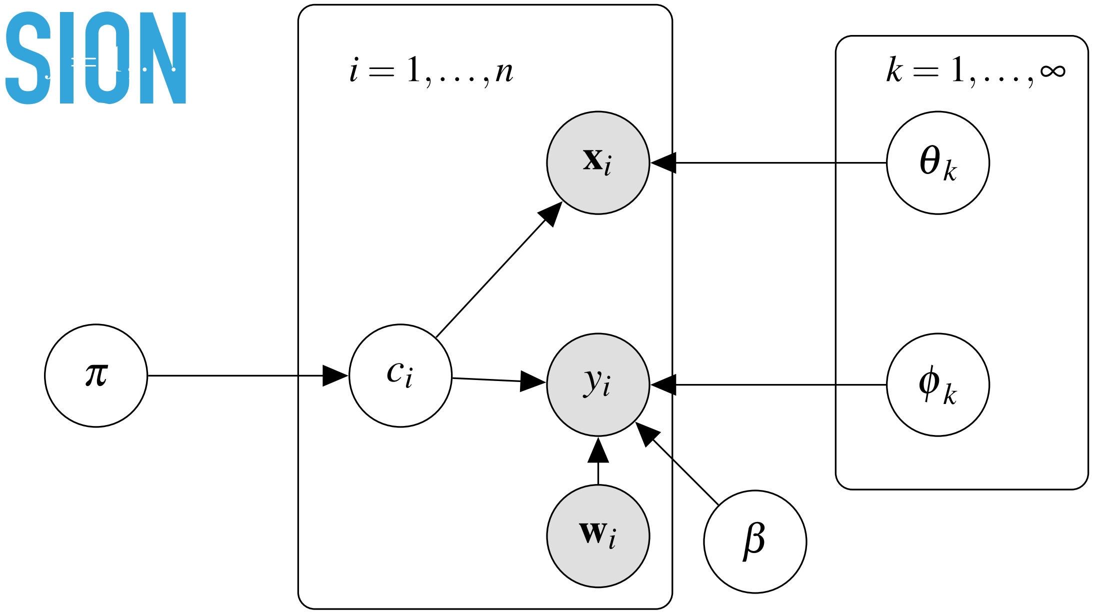


$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n, y_1, \ldots, y_n, c_1, \ldots, c_n, \theta_1, \ldots, \theta_K, \phi_1, \ldots, \phi_K, \pi_1, \ldots, \pi_K, \beta | \mathbf{w}_1, \ldots, \mathbf{w}_n)$$

$$= \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) p(c_i | \pi_1, \ldots, \pi_K) \right) p(\pi_1, \ldots, \pi_K) p(\beta) \left(\prod_{k=1}^{K} p(\phi_k) p(\theta_k) \right).$$

MIXTURE MODELS: PROFILE REGRESSION

THE JOINT MODEL

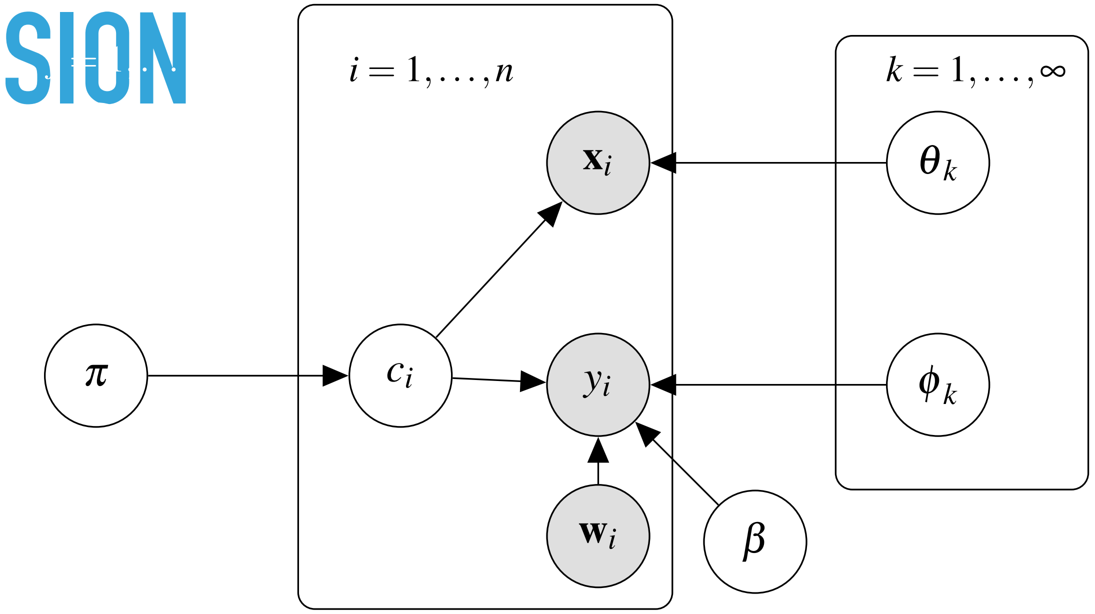


$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n, y_1, \ldots, y_n, c_1, \ldots, c_n, \theta_1, \ldots, \theta_K, \phi_1, \ldots, \phi_K, \pi_1, \ldots, \pi_K, \beta | \mathbf{w}_1, \ldots, \mathbf{w}_n)$$
$$= \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) p(c_i | \pi_1, \ldots, \pi_K) \right) p(\pi_1, \ldots, \pi_K) p(\beta) \left(\prod_{k=1}^{K} p(\phi_k) p(\theta_k) \right).$$

CAN AGAIN PERFORM INFERENCE VIA MCMC

MIXTURE MODELS: PROFILE REGRESSION

THE JOINT MODEL



$$
\begin{aligned}
&p(\mathbf{x}_1, \dots, \mathbf{x}_n, y_1, \dots, y_n, c_1, \dots, c_n, \theta_1, \dots, \theta_K, \phi_1, \dots, \phi_K, \pi_1, \dots, \pi_K, \beta | \mathbf{w}_1, \dots, \mathbf{w}_n) \\
&\quad = \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) p(c_i | \pi_1, \dots, \pi_K) \right) p(\pi_1, \dots, \pi_K) p(\beta) \left(\prod_{k=1}^{K} p(\phi_k) p(\theta_k) \right).
\end{aligned}
$$

CAN AGAIN PERFORM INFERENCE VIA MCMC

BUT LET'S MOVE ON TO VARIABLE SELECTION

PROFILE REGRESSION: VARIABLE SELECTION

PROFILE REGRESSION: VARIABLE SELECTION

PROFILE REGRESSION: VARIABLE SELECTION

RECALL THE JOINT MODEL:

PROFILE REGRESSION: VARIABLE SELECTION

RECALL THE JOINT MODEL:

$$
\begin{aligned}
p(\mathbf{x}_1,\ldots,\mathbf{x}_n, y_1,\ldots,y_n, c_1,\ldots,c_n, \theta_1,\ldots,\theta_K, \phi_1,\ldots,\phi_K, \pi_1,\ldots,\pi_K, \beta \,|\, \mathbf{w}_1,\ldots,\mathbf{w}_n) \\
= \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i|\theta_{c_i}) f_{\mathbf{y}}(y_i|\phi_{c_i}, \mathbf{w}_i, \beta) p(c_i|\pi_1,\ldots,\pi_K)\right) p(\pi_1,\ldots,\pi_K) p(\beta) \left(\prod_{k=1}^{K} p(\phi_k) p(\theta_k)\right).
\end{aligned}
$$

PROFILE REGRESSION: VARIABLE SELECTION

RECALL THE JOINT MODEL:

$$
\begin{aligned}
p(\mathbf{x}_1,\ldots,\mathbf{x}_n,y_1,\ldots,y_n,c_1,\ldots,c_n,\theta_1,\ldots,\theta_K,\phi_1,\ldots,\phi_K,\pi_1,\ldots,\pi_K,\beta|\mathbf{w}_1,\ldots,\mathbf{w}_n) \\
= \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i|\theta_{c_i}) f_{\mathbf{y}}(y_i|\phi_{c_i},\mathbf{w}_i,\beta) p(c_i|\pi_1,\ldots,\pi_K)\right) p(\pi_1,\ldots,\pi_K)p(\beta)\left(\prod_{k=1}^{K} p(\phi_k)p(\theta_k)\right).
\end{aligned}
$$

- Let’s focus on the highlighted term

PROFILE REGRESSION: VARIABLE SELECTION

RECALL THE JOINT MODEL:

$$p(\mathbf{x}_1,\dots,\mathbf{x}_n, y_1,\dots,y_n, c_1,\dots,c_n, \theta_1,\dots,\theta_K, \phi_1,\dots,\phi_K, \pi_1,\dots,\pi_K, \beta|\mathbf{w}_1,\dots,\mathbf{w}_n)$$
$$= \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i|\theta_{c_i}) f_{\mathbf{y}}(y_i|\phi_{c_i}, \mathbf{w}_i, \beta) p(c_i|\pi_1,\dots,\pi_K)\right) p(\pi_1,\dots,\pi_K) p(\beta) \left(\prod_{k=1}^{K} p(\phi_k) p(\theta_k)\right).$$

- Let's focus on the highlighted term
- This is just the likelihood associated with $\mathbf{x}_i$ and y_i

$$f(\mathbf{x}_i, y_i|\cdots) = f_{\mathbf{x}}(\mathbf{x}_i|\theta_{c_i}) f_{\mathbf{y}}(y_i|\phi_{c_i}, \mathbf{w}_i, \beta)$$

PROFILE REGRESSION: VARIABLE SELECTION

RECALL THE JOINT MODEL:

$$\begin{aligned} p(\mathbf{x}_1,\ldots,\mathbf{x}_n,y_1,\ldots,y_n,c_1,\ldots,c_n,\theta_1,\ldots,\theta_K,\phi_1,\ldots,\phi_K,\pi_1,\ldots,\pi_K,\beta|\mathbf{w}_1,\ldots,\mathbf{w}_n) \\ = \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i|\theta_{c_i})f_{\mathbf{y}}(y_i|\phi_{c_i},\mathbf{w}_i,\beta)p(c_i|\pi_1,\ldots,\pi_K)\right)p(\pi_1,\ldots,\pi_K)p(\beta)\left(\prod_{k=1}^{K}p(\phi_k)p(\theta_k)\right).\end{aligned}$$

- Let's focus on the highlighted term
- This is just the likelihood associated with $\mathbf{x}_i$ and y_i
- Let's assume that the variables (elements of $\mathbf{x}_i$) are conditionally independent, given the component allocation, c_i

$$f(\mathbf{x}_i,y_i|\cdots) = f_{\mathbf{x}}(\mathbf{x}_i|\theta_{c_i})f_{\mathbf{y}}(y_i|\phi_{c_i},\mathbf{w}_i,\beta)$$

PROFILE REGRESSION: VARIABLE SELECTION

RECALL THE JOINT MODEL:

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_n,y_1,\ldots,y_n,c_1,\ldots,c_n,\theta_1,\ldots,\theta_K,\phi_1,\ldots,\phi_K,\pi_1,\ldots,\pi_K,\beta|\mathbf{w}_1,\ldots,\mathbf{w}_n)$$
$$= \left(\prod_{i=1}^{n} f_{\mathbf{x}}(\mathbf{x}_i|\theta_{c_i}) f_{\mathbf{y}}(y_i|\phi_{c_i},\mathbf{w}_i,\beta) p(c_i|\pi_1,\ldots,\pi_K)\right) p(\pi_1,\ldots,\pi_K)p(\beta)\left(\prod_{k=1}^{K} p(\phi_k)p(\theta_k)\right).$$

- Let's focus on the highlighted term
- This is just the likelihood associated with $\mathbf{x}_i$ and y_i
- Let's assume that the variables (elements of $\mathbf{x}_i$) are conditionally independent, given the component allocation, c_i

$$\begin{aligned} f(\mathbf{x}_i, y_i|\cdots) &= f_{\mathbf{x}}(\mathbf{x}_i|\theta_{c_i}) f_{\mathbf{y}}(y_i|\phi_{c_i},\mathbf{w}_i,\beta) \\ &= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij}|\theta_{c_i})\right) f_{\mathbf{y}}(y_i|\phi_{c_i},\mathbf{w}_i,\beta) \end{aligned}$$

PROFILE REGRESSION: VARIABLE SELECTION

$$\begin{aligned} f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \\ &= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij} | \theta_{c_i}) \right) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \end{aligned}$$

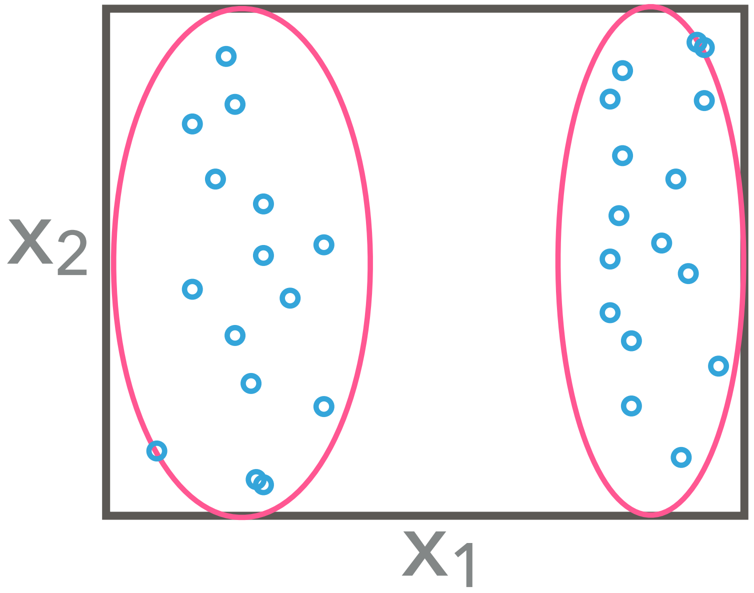
PROFILE REGRESSION: VARIABLE SELECTION

$$\begin{aligned} f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \\ &= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij} | \theta_{c_i}) \right) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \end{aligned}$$

- Not all of these variables will be informative about the clustering structure

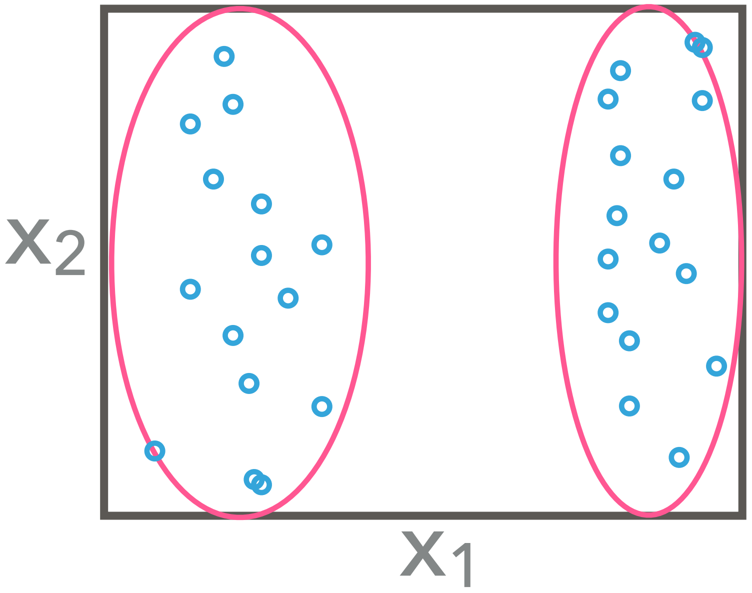
PROFILE REGRESSION: VARIABLE SELECTION

$$
\begin{aligned}
f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \\
&= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij} | \theta_{c_i}) \right) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta)
\end{aligned}
$$



- Not all of these variables will be informative about the clustering structure

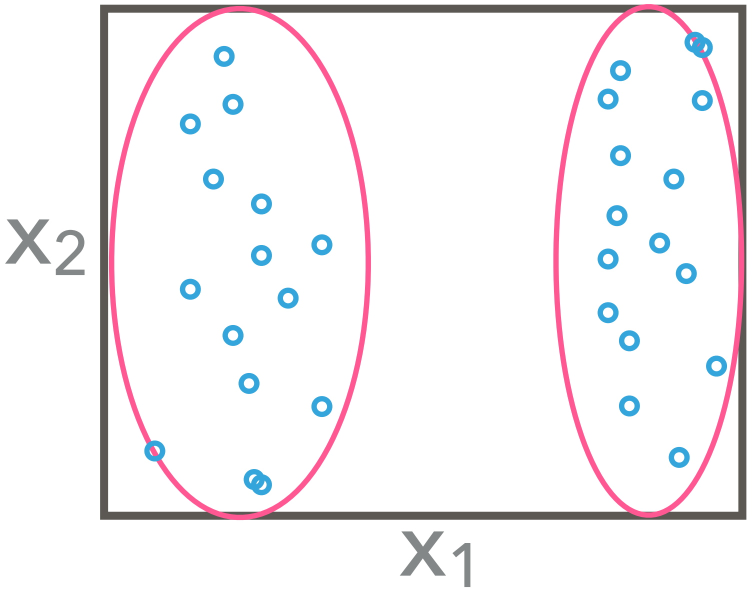
PROFILE REGRESSION: VARIABLE SELECTION



$$\begin{aligned} f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \\ &= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij} | \theta_{c_i}) \right) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \end{aligned}$$

- Not all of these variables will be informative about the clustering structure
- We introduce binary variable relevance indicators γ_j , such that:

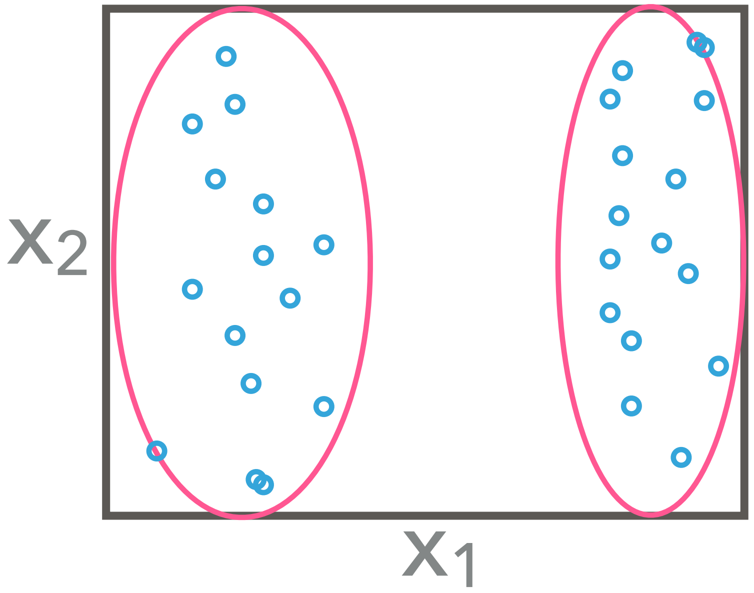
PROFILE REGRESSION: VARIABLE SELECTION



$$
\begin{aligned}
f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \\
&= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij} | \theta_{c_i}) \right) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta)
\end{aligned}
$$

- Not all of these variables will be informative about the clustering structure
- We introduce binary variable relevance indicators γ_j , such that:
  - $\gamma_j = 1$ if the j-th variable is relevant; and
  - $\gamma_j = 0$ if the j-th variable is irrelevant

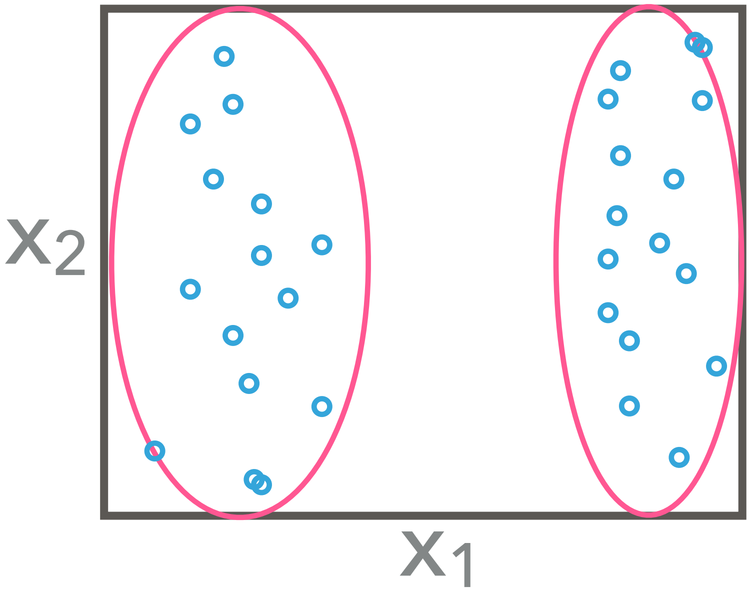
PROFILE REGRESSION: VARIABLE SELECTION



$$
\begin{aligned}
f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \\
&= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij} | \theta_{c_i}) \right) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta)
\end{aligned}
$$

- Not all of these variables will be informative about the clustering structure
- We introduce binary variable relevance indicators γ_j , such that:
  - $\gamma_j = 1$ if the j-th variable is relevant; and
  - $\gamma_j = 0$ if the j-th variable is irrelevant
- $\gamma_j = 0$ if and only if $\mathbf{x}_i$ is independent of c_i

PROFILE REGRESSION: VARIABLE SELECTION



$$
\begin{aligned}
f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i|\theta_{c_i}) f_{\mathbf{y}}(y_i|\phi_{c_i}, \mathbf{w}_i, \beta) \\
&= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij}|\theta_{c_i}) \right) f_{\mathbf{y}}(y_i|\phi_{c_i}, \mathbf{w}_i, \beta)
\end{aligned}
$$

- Not all of these variables will be informative about the clustering structure
- We introduce binary variable relevance indicators γ_j , such that:
  - $\gamma_j = 1$ if the j-th variable is relevant; and
  - $\gamma_j = 0$ if the j-th variable is irrelevant
- $\gamma_j = 0$ if and only if $\mathbf{x}_i$ is independent of c_i
- We modify our likelihood function to allow us to model this

PROFILE REGRESSION: VARIABLE SELECTION

$$\begin{aligned} f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \\ &= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij} | \theta_{c_i}) \right) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \end{aligned}$$

PROFILE REGRESSION: VARIABLE SELECTION

$$\begin{aligned} f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \\ &= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij} | \theta_{c_i}) \qquad\qquad\qquad\qquad\qquad\qquad \right) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \end{aligned}$$

PROFILE REGRESSION: VARIABLE SELECTION

$$
\begin{aligned}
f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \\
&= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij} | \theta_{c_i})^{\mathbb{I}(\gamma_j = 1)} \qquad\qquad\qquad\qquad \right) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta)
\end{aligned}
$$

PROFILE REGRESSION: VARIABLE SELECTION

$$
\begin{aligned}
f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \\
&= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij} | \theta_{c_i})^{\mathbb{I}(\gamma_j = 1)} f_{\mathbf{x}}(x_{ij} | \theta_0)^{\mathbb{I}(\gamma_j = 0)} \right) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta)
\end{aligned}
$$

PROFILE REGRESSION: VARIABLE SELECTION

$$
\begin{aligned}
f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i|\theta_{c_i}) f_{\mathbf{y}}(y_i|\phi_{c_i}, \mathbf{w}_i, \beta) \\
&= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij}|\theta_{c_i})^{\mathbb{I}(\gamma_j = 1)} f_{\mathbf{x}}(x_{ij}|\theta_0)^{\mathbb{I}(\gamma_j = 0)} \right) f_{\mathbf{y}}(y_i|\phi_{c_i}, \mathbf{w}_i, \beta)
\end{aligned}
$$

INFERRING THE VARIABLE RELEVANCE INDICATORS:

PROFILE REGRESSION: VARIABLE SELECTION

$$
\begin{aligned}
f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \\
&= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij} | \theta_{c_i})^{\mathbb{I}(\gamma_j = 1)} f_{\mathbf{x}}(x_{ij} | \theta_0)^{\mathbb{I}(\gamma_j = 0)} \right) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta)
\end{aligned}
$$

INFERRING THE VARIABLE RELEVANCE INDICATORS:

- In practice, we have to infer the variable relevance indicators, γ_j

PROFILE REGRESSION: VARIABLE SELECTION

$$
\begin{aligned}
f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \\
&= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij} | \theta_{c_i})^{\mathbb{I}(\gamma_j = 1)} f_{\mathbf{x}}(x_{ij} | \theta_0)^{\mathbb{I}(\gamma_j = 0)} \right) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta)
\end{aligned}
$$

INFERRING THE VARIABLE RELEVANCE INDICATORS:

- In practice, we have to infer the variable relevance indicators, γ_j
- The conditional is straightforward to right down:

PROFILE REGRESSION: VARIABLE SELECTION

$$
\begin{aligned}
f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \\
&= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij} | \theta_{c_i})^{\mathbb{I}(\gamma_j = 1)} f_{\mathbf{x}}(x_{ij} | \theta_0)^{\mathbb{I}(\gamma_j = 0)} \right) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta)
\end{aligned}
$$

INFERRING THE VARIABLE RELEVANCE INDICATORS:

- In practice, we have to infer the variable relevance indicators, γ_j
- The conditional is straightforward to right down:

$$
\begin{aligned}
p(\gamma_j = 1 | \cdots) &\propto p(\gamma_j = 1) \prod_{i=1}^{n} f_{\mathbf{x}}(x_{ij} | \theta_{c_i}) \\
p(\gamma_j = 0 | \cdots) &\propto p(\gamma_j = 0) \prod_{i=1}^{n} f_{\mathbf{x}}(x_{ij} | \theta_0)
\end{aligned}
$$

PROFILE REGRESSION: VARIABLE SELECTION

$$\begin{aligned} f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \\ &= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij} | \theta_{c_i})^{\mathbb{I}(\gamma_j = 1)} f_{\mathbf{x}}(x_{ij} | \theta_0)^{\mathbb{I}(\gamma_j = 0)} \right) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \end{aligned}$$

INFERRING THE VARIABLE RELEVANCE INDICATORS:

- In practice, we have to infer the variable relevance indicators, γ_j
- The conditional is straightforward to right down:

$$p(\gamma_j = 1 | \cdots) = \frac{p(\gamma_j = 1) \prod_{i=1}^{n} f_{\mathbf{x}}(x_{ij} | \theta_{c_i})}{p(\gamma_j = 1) \prod_{i=1}^{n} f_{\mathbf{x}}(x_{ij} | \theta_{c_i}) + p(\gamma_j = 0) \prod_{i=1}^{n} f_{\mathbf{x}}(x_{ij} | \theta_0)}$$

$$p(\gamma_j = 0 | \cdots) = \frac{p(\gamma_j = 0) \prod_{i=1}^{n} f_{\mathbf{x}}(x_{ij} | \theta_0)}{p(\gamma_j = 1) \prod_{i=1}^{n} f_{\mathbf{x}}(x_{ij} | \theta_{c_i}) + p(\gamma_j = 0) \prod_{i=1}^{n} f_{\mathbf{x}}(x_{ij} | \theta_0)}$$

PLATE DIAGRAM: INITIAL BASIC CASE

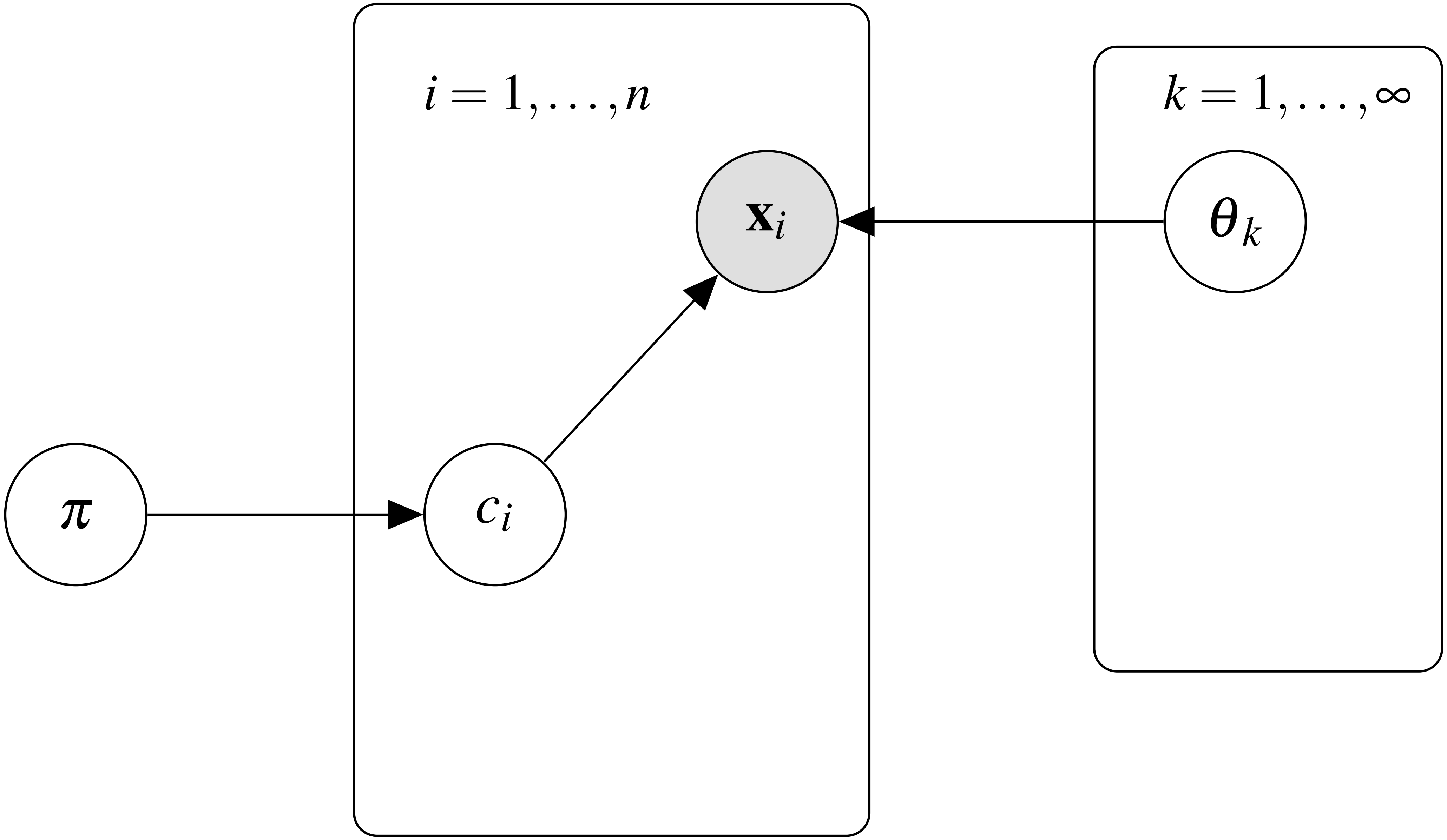


PLATE DIAGRAM: PROFILE REGRESSION

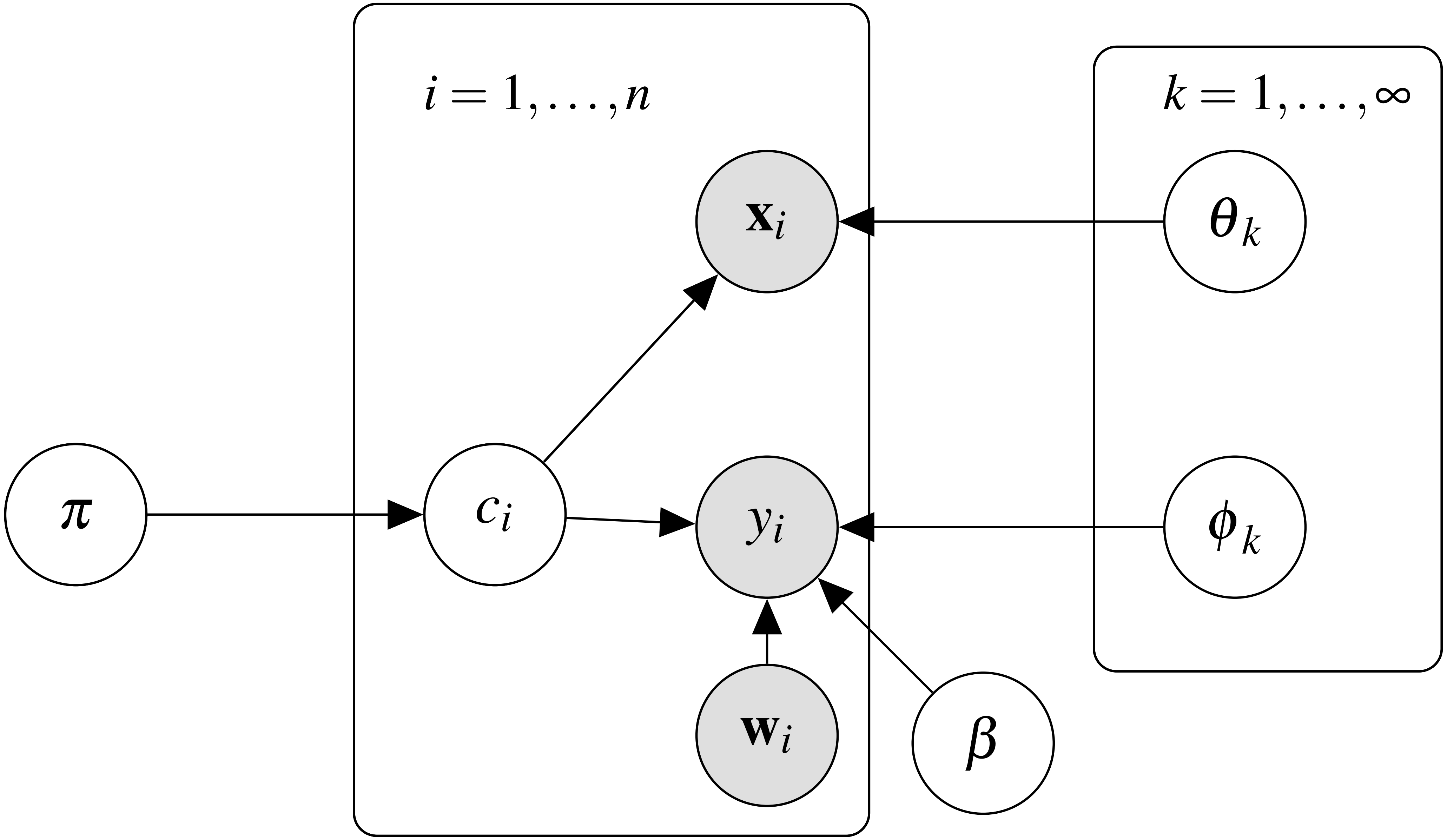
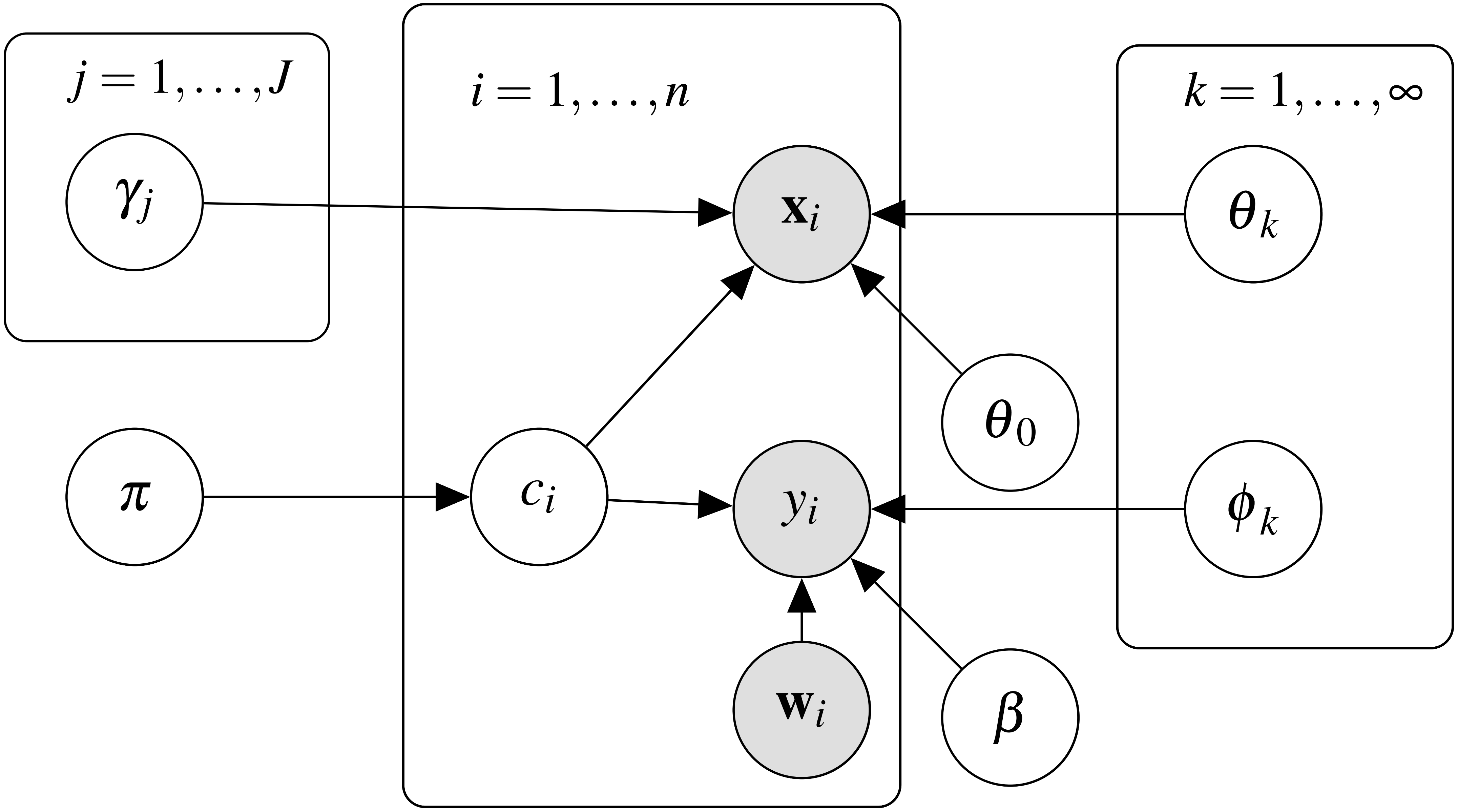


PLATE DIAGRAM: VARIABLE SELECTION



PROFILE REGRESSION WITH VARIABLE SELECTION:

REFERENCES:

- Molitor, J., Papathomas, M., Jerrett, M., & Richardson, S. (2010). Bayesian profile regression with an application to the National Survey of Children's Health. Biostatistics (Oxford, England), 11(3), 484–498.

- Papathomas, M., Molitor, J., Richardson, S., Riboli, E., & Vineis, P. (2011). Examining the joint effect of multiple risk factors using exposure risk profiles: lung cancer in nonsmokers. Environmental Health Perspectives, 119(1), 84–91.

- Papathomas, M., Molitor, J., Hoggart, C., Hastie, D., & 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(6), 663–674.

- Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M., & Richardson, S. (2015). PReMiuM: An R Package for Profile Regression Mixture Models Using Dirichlet Processes. Journal of Statistical Software, 64(7).

PROFILE REGRESSION: SIMULATION STUDY

- 6 equally-sized clusters
- y: 1 binary response
- **x**: 10 categorical variables (each with 3 categories)
- For the k-th cluster, we have:

$$x_{i,j}|k \overset{iid}{\sim} w\text{Categorical}([\pi^{(k)}_{j,1}, \pi^{(k)}_{j,2}, \pi^{(k)}_{j,3}]^\top) + (1-w)\text{Categorical}(\left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]^\top),$$

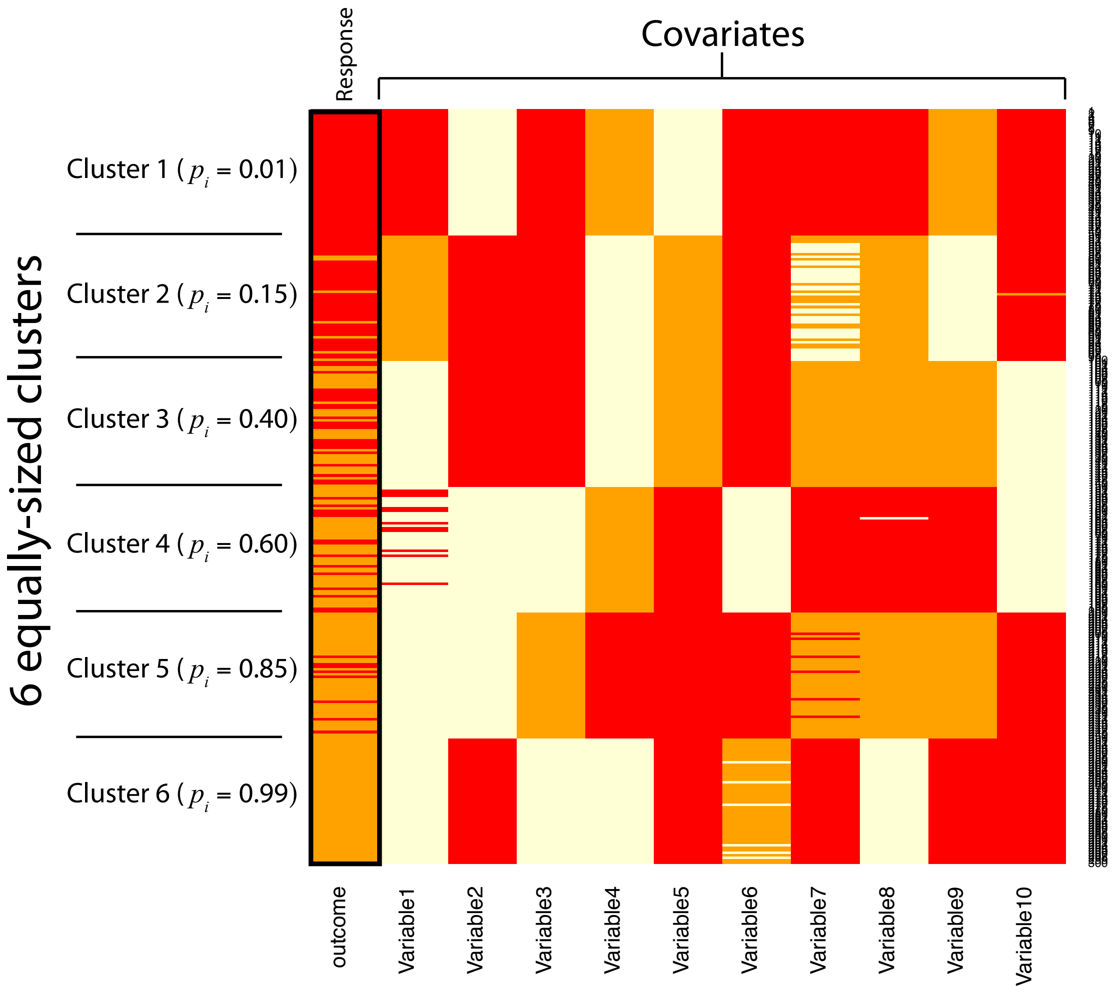
where

$$\pi^{(k)}_{j,1}, \pi^{(k)}_{j,2}, \pi^{(k)}_{j,3} \sim \text{Dirichlet}(0.01),$$

and $w \in [0,1]$.

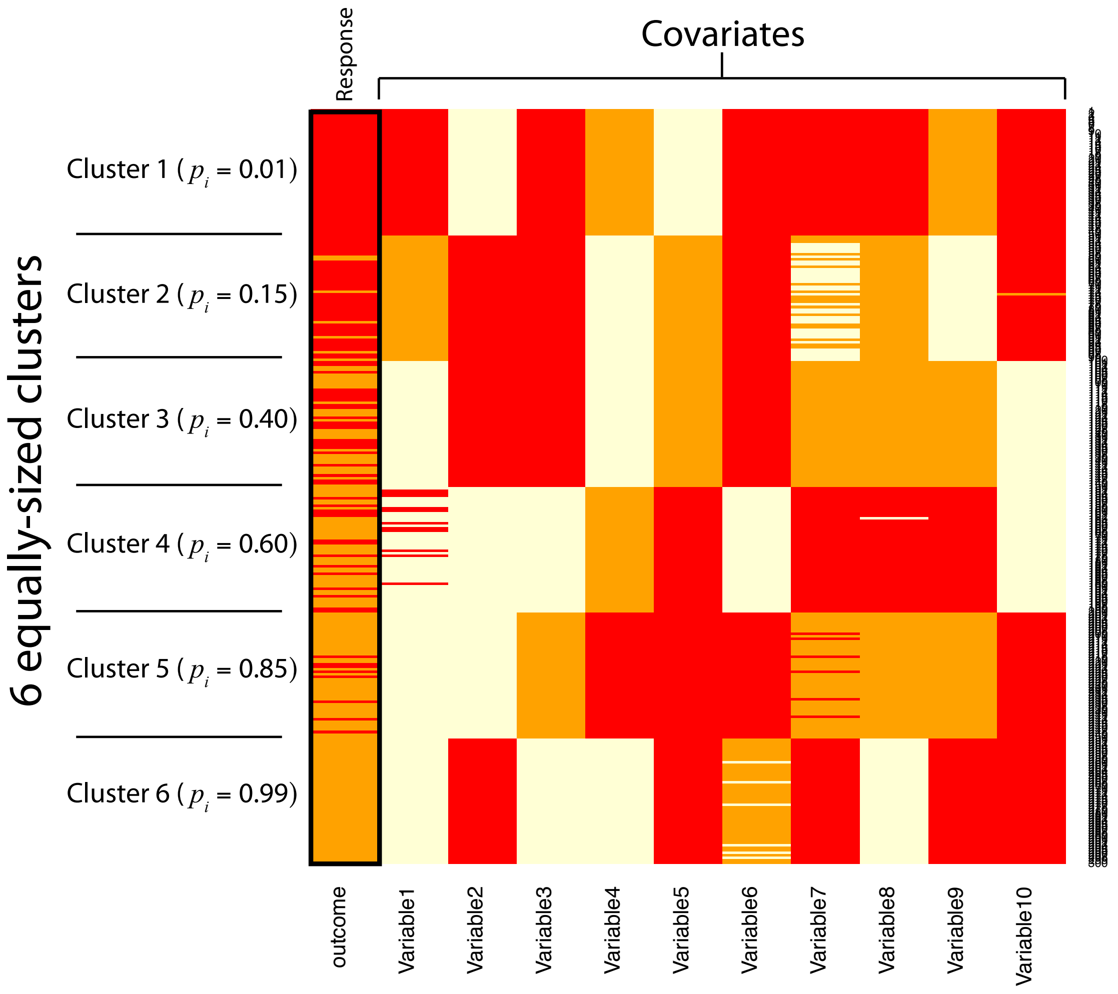
- Thus, w controls how separable the clusters are (w = 0 implies not separable)

PROFILE REGRESSION: SIMULATION STUDY

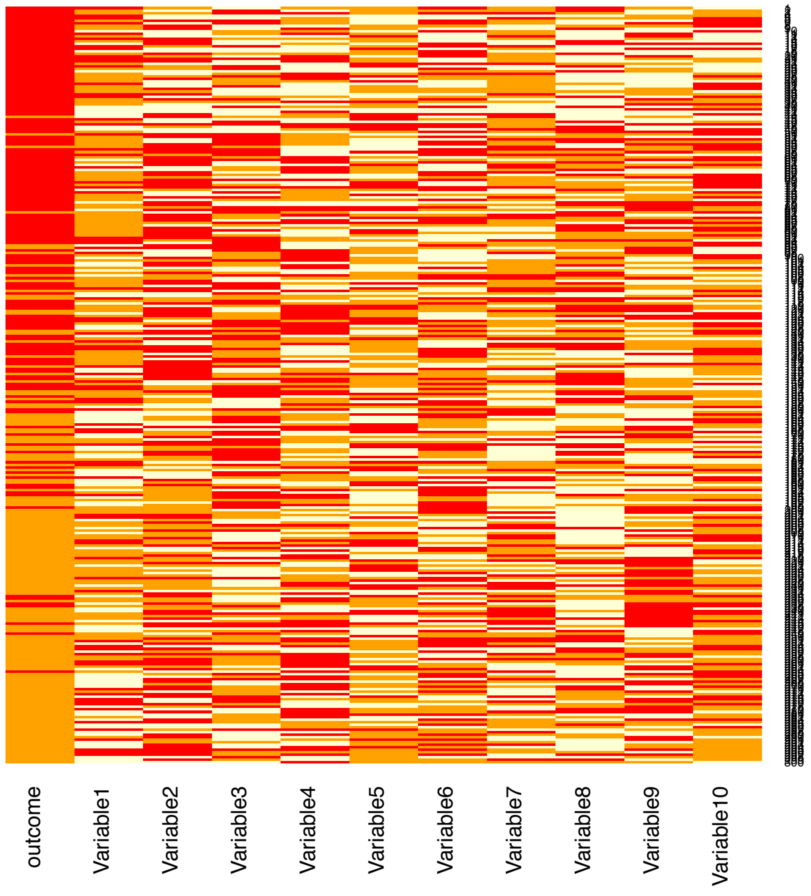


Large w

PROFILE REGRESSION: SIMULATION STUDY



Large w



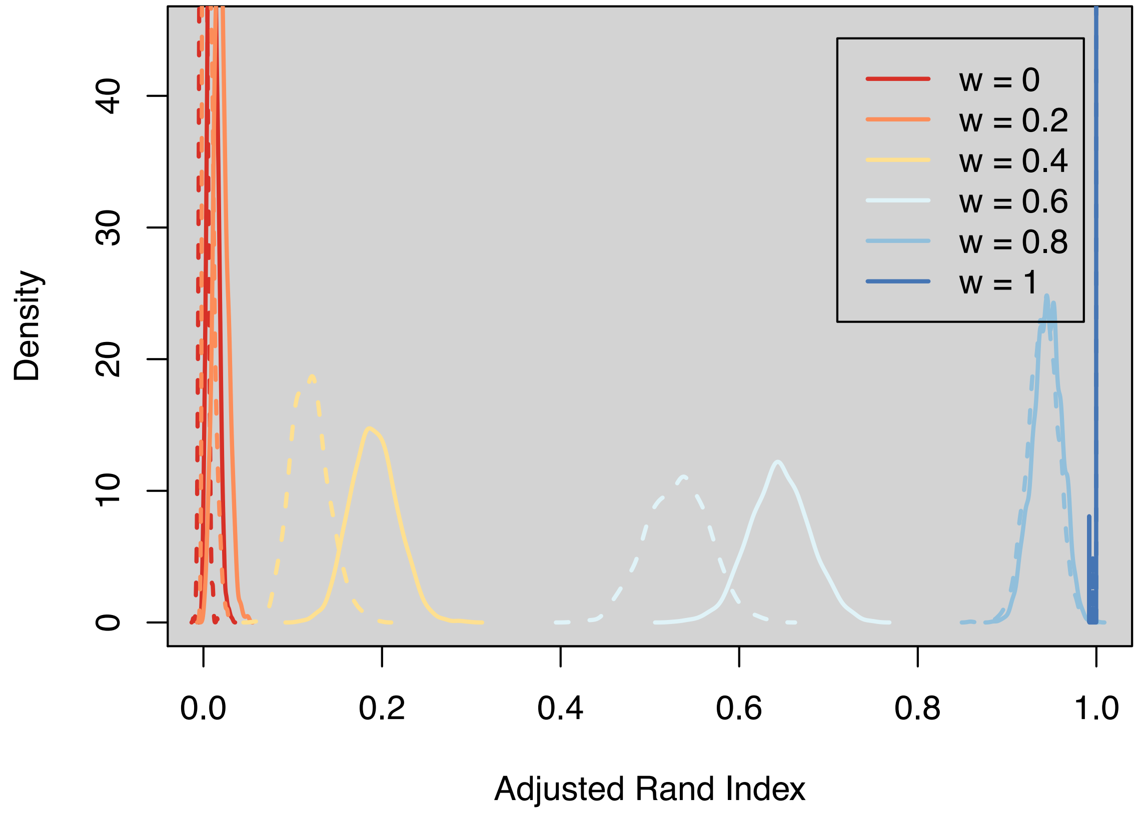
Small w

PROFILE REGRESSION: SIMULATION STUDY

- Fit a profile regression clustering model for w = 0, 0.2, 0.4, 0.6, 0.8, 1
- For each of the posterior sampled clusterings, calculate the **adjusted Rand index** (ARI) between the sampled clustering and the true clustering structure
    - ARI scores clustering quality, with values between 0 (bad) and 1 (good).

PROFILE REGRESSION: SIMULATION STUDY

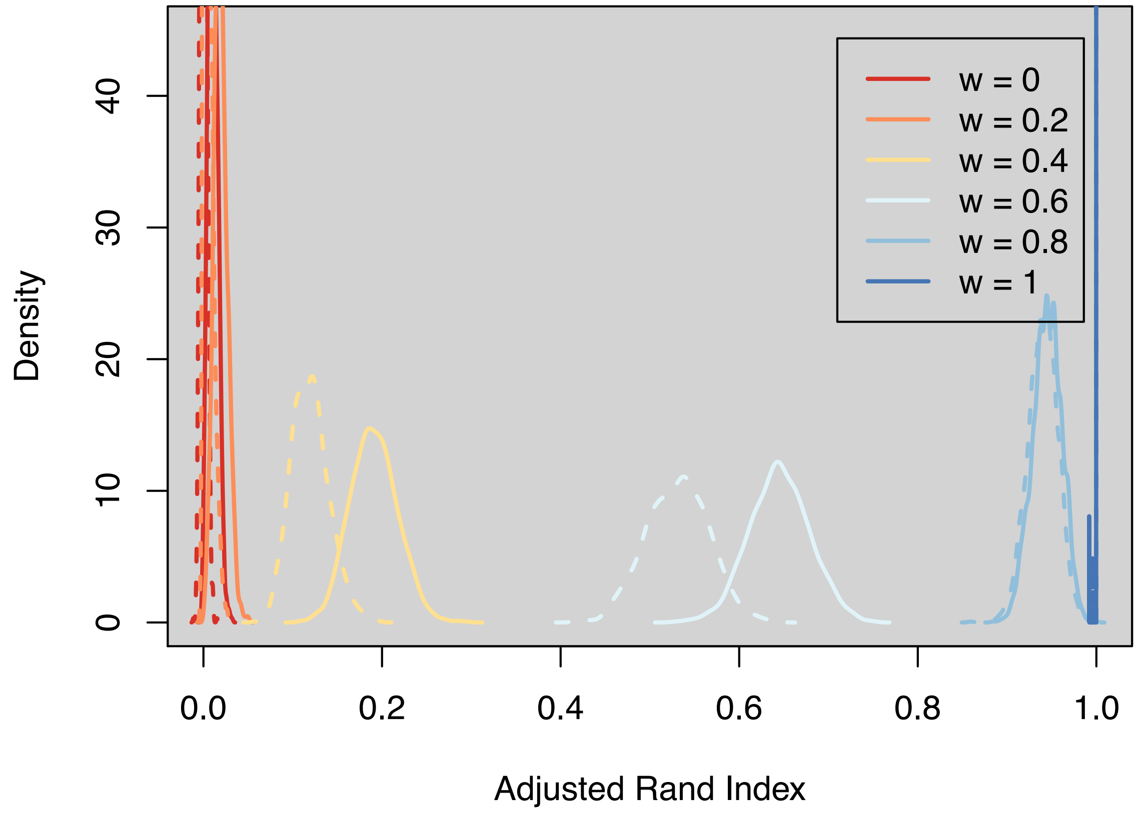
- Fit a profile regression clustering model for w = 0, 0.2, 0.4, 0.6, 0.8, 1
- For each of the posterior sampled clusterings, calculate the **adjusted Rand index** (ARI) between the sampled clustering and the true clustering structure
  - ARI scores clustering quality, with values between 0 (bad) and 1 (good).



- Solid line is the distribution of ARI values obtained if response is included
- Dotted line is the distribution of ARI values obtained if response is excluded

PROFILE REGRESSION: SIMULATION STUDY

- Fit a profile regression clustering model for w = 0, 0.2, 0.4, 0.6, 0.8, 1
- For each of the posterior sampled clusterings, calculate the **adjusted Rand index** (ARI) between the sampled clustering and the true clustering structure
  - ARI scores clustering quality, with values between 0 (bad) and 1 (good).



- Solid line is the distribution of ARI values obtained if response is included
- Dotted line is the distribution of ARI values obtained if response is excluded
- Including the response improves the clustering quality

PROFILE REGRESSION: SUMMARY

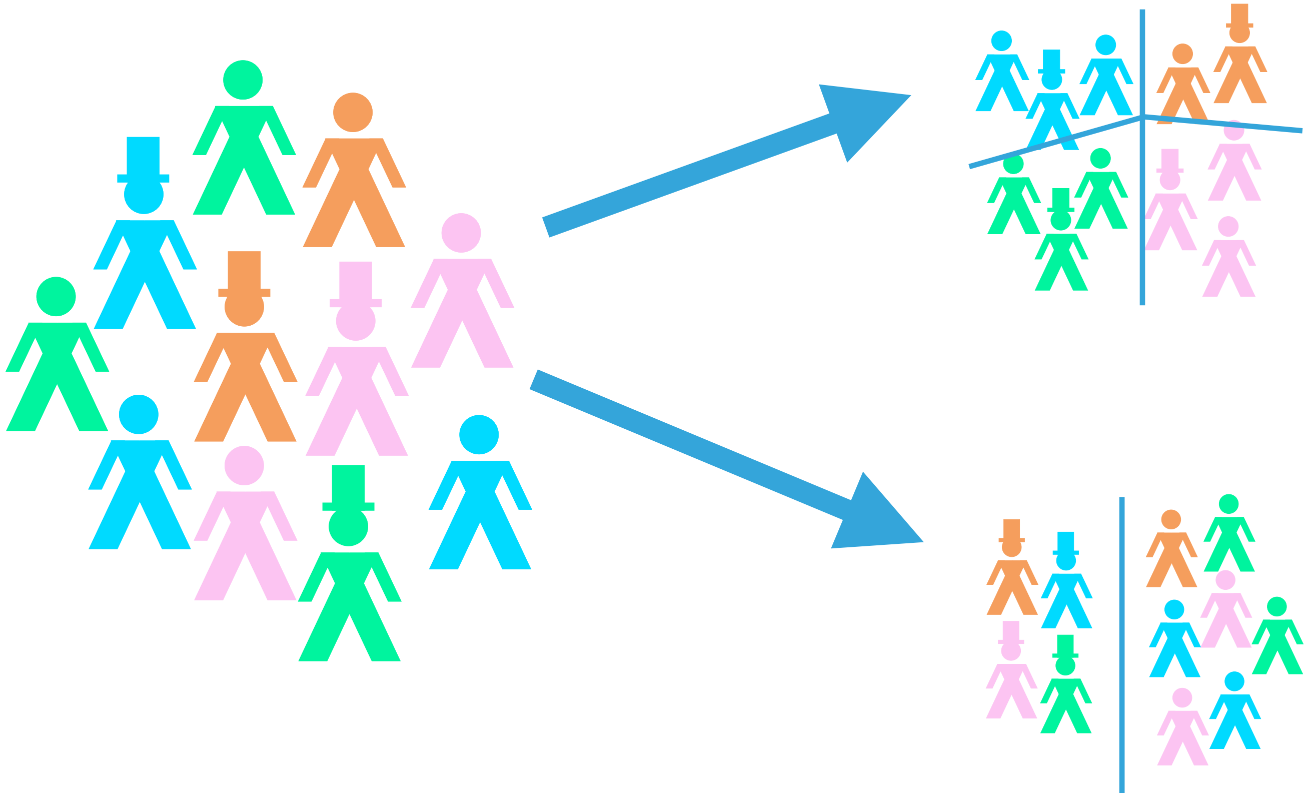
- Profile regression provides us with a way to perform semi-supervised clustering
    - Allows us to use a “response” to guide the clustering

PROFILE REGRESSION: SUMMARY

- Profile regression provides us with a way to perform semi-supervised clustering
  - Allows us to use a “response” to guide the clustering
- Variable selection removes variables that do not contribute to the clustering structure

PROFILE REGRESSION: SUMMARY

- Profile regression provides us with a way to perform semi-supervised clustering
  - Allows us to use a “response” to guide the clustering
- Variable selection removes variables that do not contribute to the clustering structure
- But does this solve the “hat problem”?

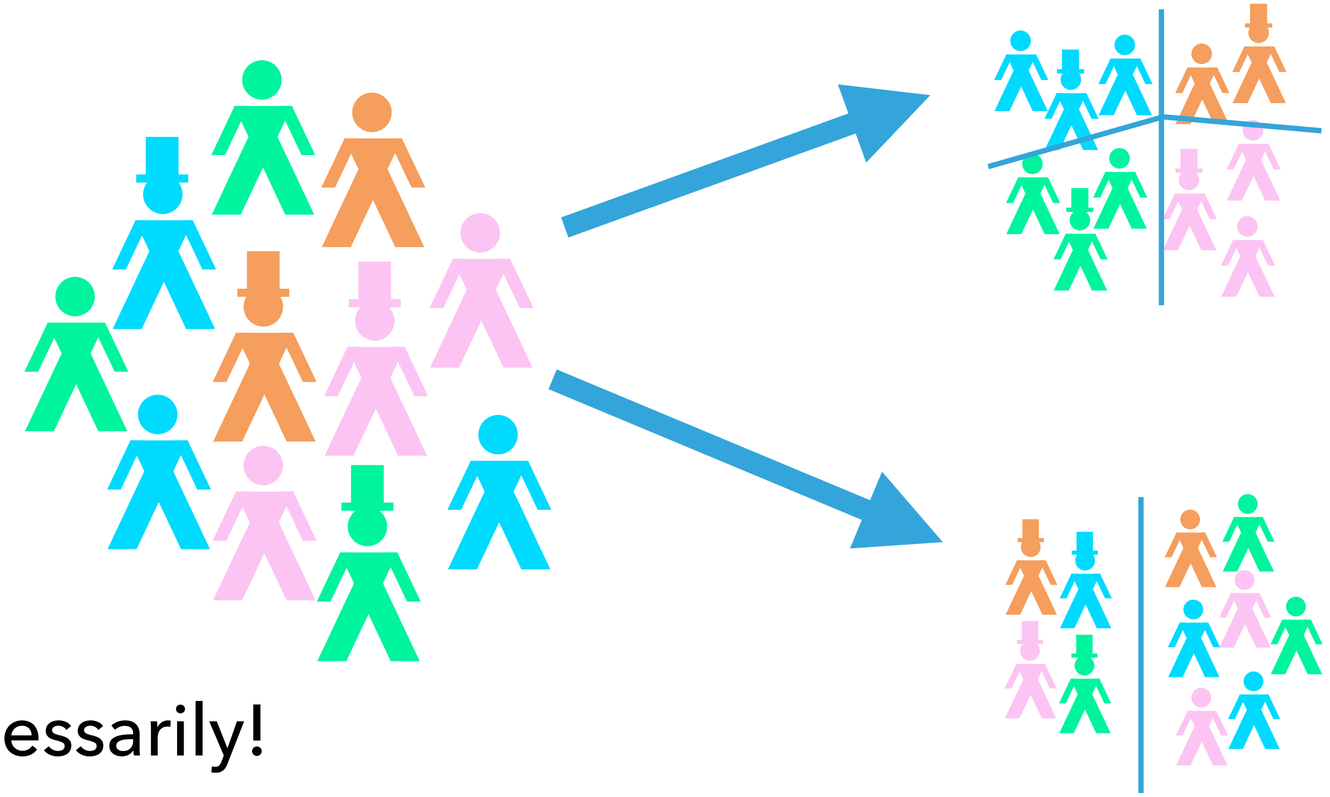


“ALL 4 GROUPS HAVE SIMILAR AVERAGE SURVIVAL TIMES”

“HAT WEARERS SURVIVE HALF AS LONG, ON AVERAGE”

PROFILE REGRESSION: SUMMARY

- Profile regression provides us with a way to perform semi-supervised clustering
  - Allows us to use a “response” to guide the clustering
- Variable selection removes variables that do not contribute to the clustering structure
- But does this solve the “hat problem”?



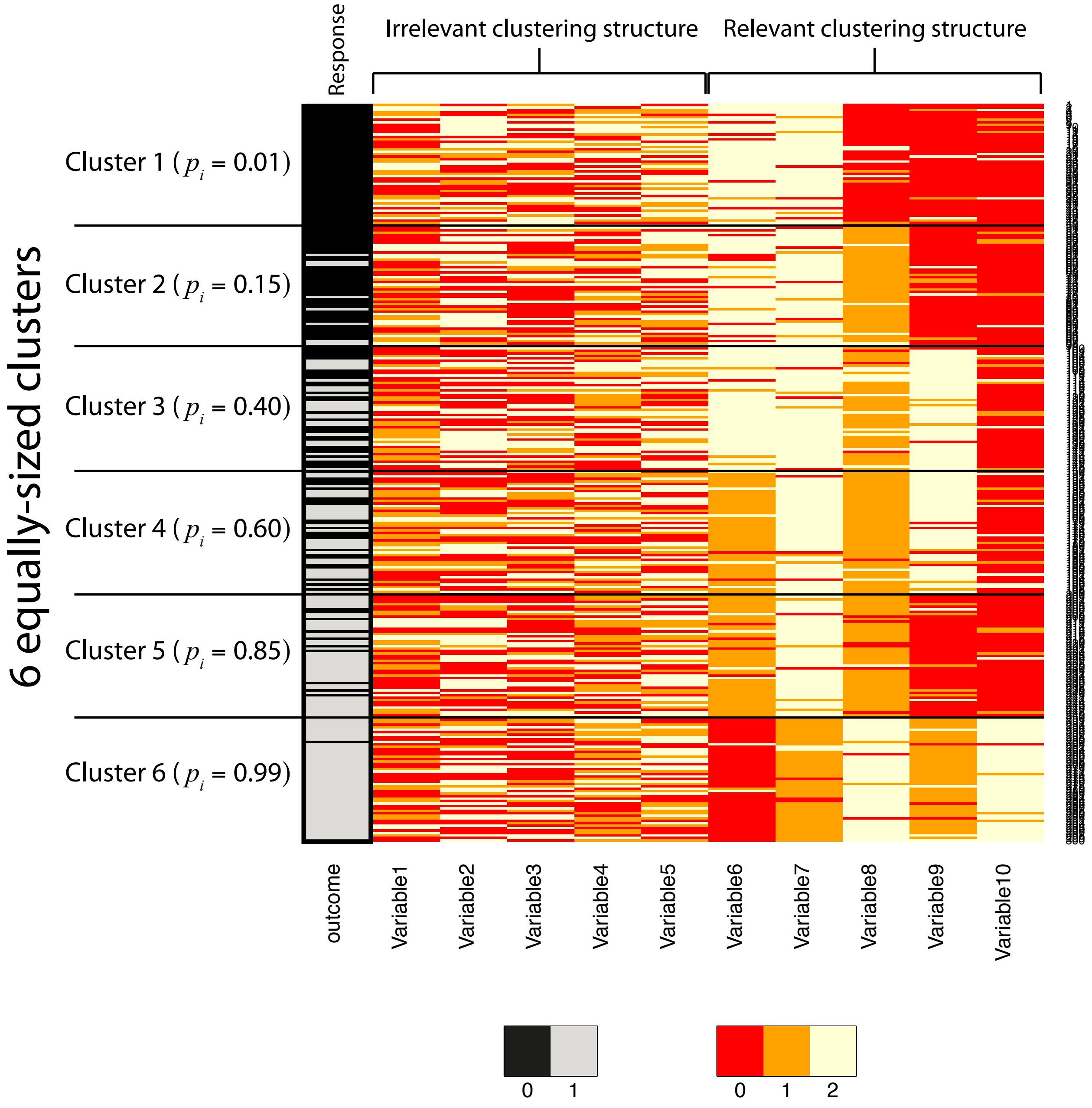
“ALL 4 GROUPS HAVE SIMILAR AVERAGE SURVIVAL TIMES”

“HAT WEARERS SURVIVE HALF AS LONG, ON AVERAGE”

- Not necessarily!
  - If we have competing clustering structures in the data, will the right one “win”?

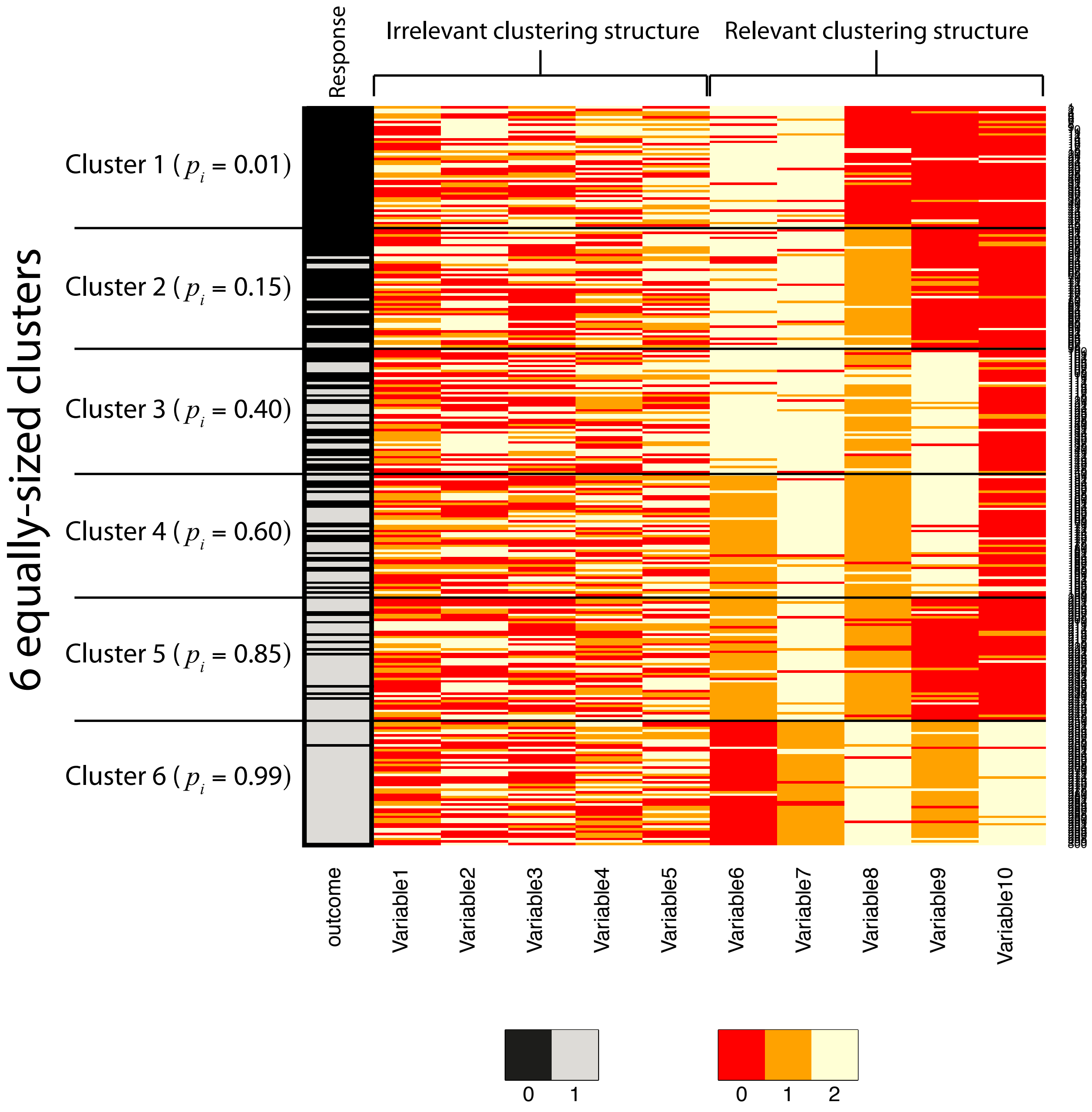
COMPETING CLUSTERING STRUCTURES

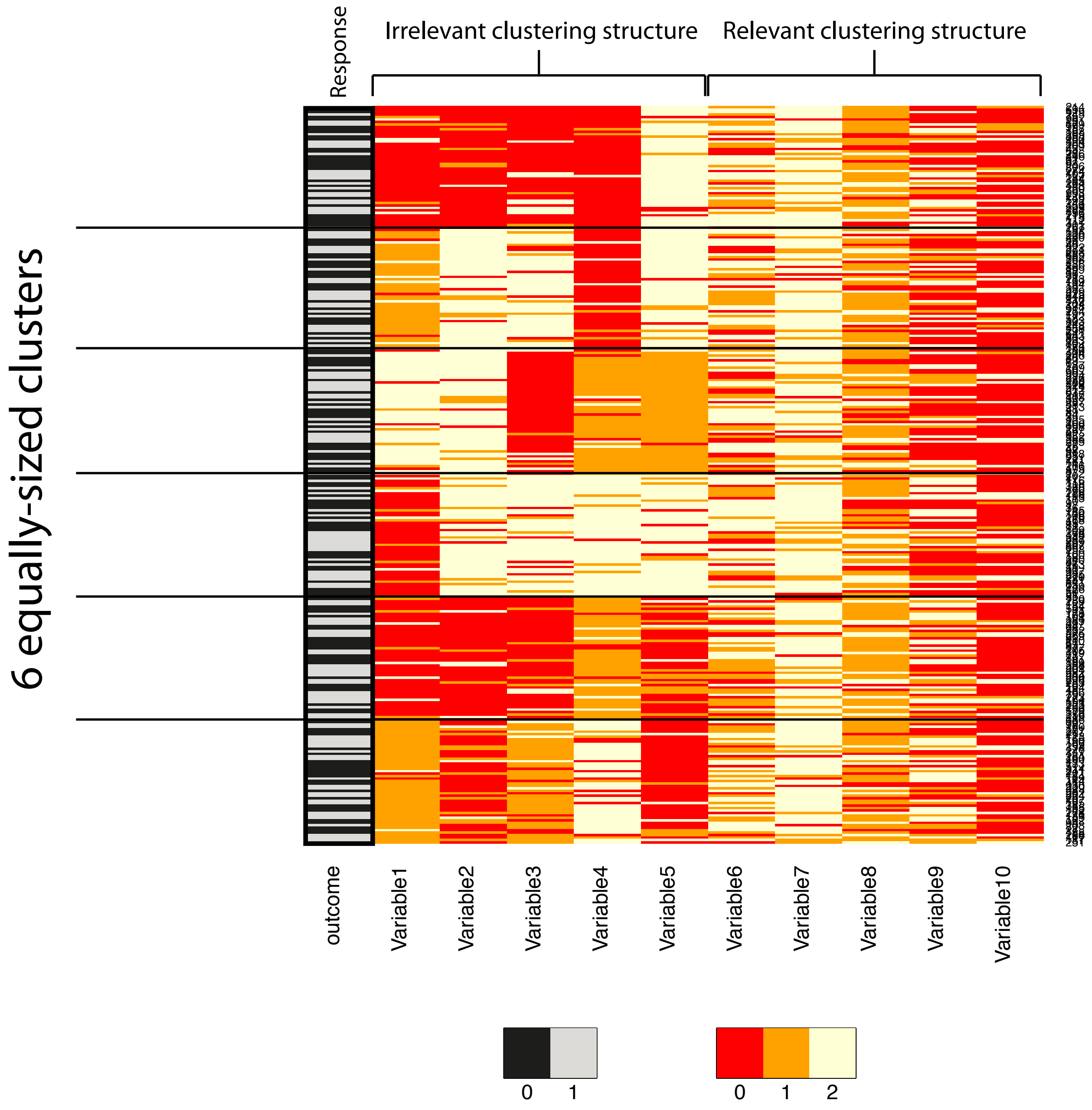
EXAMPLE



COMPETING CLUSTERING STRUCTURES

EXAMPLE



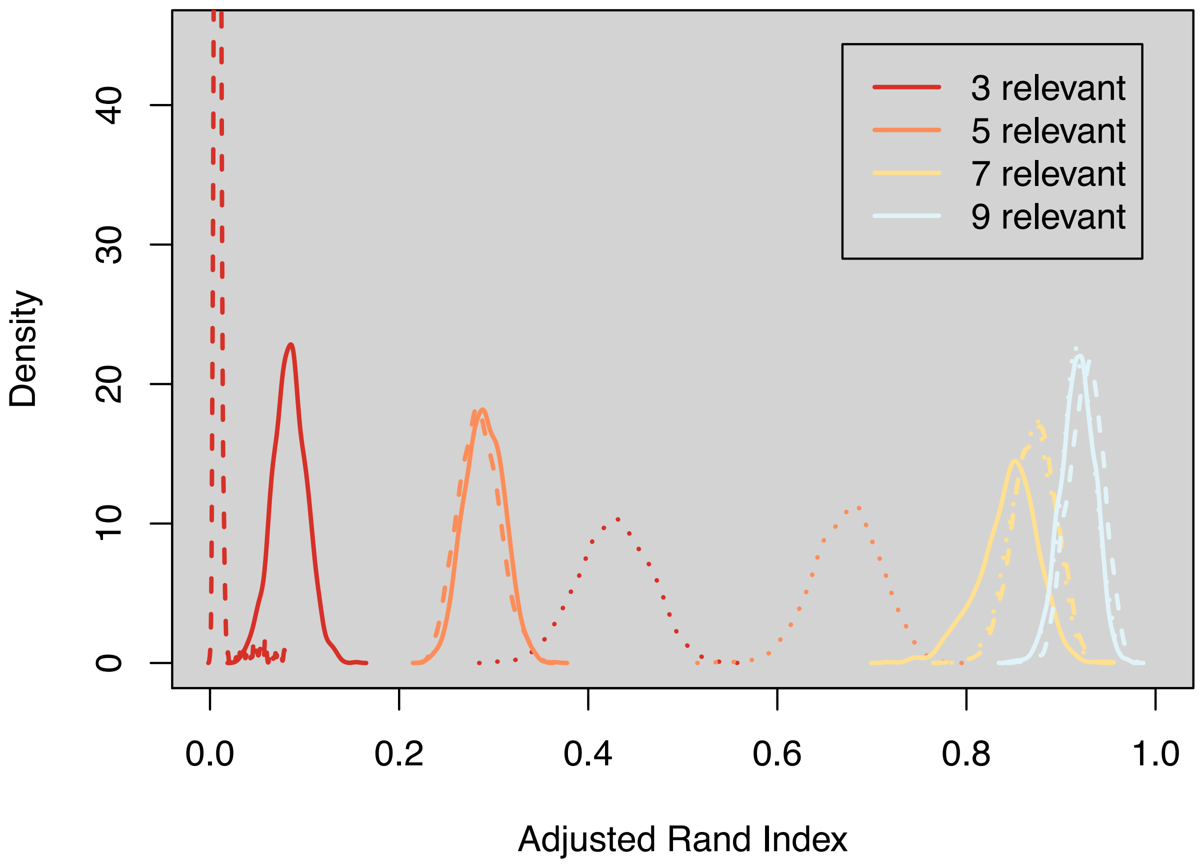


PROFILE REGRESSION: COMPETING CLUSTERINGS SIM STUDY

- 1 binary response, 10 categorical variables (as before)
- M of these possess a "relevant" clustering structure, 10 - M possess an irrelevant structure
- We consider
  - Fitting a profile regression model to the full dataset (solid line)
  - Fitting a profile regression model **with variable selection** to the full dataset (dashed line)
  - Fitting a profile regression model using just the relevant variables (dotted line)

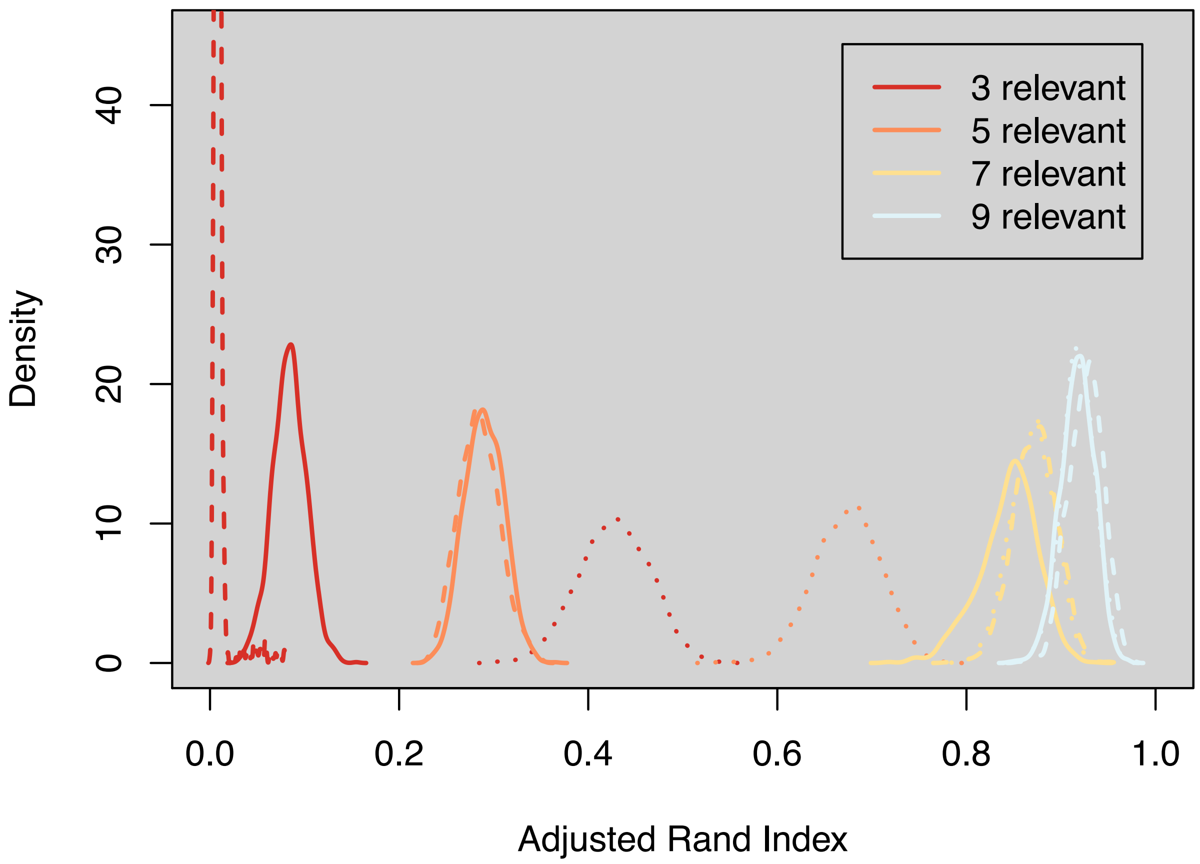
PROFILE REGRESSION: COMPETING CLUSTERINGS SIM STUDY

- 1 binary response, 10 categorical variables (as before)
- M of these possess a "relevant" clustering structure, 10 - M possess an irrelevant structure
- We consider
  - Fitting a profile regression model to the full dataset (solid line)
  - Fitting a profile regression model **with variable selection** to the full dataset (dashed line)
  - Fitting a profile regression model using just the relevant variables (dotted line)



PROFILE REGRESSION: COMPETING CLUSTERINGS SIM STUDY

- 1 binary response, 10 categorical variables (as before)
- M of these possess a “relevant” clustering structure, 10 - M possess an irrelevant structure
- We consider
  - Fitting a profile regression model to the full dataset (solid line)
  - Fitting a profile regression model **with variable selection** to the full dataset (dashed line)
  - Fitting a profile regression model using just the relevant variables (dotted line)



- Variable selection picks out the variables that define the dominant clustering structure, **not the most relevant one**

COMPETING CLUSTERING STRUCTURES: THE PROBLEM

- Typically the **dominant** clustering structure will “win”, regardless of y
- Recall the likelihood function:

COMPETING CLUSTERING STRUCTURES: THE PROBLEM

- Typically the **dominant** clustering structure will “win”, regardless of y
- Recall the likelihood function:

$$\begin{aligned} f(\mathbf{x}_i, y_i|\cdots) &= f_{\mathbf{x}}(\mathbf{x}_i|\theta_{c_i}) f_{\mathbf{y}}(y_i|\phi_{c_i}, \mathbf{w}_i, \beta) \\ &= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij}|\theta_{c_i})^{\mathbb{I}(\gamma_j=1)} f_{\mathbf{x}}(x_{ij}|\theta_0)^{\mathbb{I}(\gamma_j=0)} \right) f_{\mathbf{y}}(y_i|\phi_{c_i}, \mathbf{w}_i, \beta) \end{aligned}$$

COMPETING CLUSTERING STRUCTURES: THE PROBLEM

- Typically the **dominant** clustering structure will “win”, regardless of y
- Recall the likelihood function:

$$
\begin{aligned}
f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i | \theta_{c_i}) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta) \\
&= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij} | \theta_{c_i})^{\mathbb{I}(\gamma_j = 1)} f_{\mathbf{x}}(x_{ij} | \theta_0)^{\mathbb{I}(\gamma_j = 0)} \right) f_{\mathbf{y}}(y_i | \phi_{c_i}, \mathbf{w}_i, \beta)
\end{aligned}
$$

- Typically, J is **large**
- This means that the contribution of **x** to the likelihood is typically much greater than the contribution of y
- The “guidance” provided by y gets swamped by contribution of **x**

COMPETING CLUSTERING STRUCTURES: THE PROBLEM

- Typically the **dominant** clustering structure will "win", regardless of y
- Recall the likelihood function:

$$\begin{aligned} f(\mathbf{x}_i, y_i | \cdots) &= f_{\mathbf{x}}(\mathbf{x}_i|\theta_{c_i}) f_{\mathbf{y}}(y_i|\phi_{c_i}, \mathbf{w}_i, \beta) \\ &= \left(\prod_{j=1}^{J} f_{\mathbf{x}}(x_{ij}|\theta_{c_i})^{\mathbb{I}(\gamma_j=1)} f_{\mathbf{x}}(x_{ij}|\theta_0)^{\mathbb{I}(\gamma_j=0)}\right) f_{\mathbf{y}}(y_i|\phi_{c_i}, \mathbf{w}_i, \beta) \end{aligned}$$

- Typically, J is **large**
- This means that the contribution of **x** to the likelihood is typically much greater than the contribution of y
- The "guidance" provided by y gets swamped by contribution of **x**

How can we overcome this?

PART 3…

PART 3…

SEMI-SUPERVISED MULTIVIEW CLUSTERING

MULTIVIEW PROFILE REGRESSION

- We model multiple clustering structures in the data
- We use “side-information” to pick out the most relevant of these clustering structures

MULTIVIEW PROFILE REGRESSION

- We model multiple clustering structures in the data
- We use "side-information" to pick out the most relevant of these clustering structures

DATA MATRIX

MULTIVIEW PROFILE REGRESSION

- We model multiple clustering structures in the data
- We use “side-information” to pick out the most relevant of these clustering structures

DATA MATRIX

individuals
(to be clustered)

MULTIVIEW PROFILE REGRESSION

- We model multiple clustering structures in the data
- We use “side-information” to pick out the most relevant of these clustering structures

DATA MATRIX

individuals
(to be clustered)

MULTIVIEW PROFILE REGRESSION

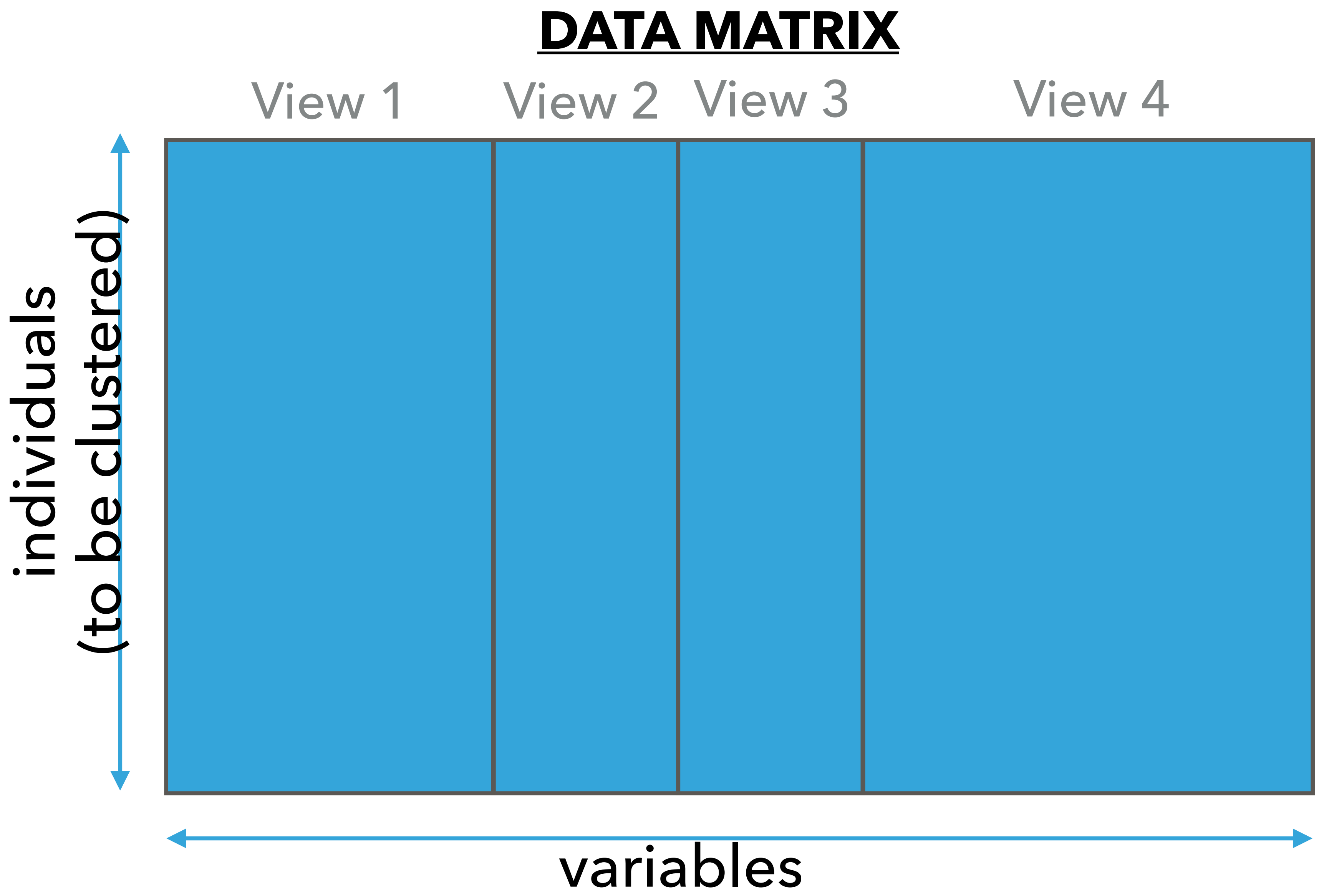
- We model multiple clustering structures in the data
- We use “side-information” to pick out the most relevant of these clustering structures

DATA MATRIX

individuals
(to be clustered)
variables

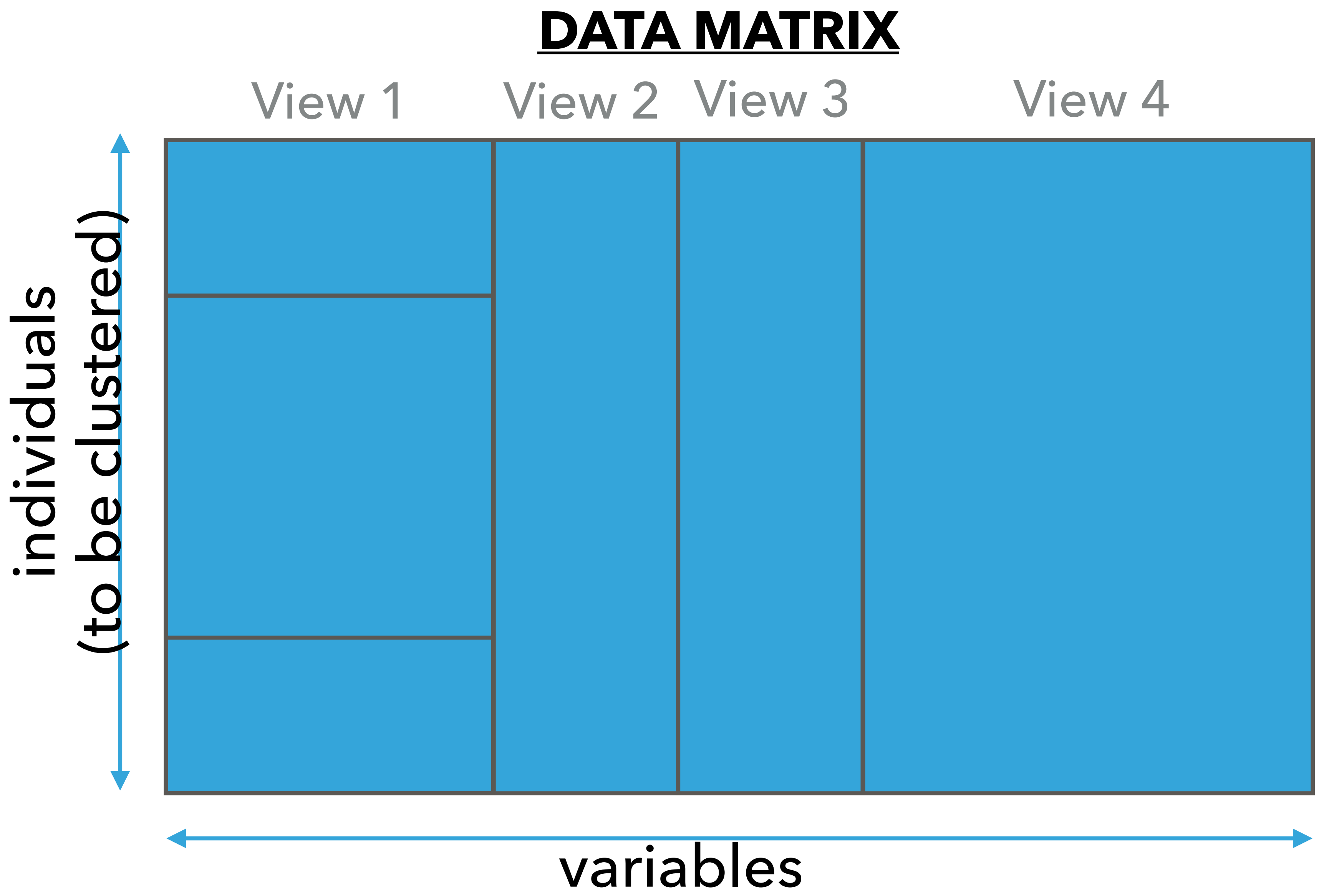
MULTIVIEW PROFILE REGRESSION

- We model multiple clustering structures in the data
- We use "side-information" to pick out the most relevant of these clustering structures



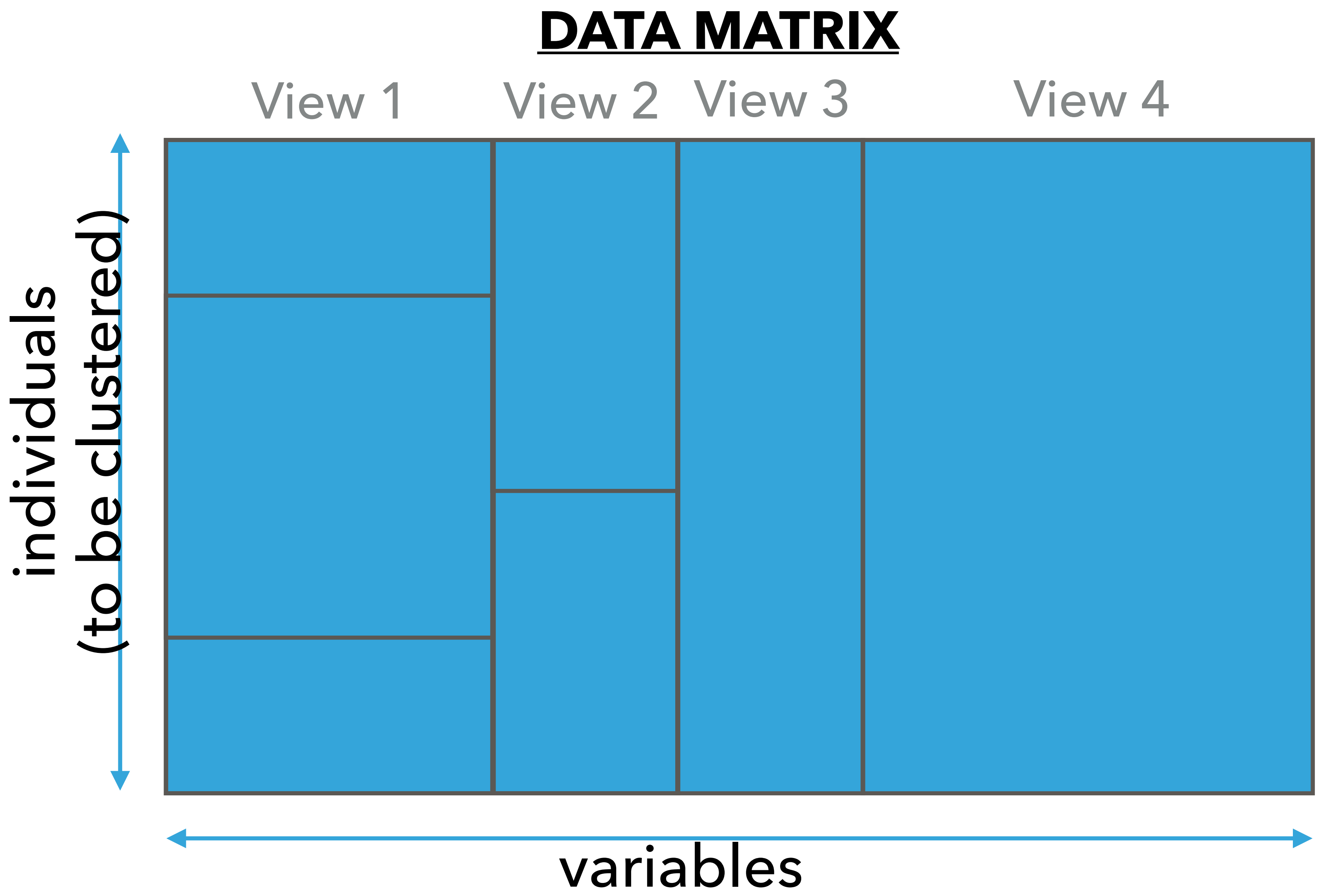
MULTIVIEW PROFILE REGRESSION

- We model multiple clustering structures in the data
- We use "side-information" to pick out the most relevant of these clustering structures



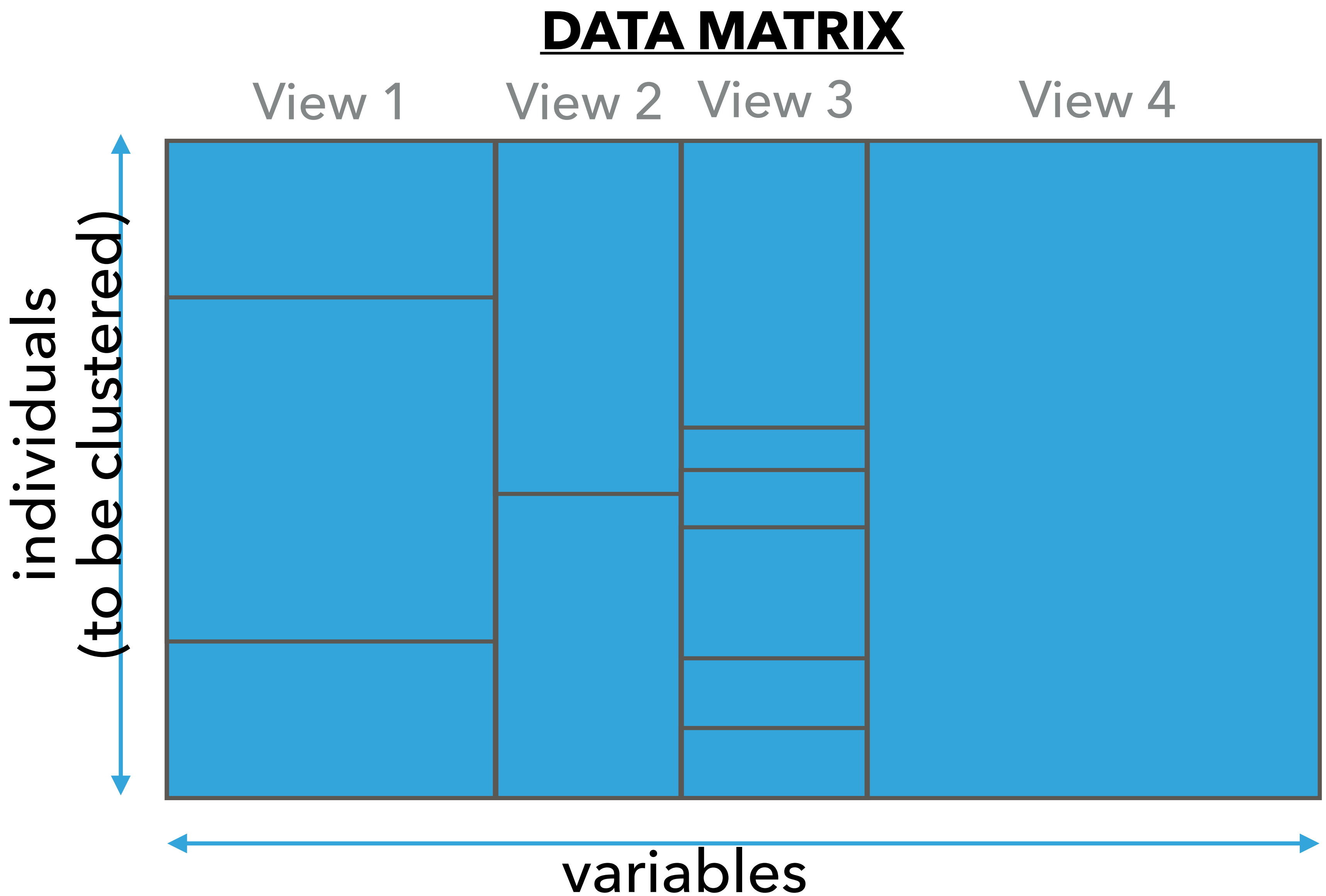
MULTIVIEW PROFILE REGRESSION

- We model multiple clustering structures in the data
- We use "side-information" to pick out the most relevant of these clustering structures



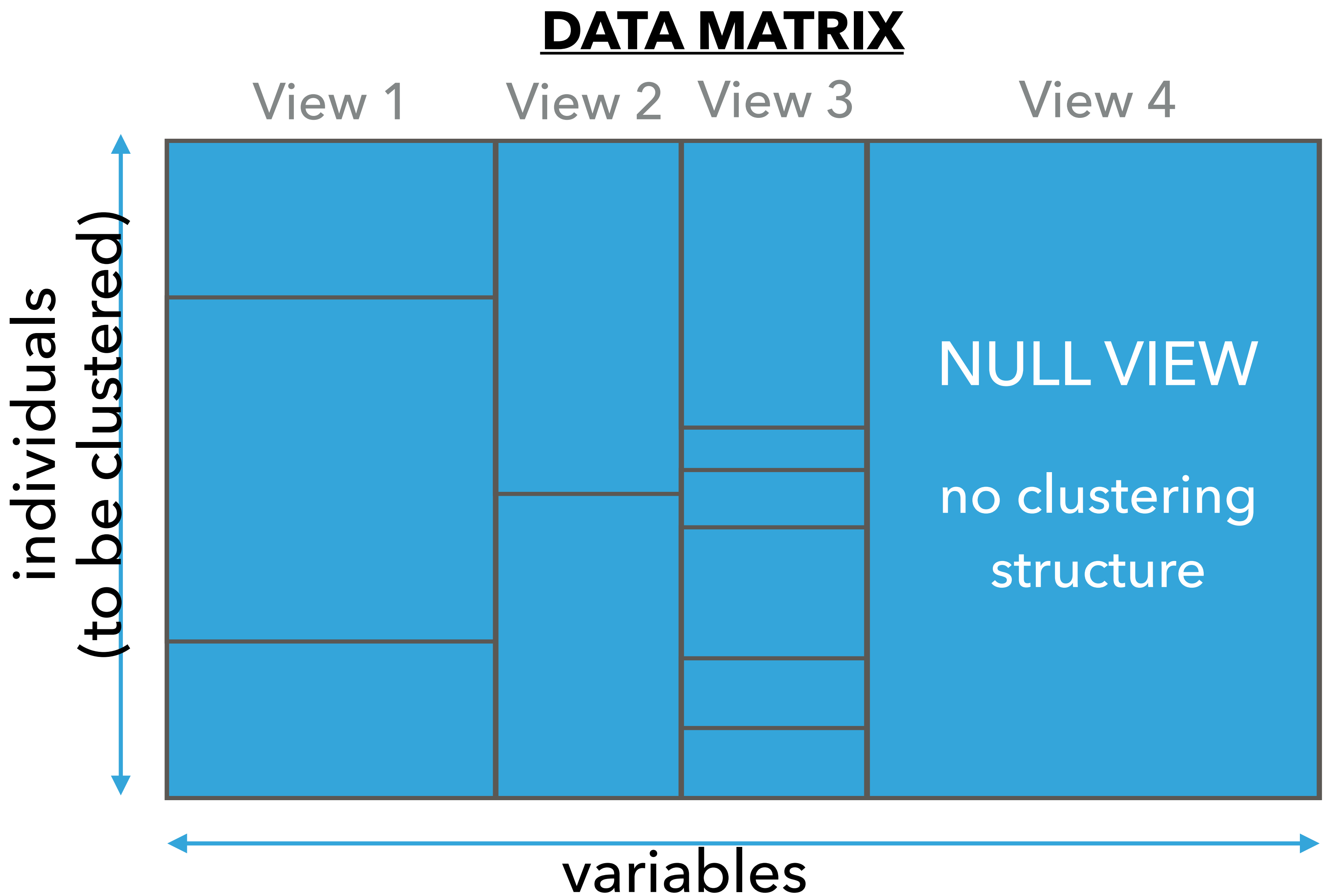
MULTIVIEW PROFILE REGRESSION

- We model multiple clustering structures in the data
- We use “side-information” to pick out the most relevant of these clustering structures



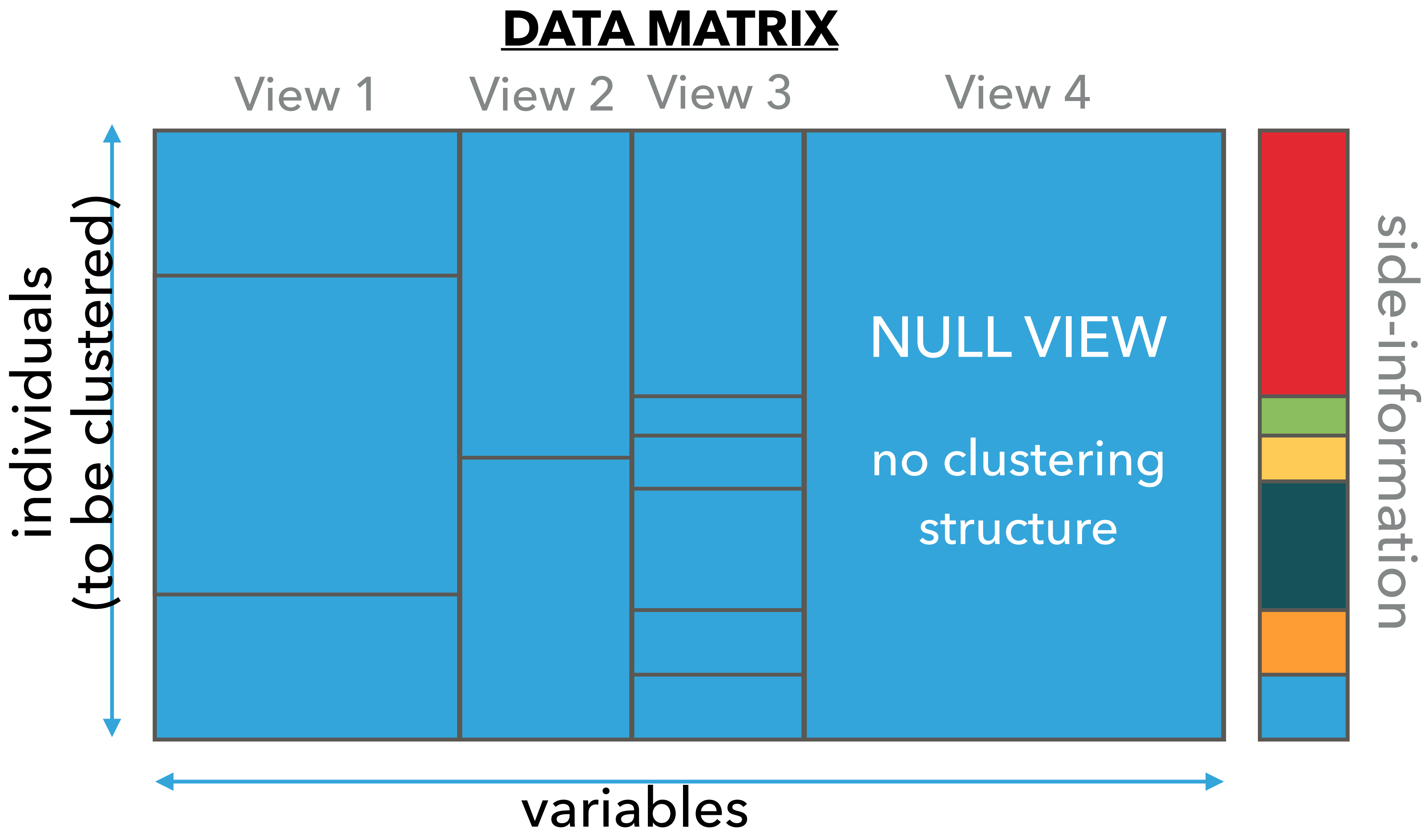
MULTIVIEW PROFILE REGRESSION

- We model multiple clustering structures in the data
- We use "side-information" to pick out the most relevant of these clustering structures



MULTIVIEW PROFILE REGRESSION

- We model multiple clustering structures in the data
- We use "side-information" to pick out the most relevant of these clustering structures



MIXTURE MODELS: MULTI-VIEW PROFILE REGRESSION

MIXTURE MODELS: MULTI-VIEW PROFILE REGRESSION

- Suppose we wish to model V different views of the data

MIXTURE MODELS: MULTI-VIEW PROFILE REGRESSION

- Suppose we wish to model V different views of the data
  - 1 null view, and V-1 views that each possess a (distinct) clustering structure

MIXTURE MODELS: MULTI-VIEW PROFILE REGRESSION

- Suppose we wish to model V different views of the data
  - 1 null view, and V-1 views that each possess a (distinct) clustering structure

- Introduce categorical "view membership" indicators, $\gamma_j \in \{0, 1, 2, \ldots, V-1\}$
  - If γ_j = 0, the j-th variable is in the null view (no clustering structure)
  - If γ_j = 1, the j-th variable is in the "relevant" clustering view (the clustering structure which is most useful for stratifying the individuals according to the response, y)
  - The remaining views mop up clustering structures present in the data that are not relevant for the present stratification task

MIXTURE MODELS: MULTIVIEW PROFILE REGRESSION

MIXTURE MODELS: MULTIVIEW PROFILE REGRESSION

- Within each non-null view, we model the data using a mixture model.

MIXTURE MODELS: MULTIVIEW PROFILE REGRESSION

- Within each non-null view, we model the data using a mixture model.
- Define $c_i^{(v)}$ to be the latent component allocation variable associated with the *i*-th individual in the *v*-th view.

MIXTURE MODELS: MULTIVIEW PROFILE REGRESSION

- Within each non-null view, we model the data using a mixture model.
- Define $c_i^{(v)}$ to be the latent component allocation variable associated with the *i*-th individual in the *v*-th view.
- Define $\theta_k^{(v)}$ to be the parameters associated with the *k*-th component in the mixture model for the *v*-th view.

MIXTURE MODELS: MULTIVIEW PROFILE REGRESSION

- Within each non-null view, we model the data using a mixture model.
- Define $c_i^{(v)}$ to be the latent component allocation variable associated with the *i*-th individual in the *v*-th view.
- Define $\theta_k^{(v)}$ to be the parameters associated with the *k*-th component in the mixture model for the *v*-th view.
- Our likelihood model becomes:

MIXTURE MODELS: MULTIVIEW PROFILE REGRESSION

- Within each non-null view, we model the data using a mixture model.
- Define $c_i^{(v)}$ to be the latent component allocation variable associated with the *i*-th individual in the *v*-th view.
- Define $\theta_k^{(v)}$ to be the parameters associated with the *k*-th component in the mixture model for the *v*-th view.
- Our likelihood model becomes:

$$f(\mathbf{x}_i, y_i | \cdots) = \left(\prod_{j=1}^{J} \left(\prod_{v=1}^{V-1} f_{\mathbf{x}}(x_{ij} | \theta^{(v)}_{c_i^{(v)}})^{\mathbb{I}(\gamma_j = v)} \right) f_{\mathbf{x}}(x_{ij} | \theta_0)^{\mathbb{I}(\gamma_j = 0)} \right) f_{\mathbf{y}}(y_i | \phi_{c_i^{(1)}}, \mathbf{w}_i, \beta)$$

MIXTURE MODELS: MULTIVIEW PROFILE REGRESSION

- Within each non-null view, we model the data using a mixture model.
- Define $c_i^{(v)}$ to be the latent component allocation variable associated with the *i*-th individual in the *v*-th view.
- Define $\theta_k^{(v)}$ to be the parameters associated with the *k*-th component in the mixture model for the *v*-th view.
- Our likelihood model becomes:

$$f(\mathbf{x}_i, y_i | \cdots) = \left(\prod_{j=1}^{J} \left(\prod_{v=1}^{V-1} f_{\mathbf{x}}(x_{ij} | \theta^{(v)}_{c_i^{(v)}})^{\mathbb{I}(\gamma_j = v)} \right) f_{\mathbf{x}}(x_{ij} | \theta_0)^{\mathbb{I}(\gamma_j = 0)} \right) f_{\mathbf{y}}(y_i | \phi_{c_i^{(1)}}, \mathbf{w}_i, \beta)$$

THIS LOOKS FAIRLY UGLY…

MIXTURE MODELS: MULTIVIEW PROFILE REGRESSION

- Within each non-null view, we model the data using a mixture model.
- Define $c_i^{(v)}$ to be the latent component allocation variable associated with the *i*-th individual in the *v*-th view.
- Define $\theta_k^{(v)}$ to be the parameters associated with the *k*-th component in the mixture model for the *v*-th view.
- Our likelihood model becomes:

$$f(\mathbf{x}_i, y_i | \cdots) = \left(\prod_{j=1}^{J} \left(\prod_{v=1}^{V-1} f_{\mathbf{x}}(x_{ij} | \theta^{(v)}_{c_i^{(v)}})^{\mathbb{I}(\gamma_j = v)} \right) f_{\mathbf{x}}(x_{ij} | \theta_0)^{\mathbb{I}(\gamma_j = 0)} \right) f_{\mathbf{y}}(y_i | \phi_{c_i^{(1)}}, \mathbf{w}_i, \beta)$$

THIS LOOKS FAIRLY UGLY…

…BUT IS ACTUALLY EASY TO DEAL WITH.

MIXTURE MODELS: MULTI-VIEW PROFILE REGRESSION

MIXTURE MODELS: MULTI-VIEW PROFILE REGRESSION

- Given the view allocation indicators, we act as if we have V datasets that we model independently

MIXTURE MODELS: MULTI-VIEW PROFILE REGRESSION

- Given the view allocation indicators, we act as if we have V datasets that we model independently
    - The null view ($v = 0$) has no clustering structure (so is easy)

MIXTURE MODELS: MULTI-VIEW PROFILE REGRESSION

- Given the view allocation indicators, we act as if we have V datasets that we model independently
    - The null view ($v = 0$) has no clustering structure (so is easy)
    - The $v = 1$ view is modelled using a profile regression model

MIXTURE MODELS: MULTI-VIEW PROFILE REGRESSION

- Given the view allocation indicators, we act as if we have V datasets that we model independently
    - The null view (v = 0) has no clustering structure (so is easy)
    - The v = 1 view is modelled using a profile regression model
    - The other views are modelled using vanilla DP mixture models

MIXTURE MODELS: MULTI-VIEW PROFILE REGRESSION

- Given the view allocation indicators, we act as if we have V datasets that we model independently
    - The null view (v = 0) has no clustering structure (so is easy)
    - The v = 1 view is modelled using a profile regression model
    - The other views are modelled using vanilla DP mixture models
- Sampling the view allocation indicators is straightforward

MIXTURE MODELS: MULTI-VIEW PROFILE REGRESSION

- Given the view allocation indicators, we act as if we have V datasets that we model independently
  - The null view (v = 0) has no clustering structure (so is easy)
  - The v = 1 view is modelled using a profile regression model
  - The other views are modelled using vanilla DP mixture models
- Sampling the view allocation indicators is straightforward

$$p(\gamma_j = 0 | \cdots) \propto p(\gamma_j = 0) \prod_{i=1}^{n} f_{\mathbf{x}}(x_{ij} | \theta_0)$$

$$p(\gamma_j = v | \cdots) \propto p(\gamma_j = v) \prod_{i=1}^{n} f_{\mathbf{x}}(x_{ij} | \theta^{(v)}_{c_i^{(v)}}) \qquad \text{for } v = 1, \ldots, V-1$$

PLATE DIAGRAM: INITIAL BASIC CASE

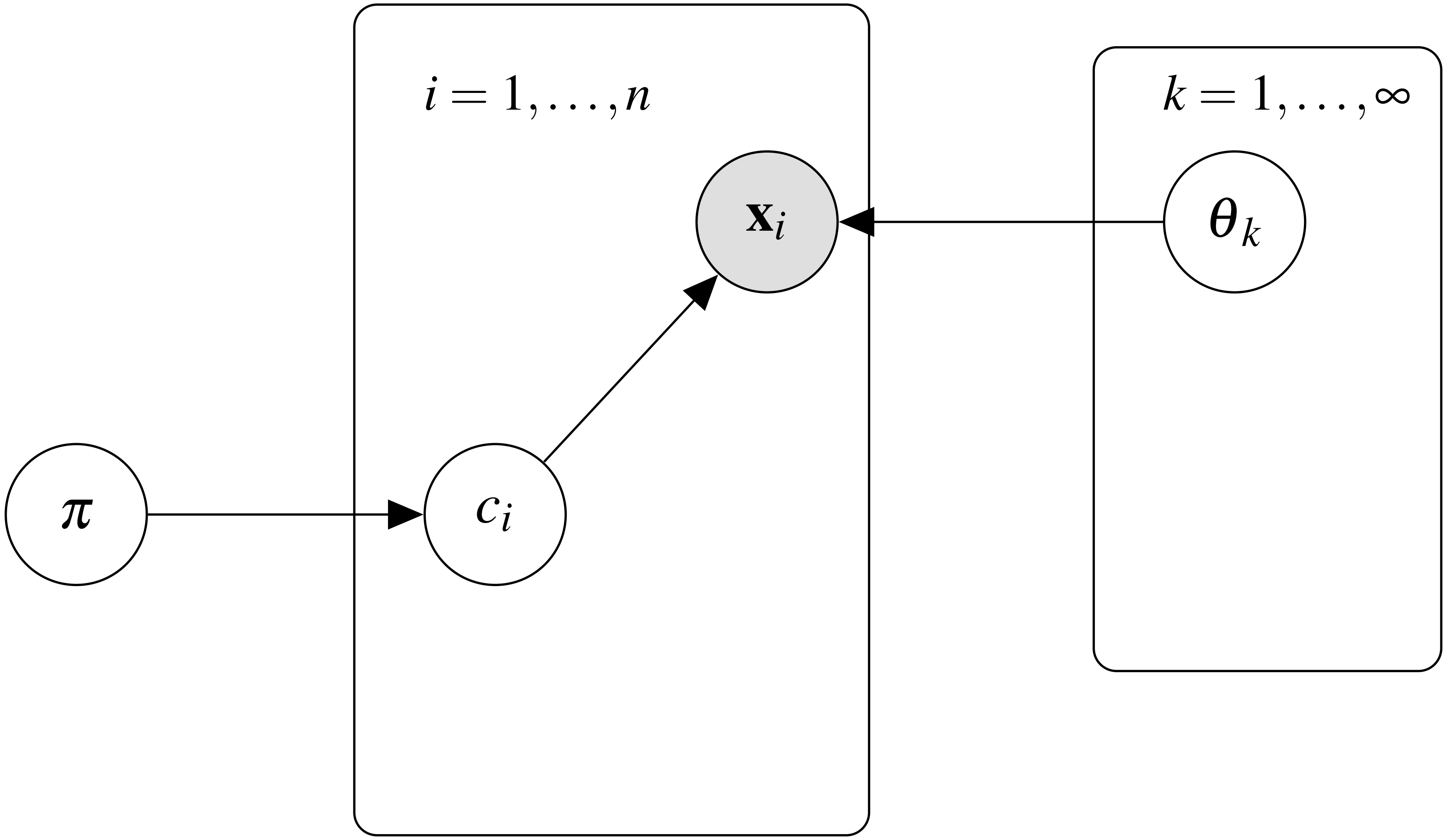


PLATE DIAGRAM: PROFILE REGRESSION

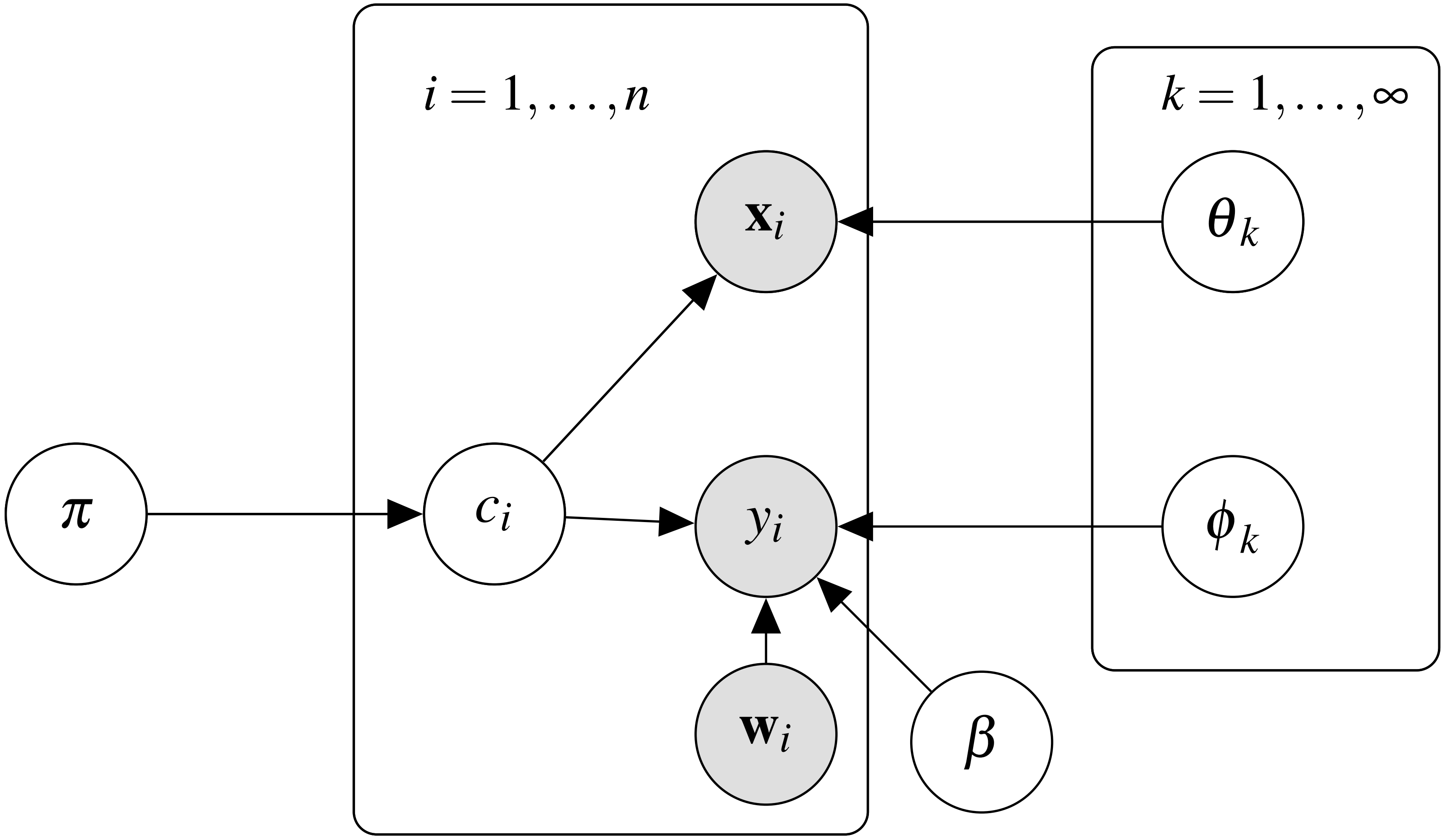


PLATE DIAGRAM: VARIABLE SELECTION

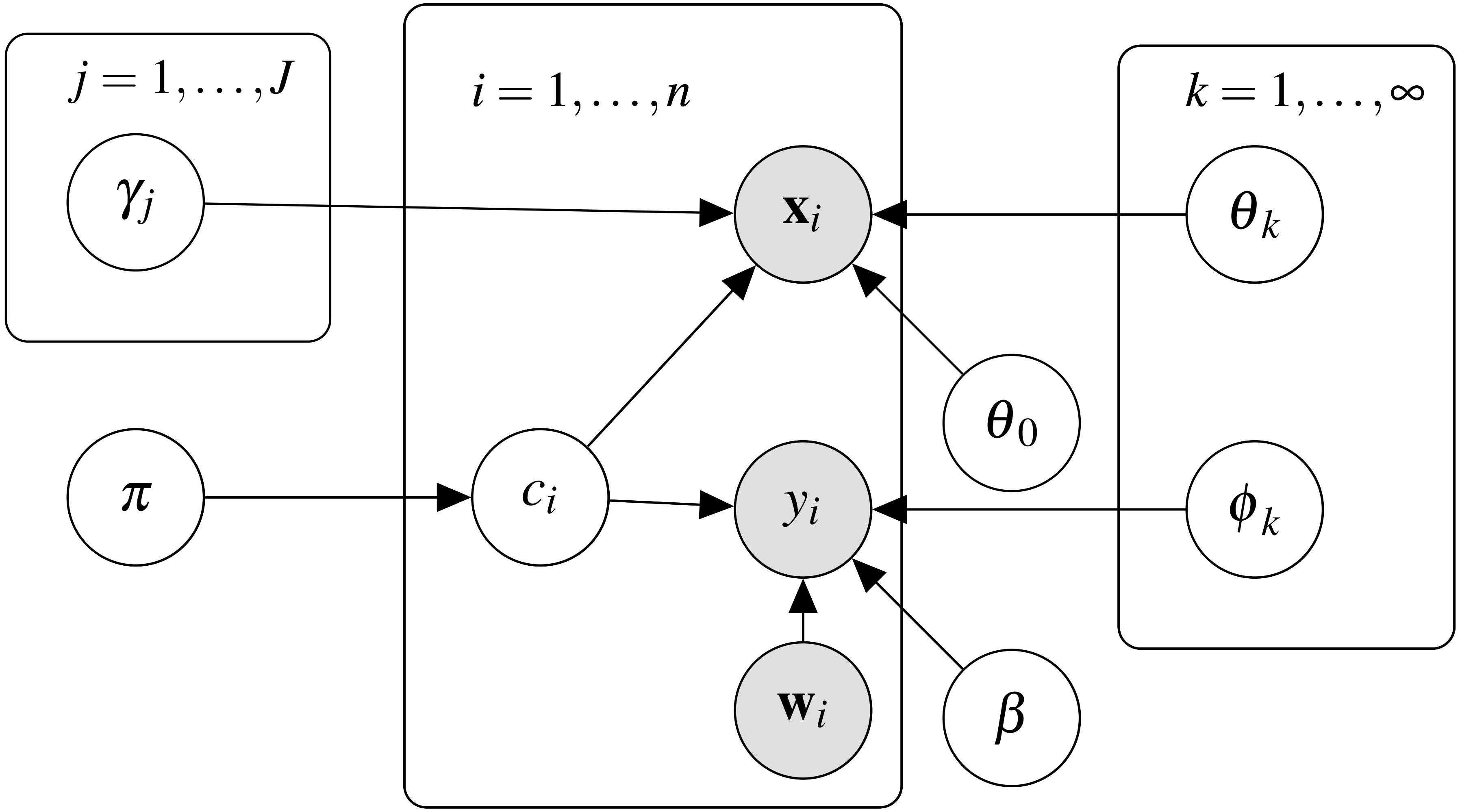
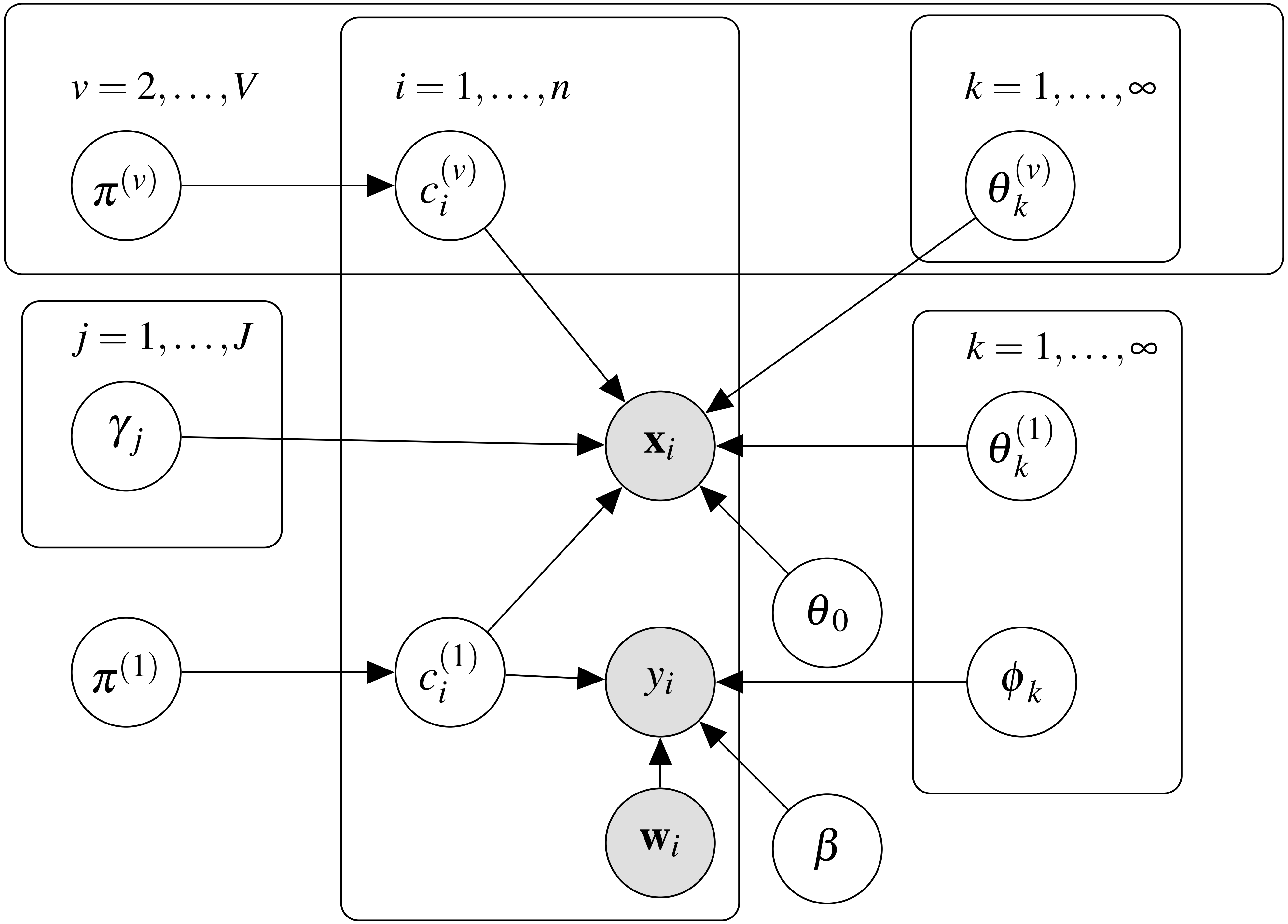


PLATE DIAGRAM: MULTI-VIEW PROFILE REGRESSION



MIXTURE MODELS: MULTI-VIEW PROFILE REGRESSION

CONSIDERATIONS

- How to determine the number of views, V.
- In the remainder, we fix V, and we set $p(\gamma_j = v)$ = 1/V for all v.
- In principle, we could instead treat the prior probabilities of view membership as parameters, and adopt a Dirichlet or Dirichlet process prior, to allow the number of views to be inferred.
  - We have explored this in the unsupervised case.
  - Can be **very** computationally costly.
- Currently investigating whether or not it is important to get the "right" number of views, if our interest is in the relevant view only.

PART 4. . .

PART 4. . .

EXAMPLES

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

- Due to the diversity of ‘omics datasets, there has been much interest in **integrative clustering** approaches, which identify common/contrasting clustering structures across multiple datasets

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

- Due to the diversity of 'omics datasets, there has been much interest in **integrative clustering** approaches, which identify common/contrasting clustering structures across multiple datasets
  - Kirk, Griffin, Savage, Ghahramani, & Wild (2012). **Bayesian correlated clustering to integrate multiple datasets**. *Bioinformatics*, *28*(24), 3290–3297.

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

- Due to the diversity of ‘omics datasets, there has been much interest in **integrative clustering** approaches, which identify common/contrasting clustering structures across multiple datasets
    - Kirk, Griffin, Savage, Ghahramani, & Wild (2012). **Bayesian correlated clustering to integrate multiple datasets.** *Bioinformatics*, *28*(24), 3290–3297.
- Multiview clustering is a biclustering approach in which we cluster variables together if they define the same stratification of patients.

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

- Due to the diversity of 'omics datasets, there has been much interest in **integrative clustering** approaches, which identify common/contrasting clustering structures across multiple datasets
  - Kirk, Griffin, Savage, Ghahramani, & Wild (2012). **Bayesian correlated clustering to integrate multiple datasets**. *Bioinformatics*, *28*(24), 3290–3297.
- Multiview clustering is a biclustering approach in which we cluster variables together if they define the same stratification of patients.
- It does not matter where the variables came from (or what data type they are), it just matters whether or not they define the same stratification

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

- Due to the diversity of 'omics datasets, there has been much interest in **integrative clustering** approaches, which identify common/contrasting clustering structures across multiple datasets
    - Kirk, Griffin, Savage, Ghahramani, & Wild (2012). **Bayesian correlated clustering to integrate multiple datasets**. *Bioinformatics*, *28*(24), 3290–3297.
- Multiview clustering is a biclustering approach in which we cluster variables together if they define the same stratification of patients.
- It does not matter where the variables came from (or what data type they are), it just matters whether or not they define the same stratification
- Thus, multiview clustering permits data integration (assuming each dataset provides information on a common set of patients):

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

- Due to the diversity of 'omics datasets, there has been much interest in **integrative clustering** approaches, which identify common/contrasting clustering structures across multiple datasets
  - Kirk, Griffin, Savage, Ghahramani, & Wild (2012). **Bayesian correlated clustering to integrate multiple datasets**. *Bioinformatics*, *28*(24), 3290–3297.
- Multiview clustering is a biclustering approach in which we cluster variables together if they define the same stratification of patients.
- It does not matter where the variables came from (or what data type they are), it just matters whether or not they define the same stratification
- Thus, multiview clustering permits data integration (assuming each dataset provides information on a common set of patients):
  - Concatenate the data matrices, and then see which variables are selected into the relevant view.

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

- n = 108 breast cancer tumour samples

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

- *n* = 108 breast cancer tumour samples
- **x :** miRNA (p = 423) and protein data (p = 171), both discretised

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

- n = 108 breast cancer tumour samples
- **x :** miRNA (p = 423) and protein data (p = 171), both discretised
- y: PAM50 tumour subtype

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

- n = 108 breast cancer tumour samples
- **x :** miRNA (p = 423) and protein data (p = 171), both discretised
- y: PAM50 tumour subtype
    - We include 2 different tumour types:

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

- n = 108 breast cancer tumour samples
- **x :** miRNA (p = 423) and protein data (p = 171), both discretised
- y: PAM50 tumour subtype
    - We include 2 different tumour types:
        - 66 Basal-like

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

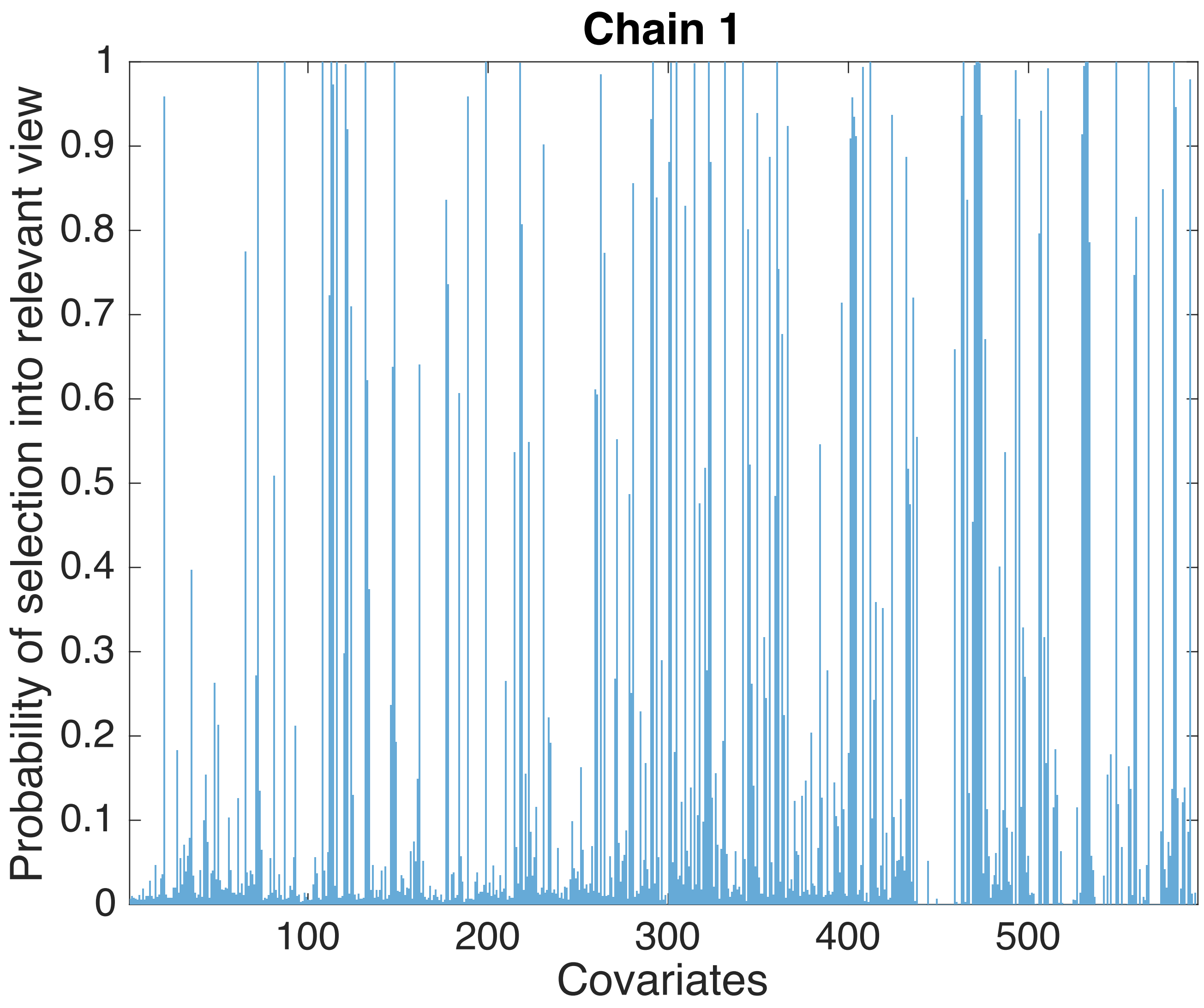
- n = 108 breast cancer tumour samples
- **x :** miRNA (p = 423) and protein data (p = 171), both discretised
- y: PAM50 tumour subtype
    - We include 2 different tumour types:
        - 66 Basal-like
        - 42 Luminal A

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

- n = 108 breast cancer tumour samples
- **x :** miRNA (p = 423) and protein data (p = 171), both discretised
- y: PAM50 tumour subtype
  - We include 2 different tumour types:
    - 66 Basal-like
    - 42 Luminal A
- **Model 3 views:** 1 relevant, 1 irrelevant, 1 null

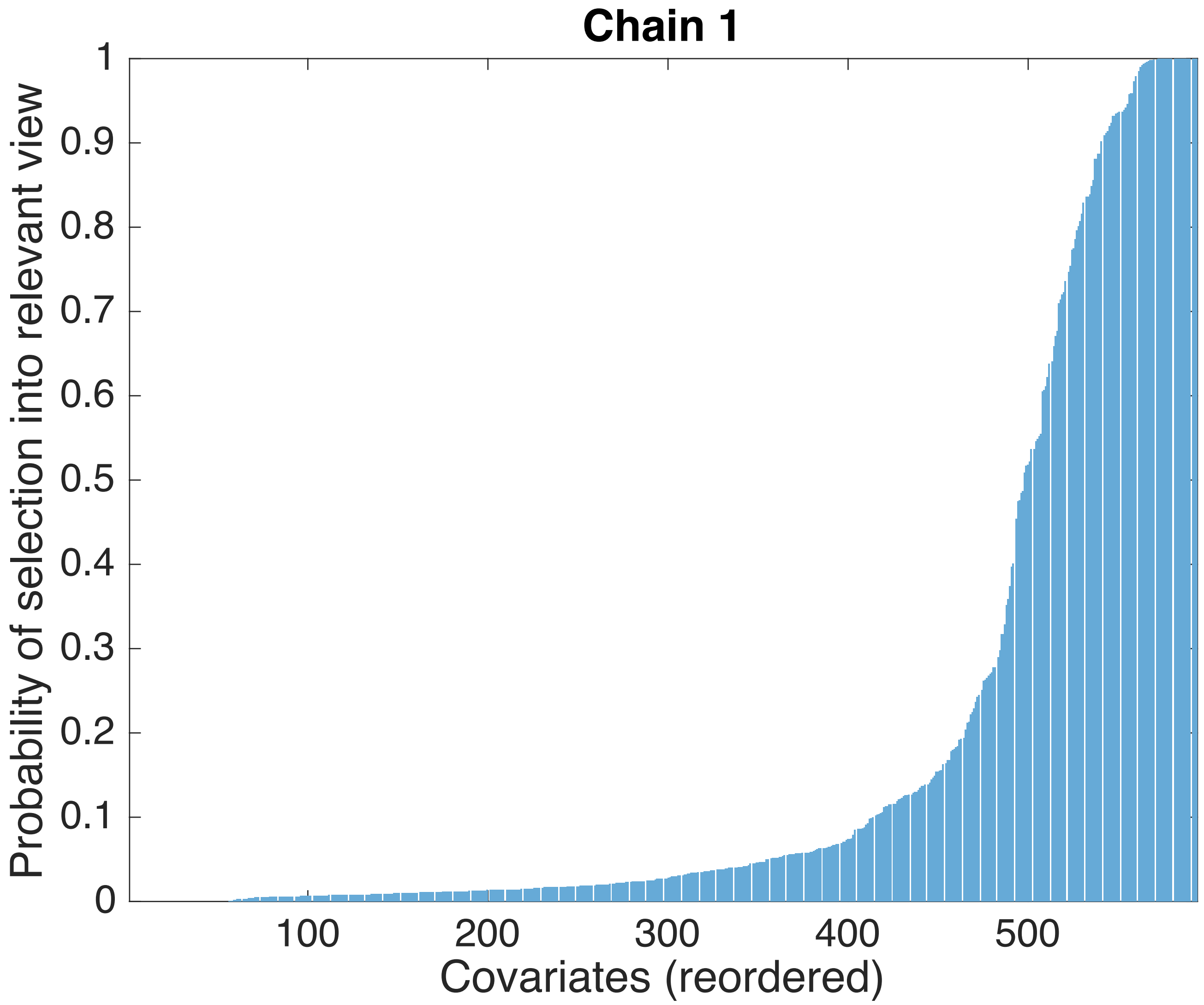
1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

SELECTION PROBABILITIES



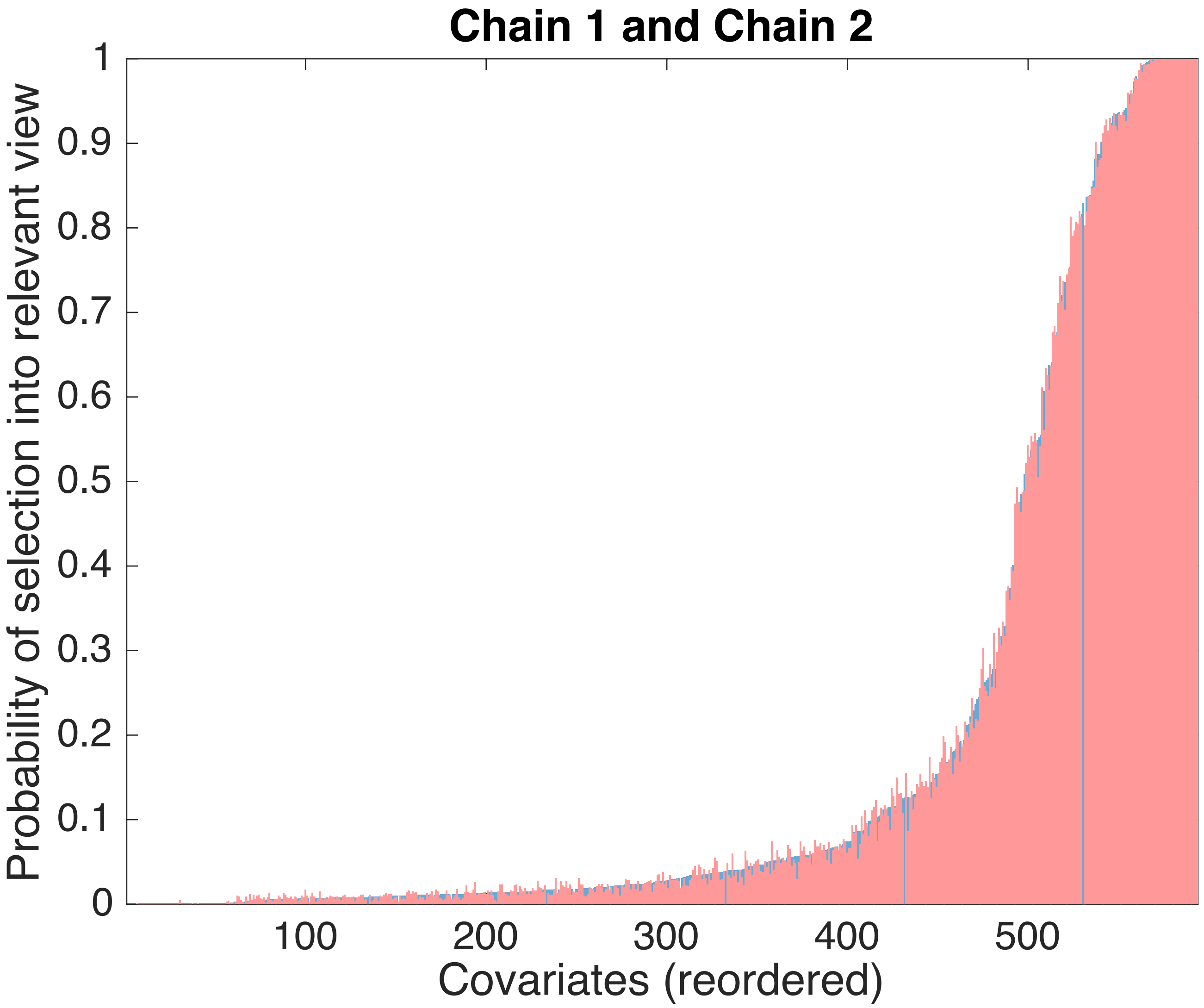
1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

SELECTION PROBABILITIES

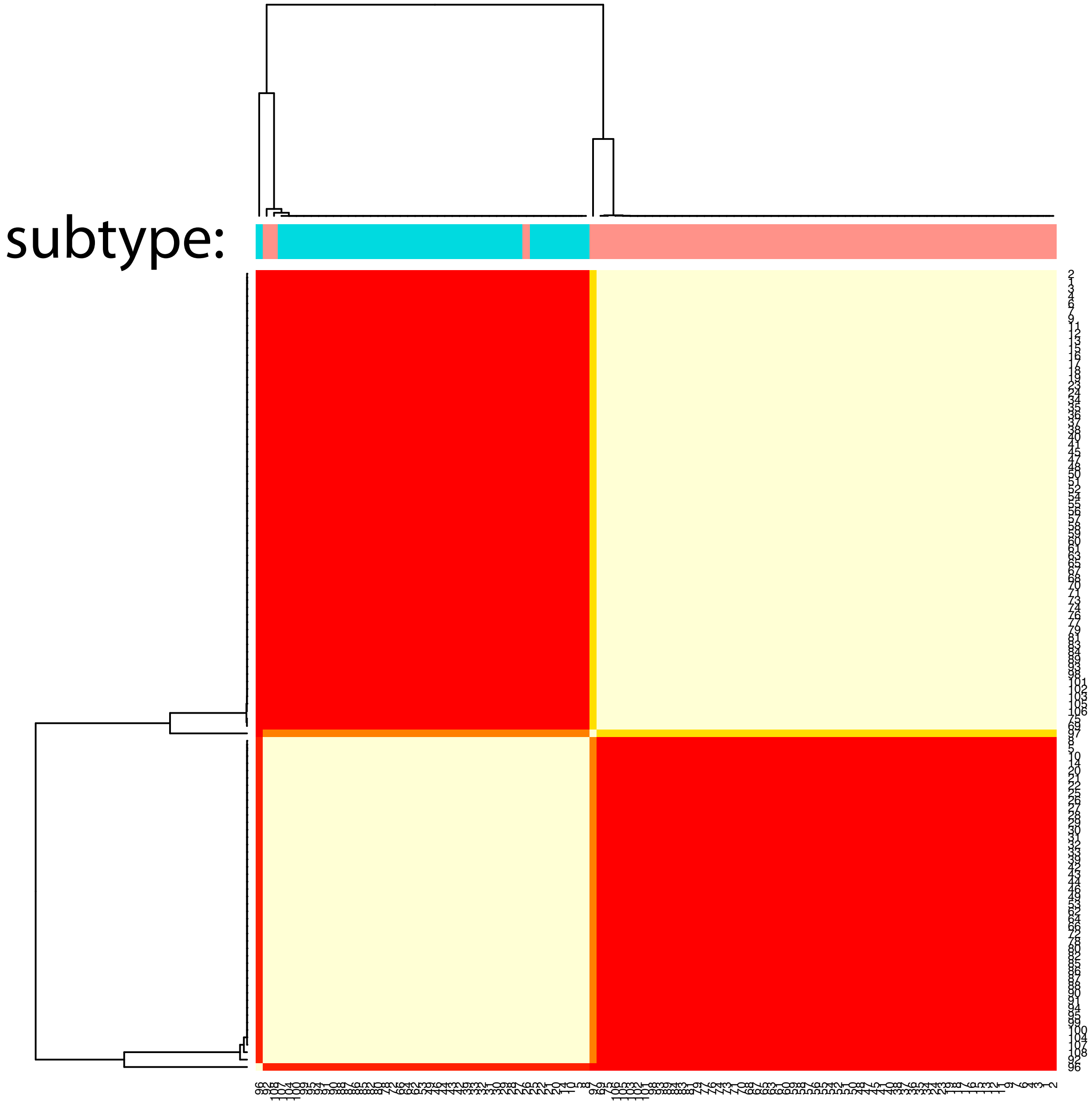


1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

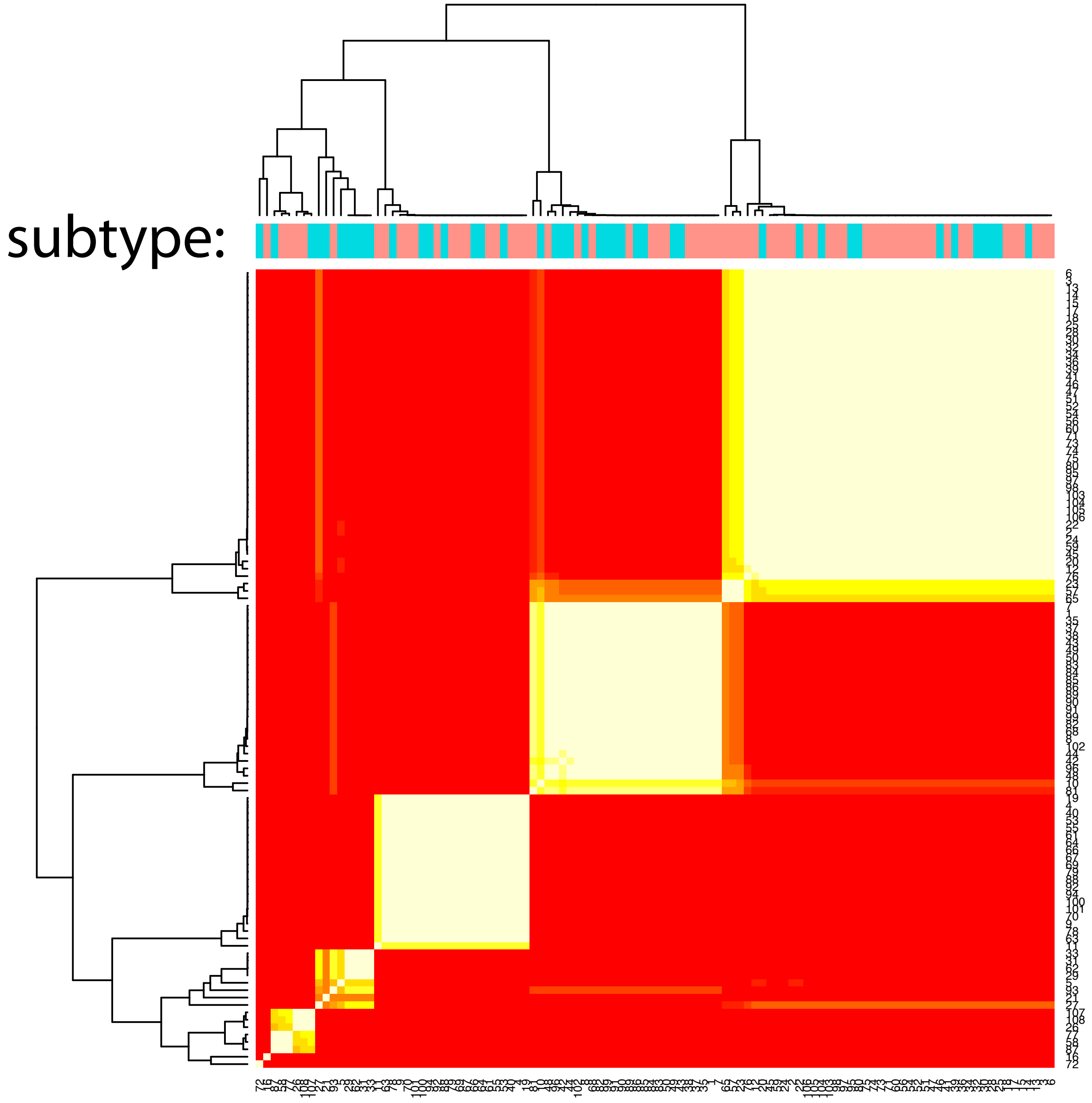
SELECTION PROBABILITIES



1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

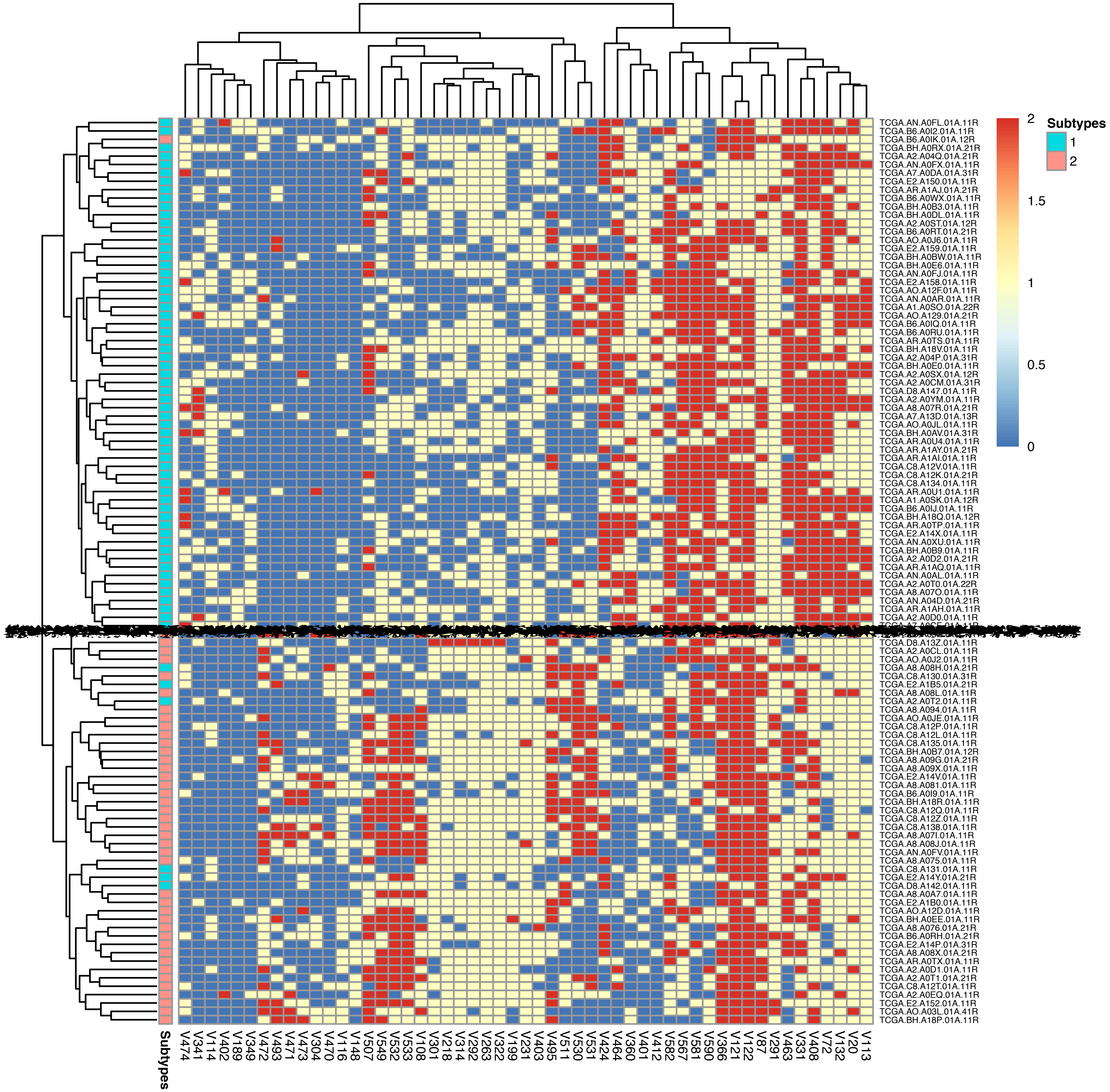


PSM, relevant view

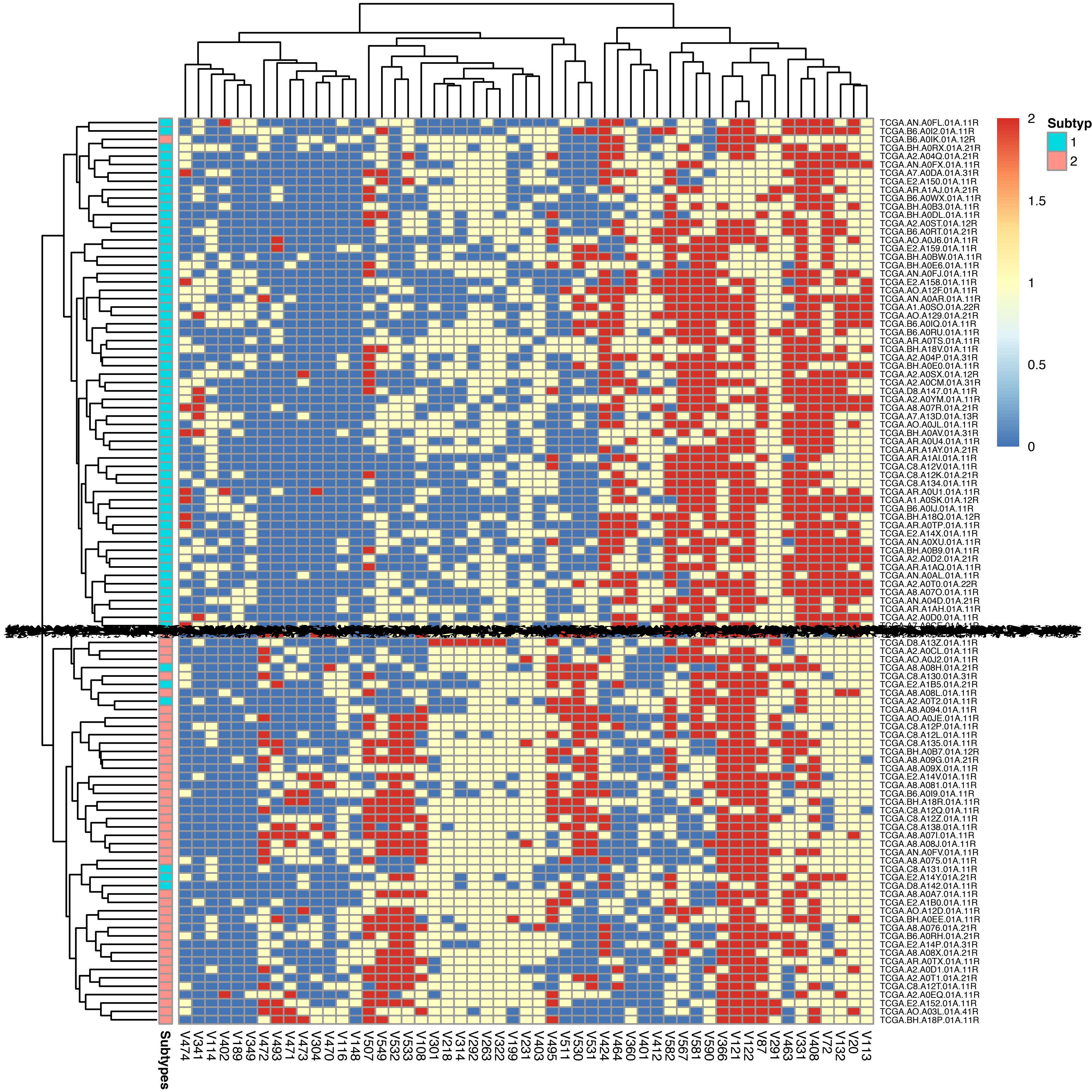


PSM, irrelevant view

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

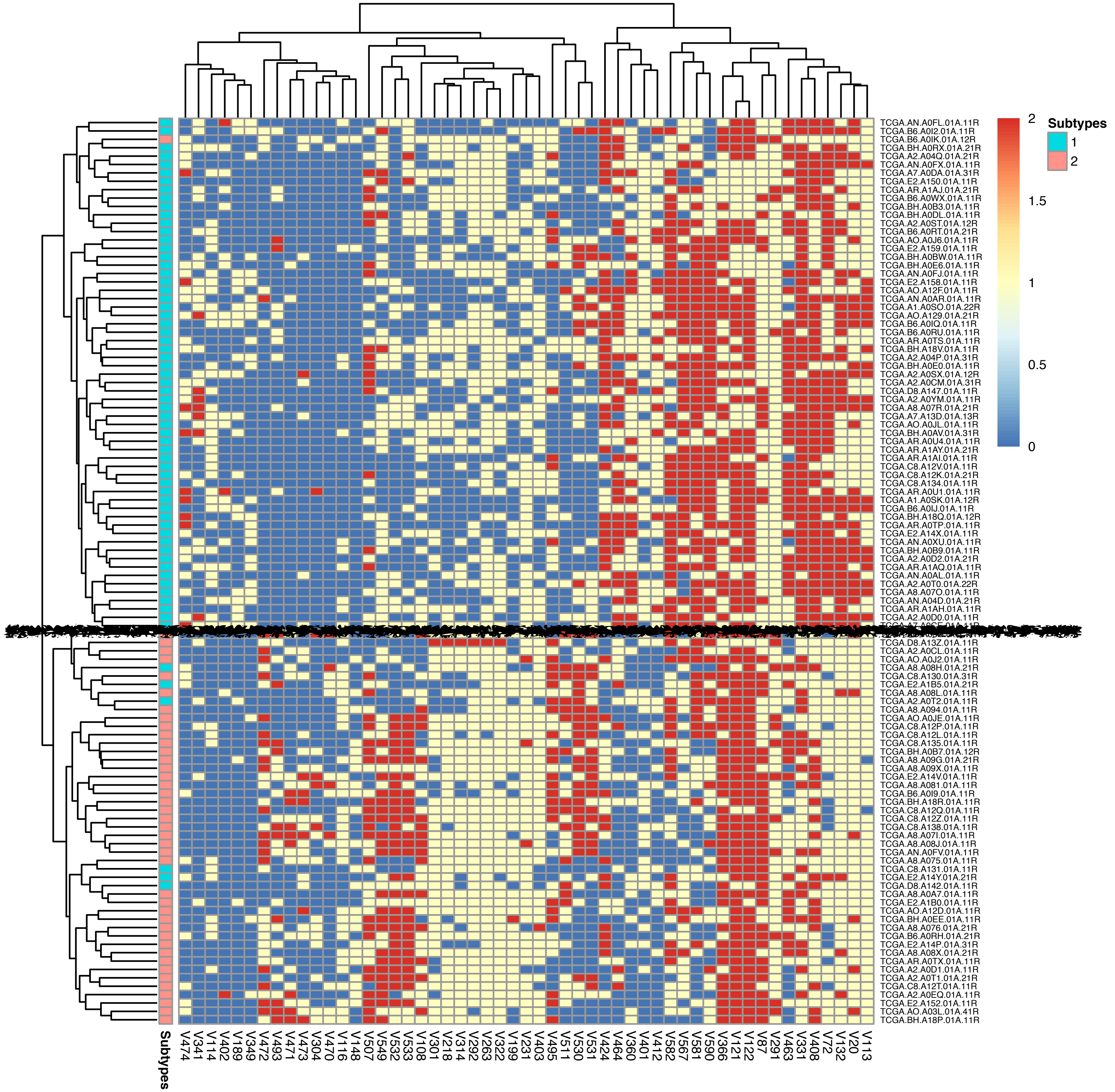


1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS



Retain only variables selected into relevant view at least 90% of the time

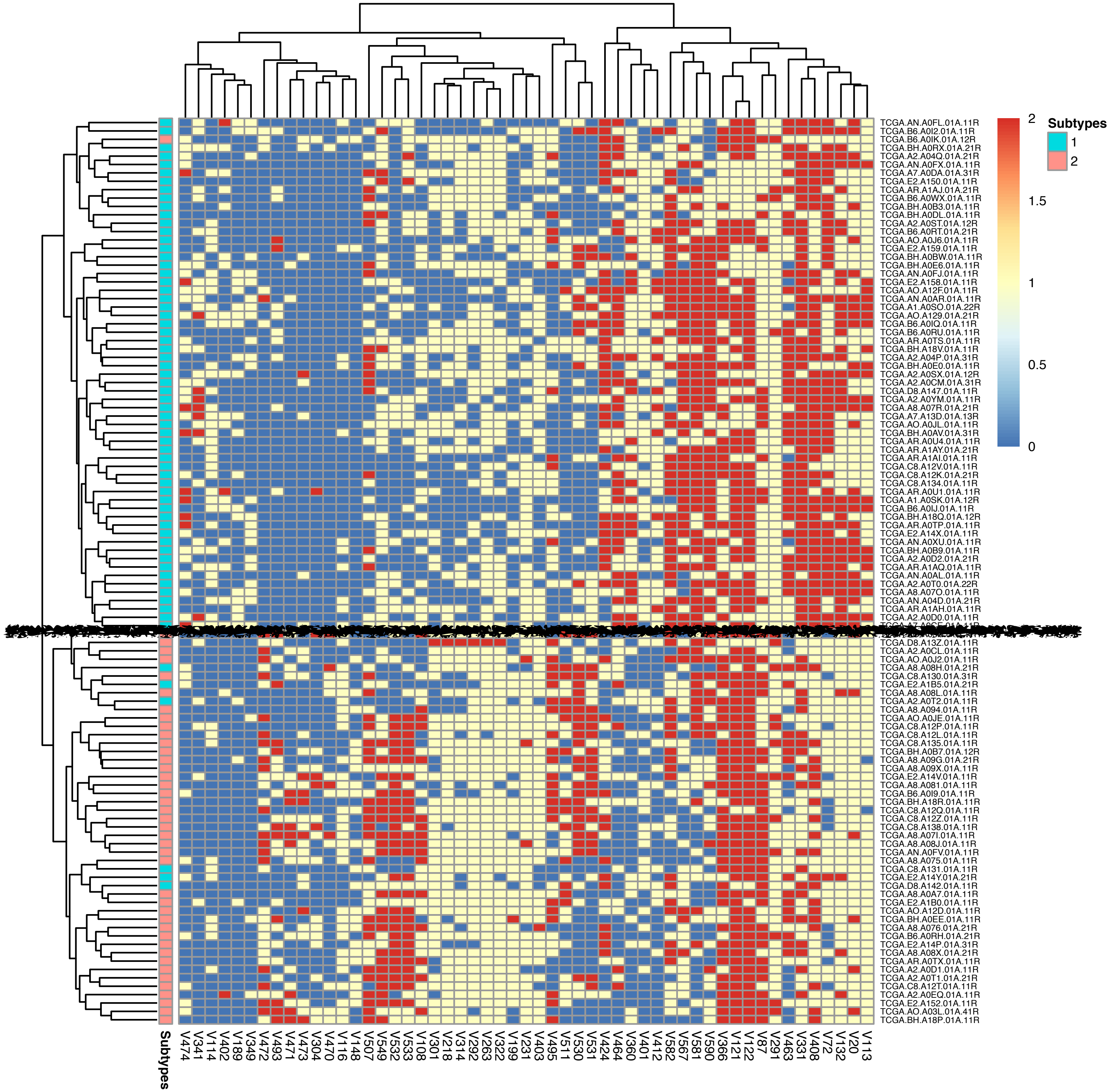
1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS



Retain only variables selected into relevant view at least 90% of the time

53 variables in total:

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

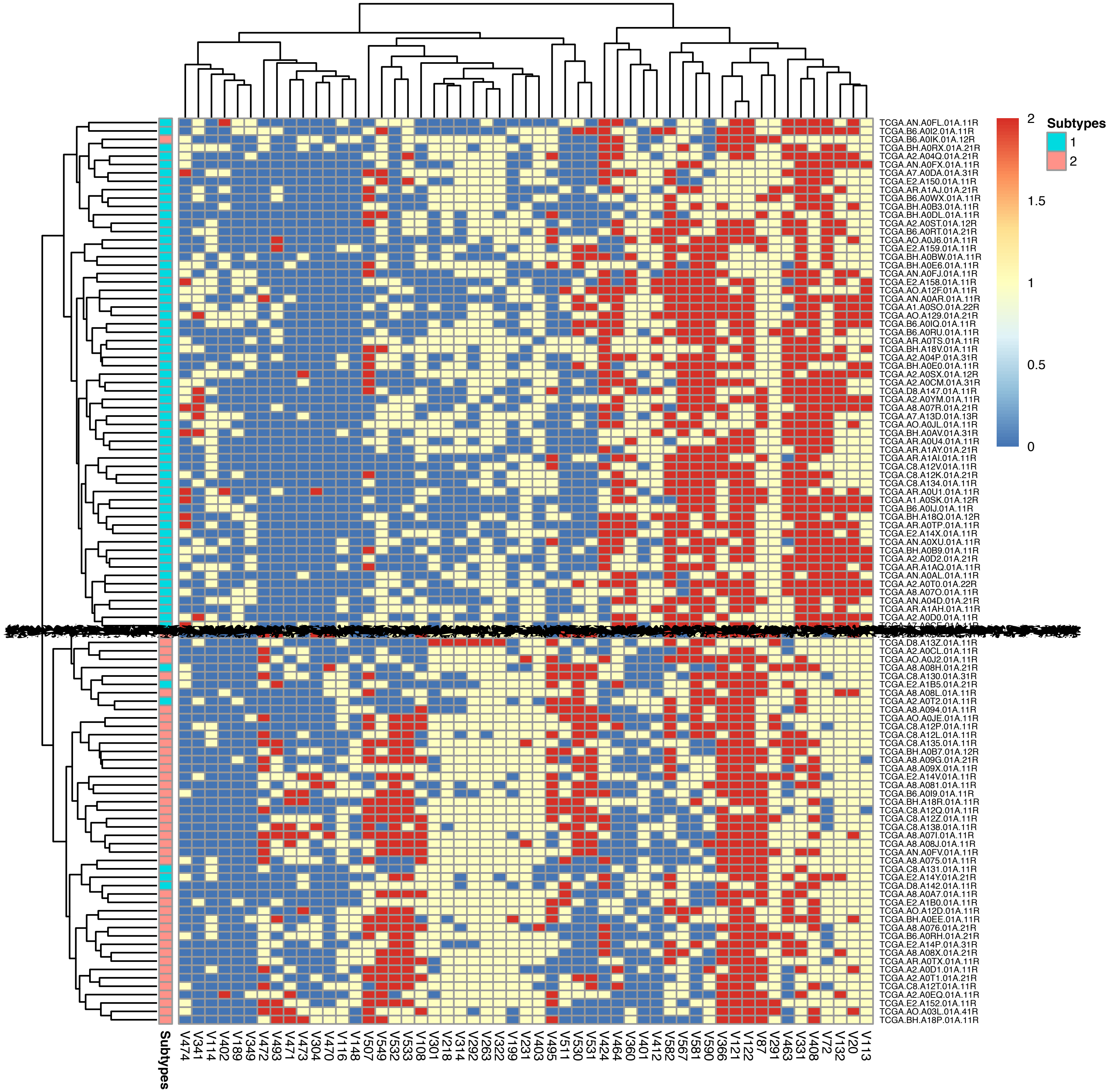


Retain only variables selected into relevant view at least 90% of the time

53 variables in total:

- **32 miRNA**

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

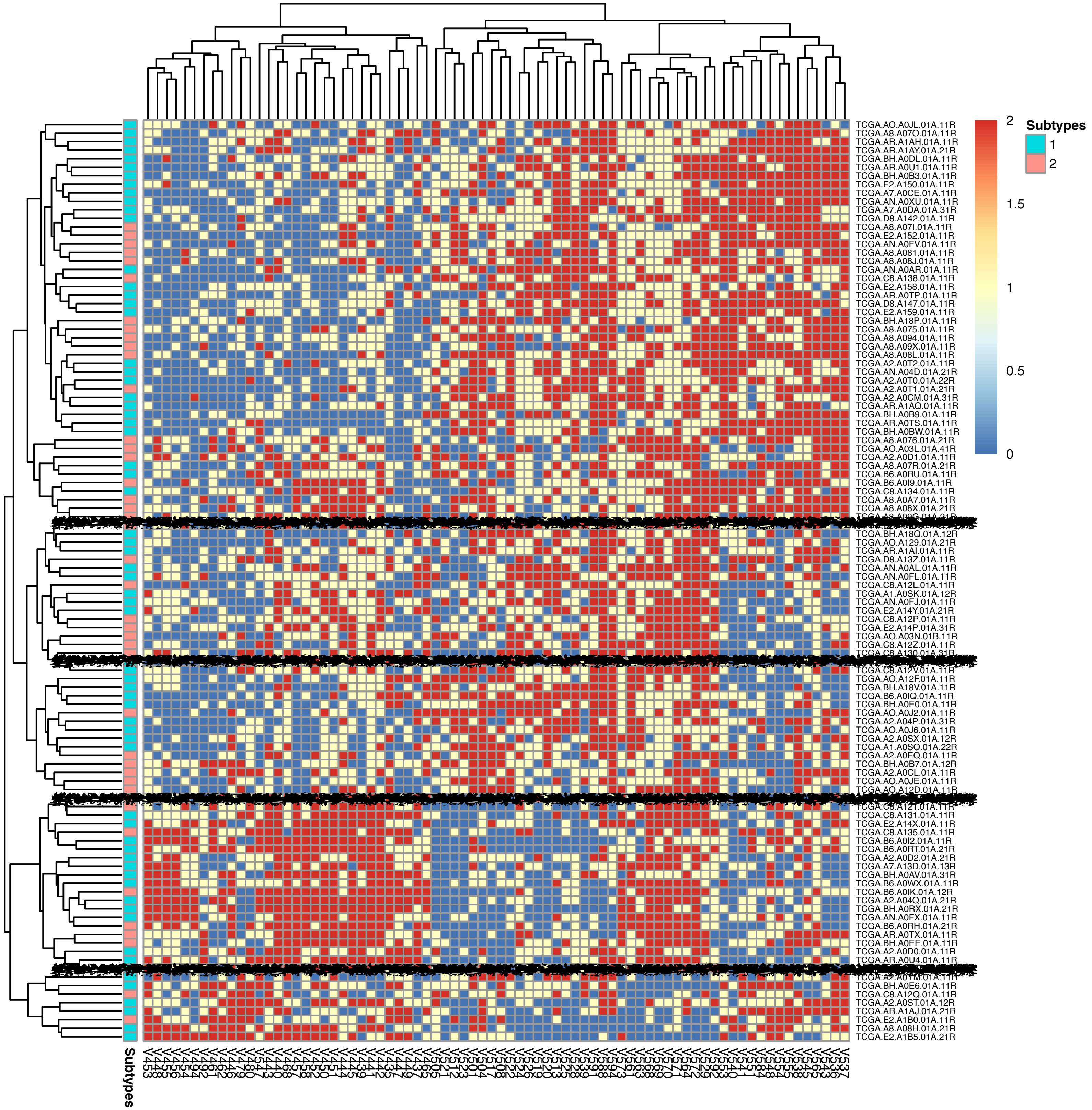


Retain only variables selected into relevant view at least 90% of the time

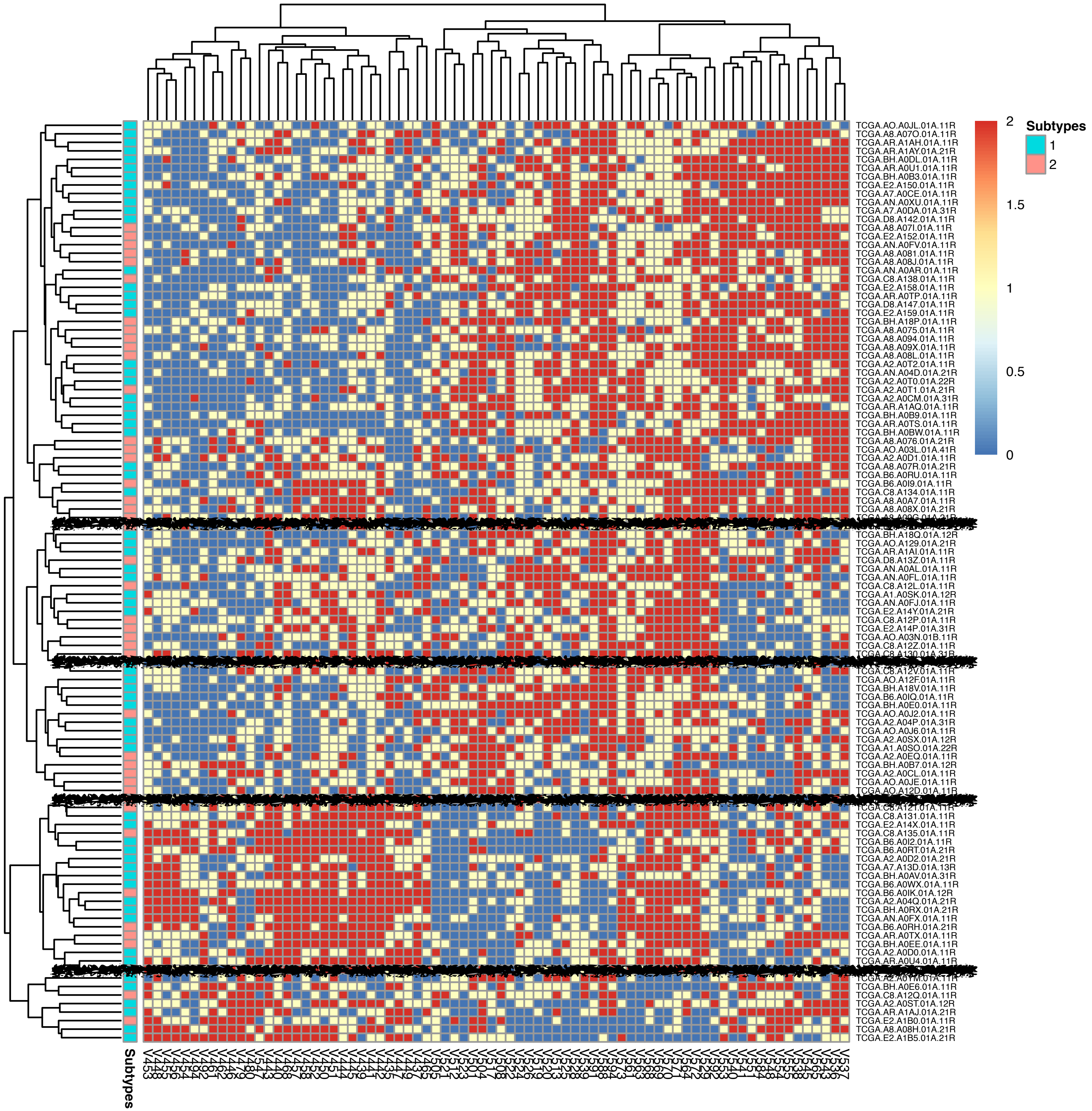
53 variables in total:

- **32 miRNA**
- **21 protein**

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

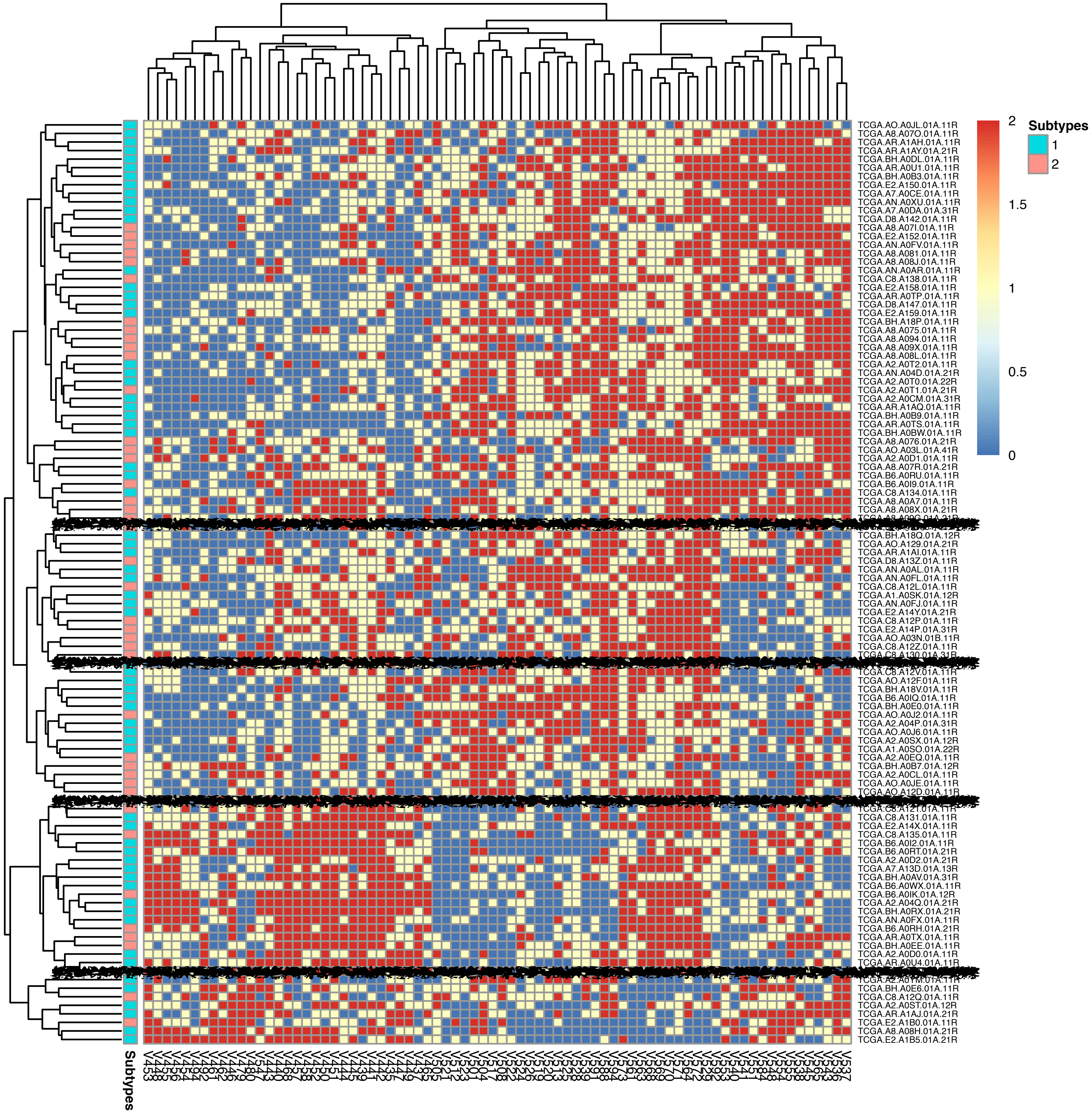


1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS



Retain only variables selected into irrelevant view at least 90% of the time

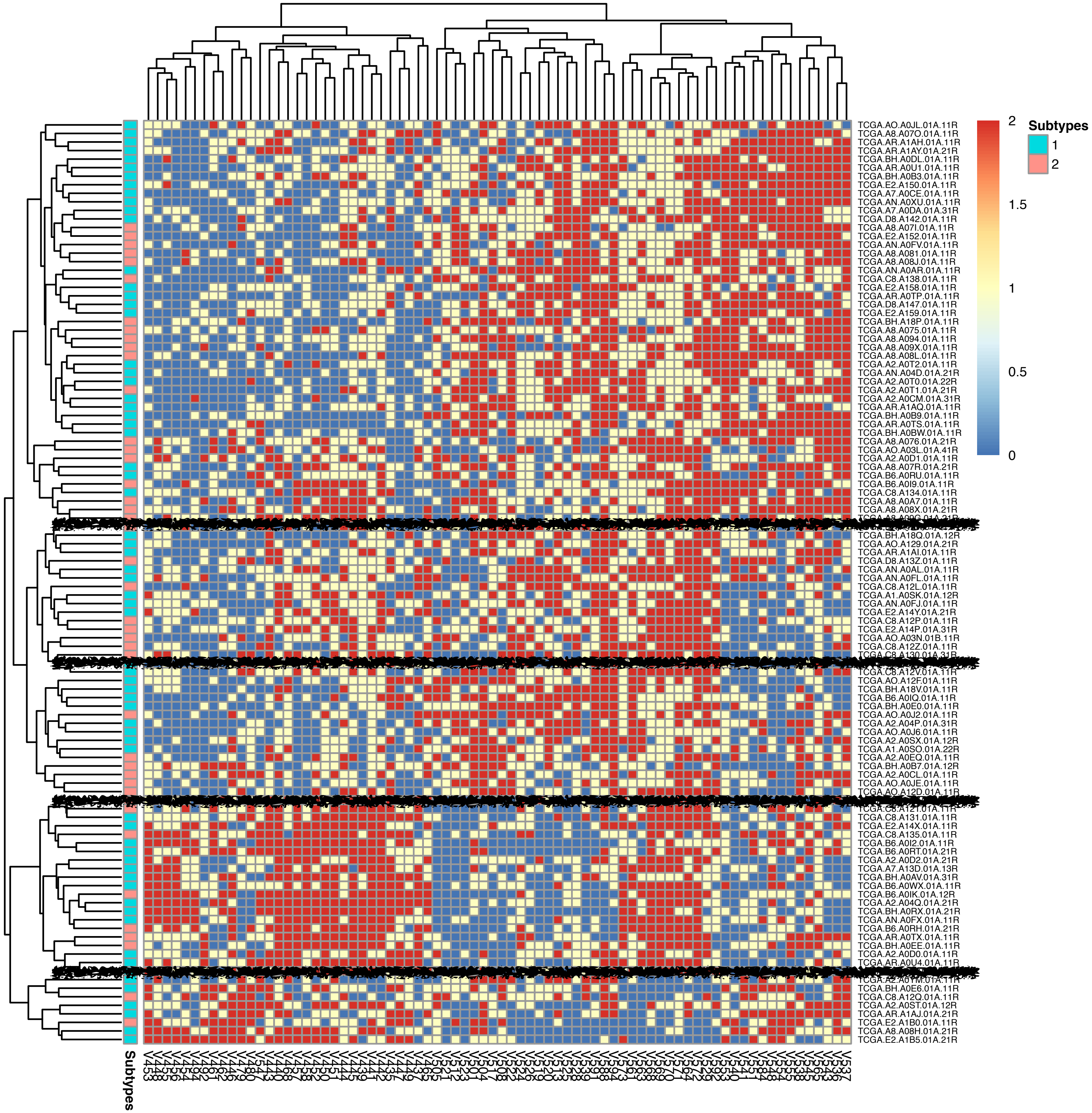
1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS



Retain only variables selected into irrelevant view at least 90% of the time

76 variables in total:

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS

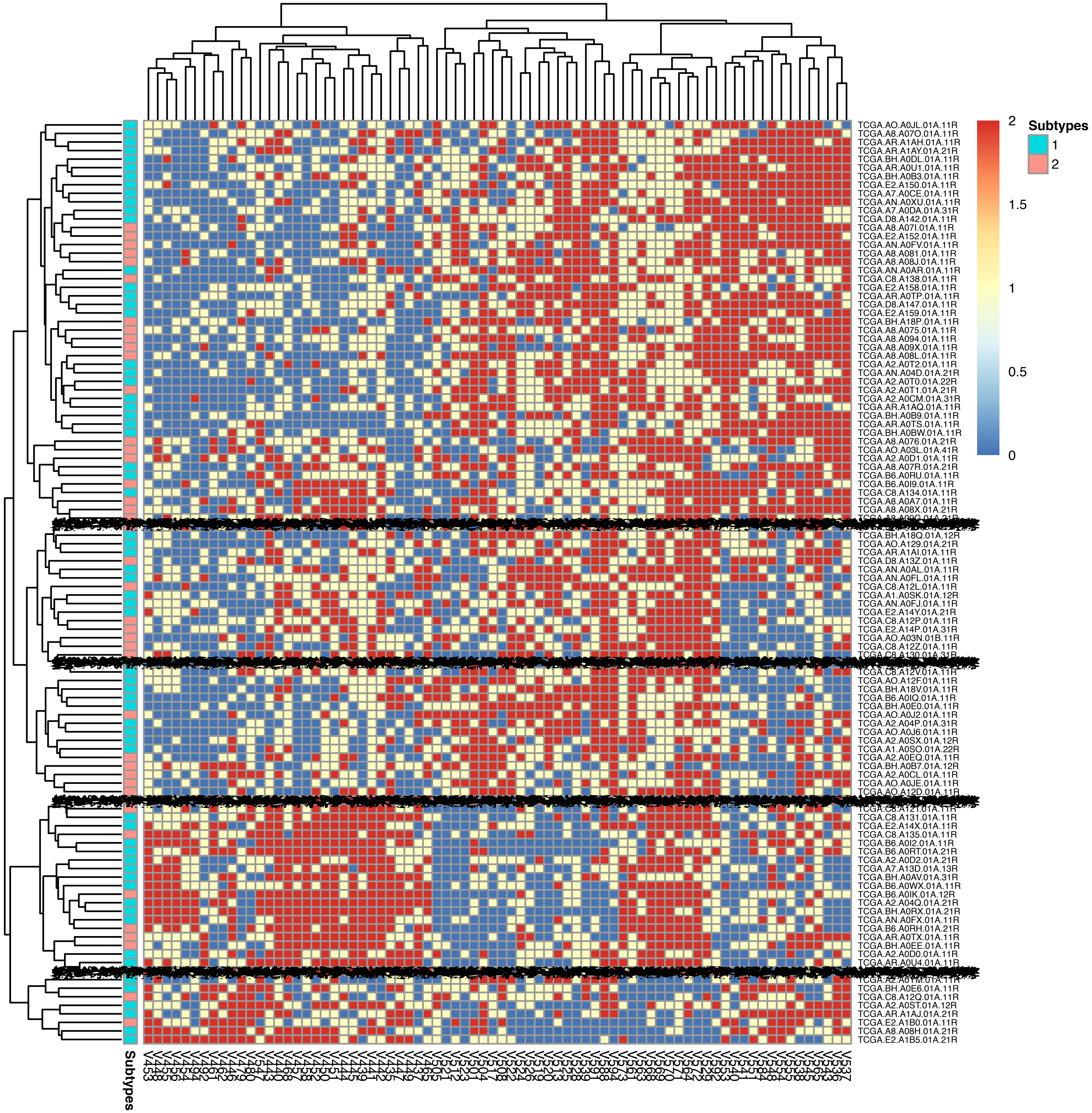


Retain only variables selected into <span style="color:#a00">**irrelevant**</span> **view at least 90% of the time**

76 variables in total:

- **0 miRNA**

1. TCGA BREAST CANCER DATA: INTEGRATING 2 DATASETS



Retain only variables selected into irrelevant view at least 90% of the time

76 variables in total:

- **0 miRNA**
- **76 protein**

PART 5. . .

PART 5. . .

WRAP UP

- In order to determine clinically actionable strata from high-dimensional 'omics data, we should **exploit "response" information** to guide clustering.

- In order to determine clinically actionable strata from high-dimensional ‘omics data, we should **exploit “response” information** to guide clustering.
- Profile regression provides one way to do this, but - for high dimensional data - the **influence of the response can get swamped** by the influence of the variables.

- In order to determine clinically actionable strata from high-dimensional ‘omics data, we should **exploit “response” information** to guide clustering.
- Profile regression provides one way to do this, but - for high dimensional data - the **influence of the response can get swamped** by the influence of the variables.
- **Semi-supervised multiview clustering provides an alternative**, which potentially overcomes this problem

- In order to determine clinically actionable strata from high-dimensional ‘omics data, we should **exploit “response” information** to guide clustering.
- Profile regression provides one way to do this, but - for high dimensional data - the **influence of the response can get swamped** by the influence of the variables.
- **Semi-supervised multiview clustering provides an alternative**, which potentially overcomes this problem
    - and can also be used **for data integration**.

- In order to determine clinically actionable strata from high-dimensional ‘omics data, we should **exploit “response” information** to guide clustering.
- Profile regression provides one way to do this, but - for high dimensional data - the **influence of the response can get swamped** by the influence of the variables.
- **Semi-supervised multiview clustering provides an alternative**, which potentially overcomes this problem
    - and can also be used **for data integration**.
- Still a work in progress!

- Scalability is an issue:

- Scalability is an issue:
    - We have one DP mixture model per view, so usual scalability issues are amplified.

- Scalability is an issue:
    - We have one DP mixture model per view, so usual scalability issues are amplified.
    - There are many opportunities for parallelisation.

- Scalability is an issue:
  - We have one DP mixture model per view, so usual scalability issues are amplified.
  - There are many opportunities for parallelisation.
  - We are also exploring fast approximate inference approaches

- Scalability is an issue:
  - We have one DP mixture model per view, so usual scalability issues are amplified.
  - There are many opportunities for parallelisation.
  - We are also exploring fast approximate inference approaches
    - Variational inference, SUGS, …

- Scalability is an issue:
  - We have one DP mixture model per view, so usual scalability issues are amplified.
  - There are many opportunities for parallelisation.
  - We are also exploring fast approximate inference approaches
    - Variational inference, SUGS, …
- Gibbs sampling can be **slow to converge** and **mix poorly**

- Scalability is an issue:
  - We have one DP mixture model per view, so usual scalability issues are amplified.
  - There are many opportunities for parallelisation.
  - We are also exploring fast approximate inference approaches
    - Variational inference, SUGS, …
- Gibbs sampling can be **slow to converge** and **mix poorly**
  - **Split-merge** procedures are useful

- Scalability is an issue:
  - We have one DP mixture model per view, so usual scalability issues are amplified.
  - There are many opportunities for parallelisation.
  - We are also exploring fast approximate inference approaches
    - Variational inference, SUGS, …
- Gibbs sampling can be **slow to converge** and **mix poorly**
  - **Split-merge** procedures are useful
- Just starting to assess the importance of getting the “right” number of views.

THANKS FOR LISTENING!

MRC | Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

@pauldwkirk

http://www.mrc-bsu.cam.ac.uk/people/in-alphabetical-order/h-to-m/paul-kirk/