Prediction of **U**nderlying **L**atent **C**lasses via **K**-means and **H**ierarchical **C**lustering **A**lgorithm

Guan-Hua Huang, Su-Mei Wang and Chung-Chu Hsu

07/07/2010

Breast cancer data

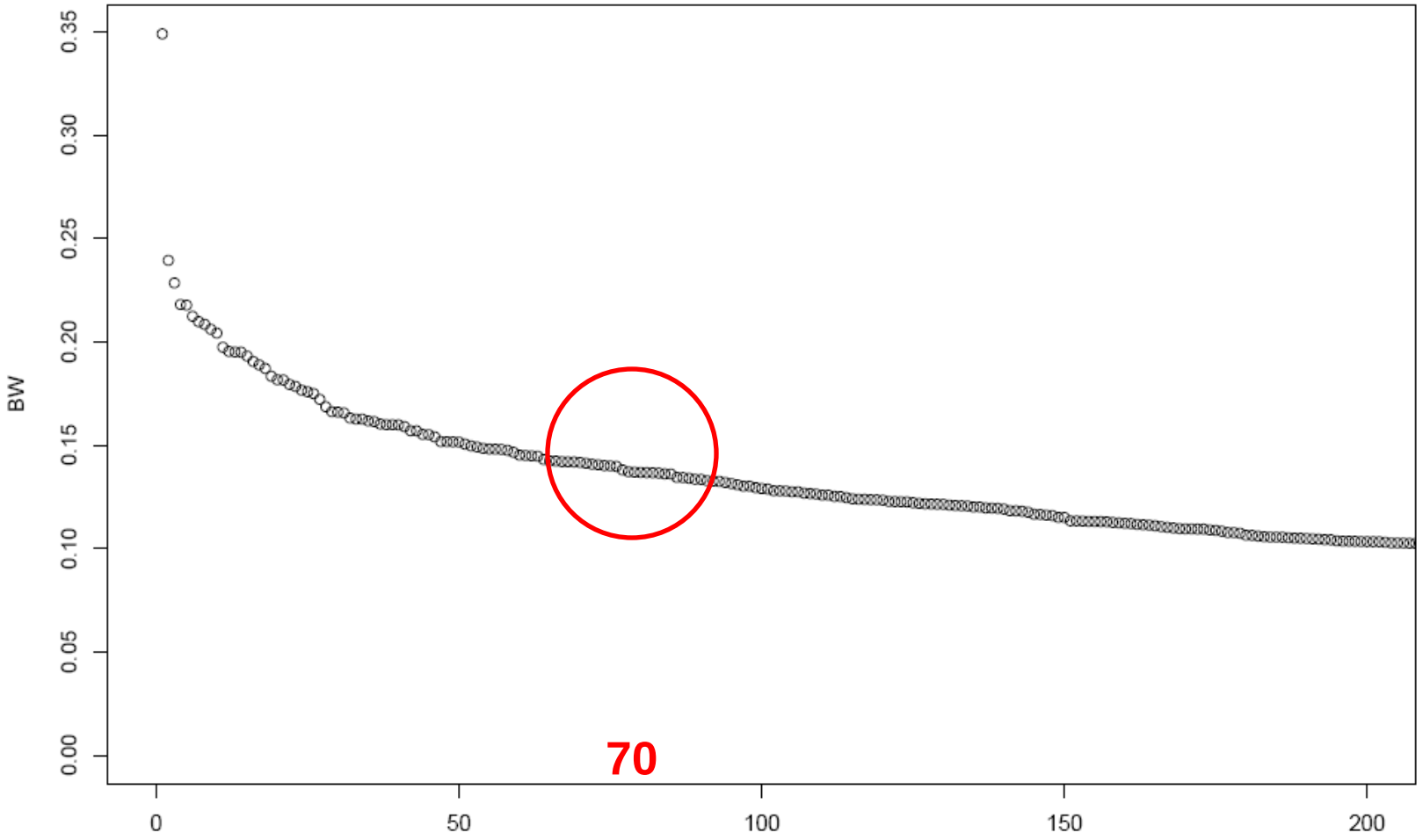
- van't Veer *et al*. Nature 2002
- The 78 sporadic lymph-node-negative breast cancer patients
  - 44 remained free of disease for an interval of at least 5 years (good prognosis group)
  - 34 had developed distant metastases within 5 years (poor prognosis group).
- Aim to predict good and poor prognostic patients through gene expression profiling

Breast cancer data (cont'd)

- A preliminary two-step gene selection process (from 24481 genes):
  - 4741 genes with the intensity ratio more than two-fold difference and the significance of regulation p-value < 0.01 in more than 3 patients
  - Apply a selection of genes based on the ratio of their between-group to within-group sums of squares

$$BW(m) = \frac{\sum_i \sum_c I(d_i = c)(\bar{y}_{cm} - \bar{y}_{.m})^2}{\sum_i \sum_c I(d_i = c)(y_{im} - \bar{y}_{cm})^2}$$

BW plot



Breast cancer data (cont'd)

- Using 70 selected gene expression ratios as observed surrogates, a finite mixture model was fitted.

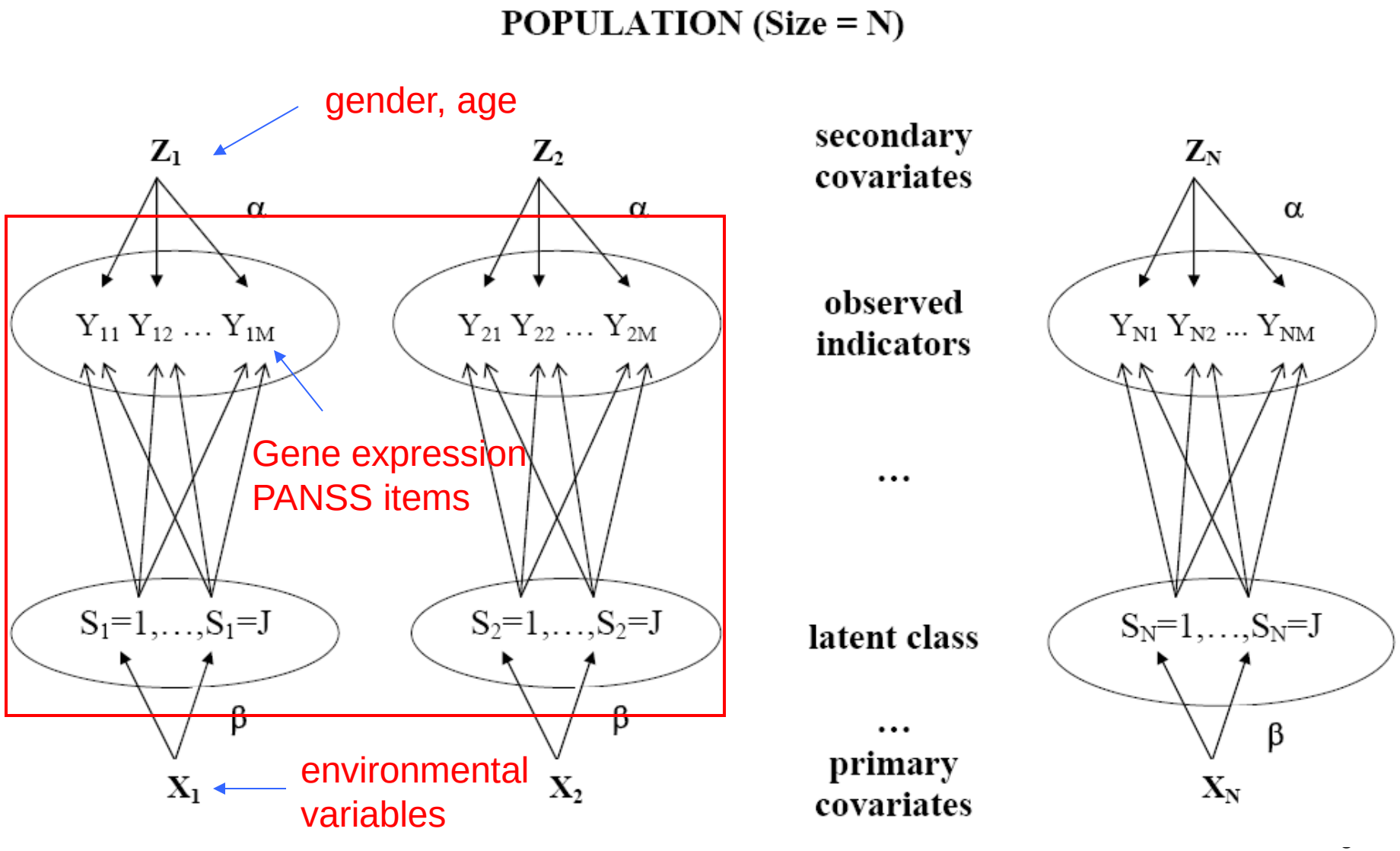
Schizophrenia data

- The data were collected from a series of projects for schizophrenia (Dr. Hai-Gwo Hwu).
- The analyzed data include
  - 169 acute patients of schizophrenia who were recruited within one week of index admission
  - 160 subsided state patients who were living with community and under family care
- Aim to
  - explore the subtypes of schizophrenia patients
  - predict patients' phases of chronicity

Schizophrenia data (cont'd)

- Schizophrenia symptoms were assessed by the PANSS:
  - 30 items and consists of three subscales: positive, negative and general psychopathology
  - Each item was originally rated on a 7-point scale (1=absent, 7=extreme), but we reduced the 7-point scale by merging the points that had the response percentages less than 10%

Models


POPULATION (Size = N)
gender, age
Z_1
Z_2
Z_N
secondary covariates
α
α
α
Y_{11} Y_{12} … Y_{1M}
Y_{21} Y_{22} … Y_{2M}
Y_{N1} Y_{N2} … Y_{NM}
observed indicators
Gene expression
PANSS items
…
S_1=1,…,S_1=J
S_2=1,…,S_2=J
S_N=1,…,S_N=J
latent class
…
β
β
β
X_1
X_2
X_N
environmental variables
primary covariates

Introduction

- Finite mixture model is an analogy of cluster analysis.
- Finite mixture model classifies objects based on their responses to a set of surrogates.
- Measured surrogates are assumed independent of one another within any category of the underlying latent variable.
- Use k-means and hierarchical clustering methods with covariance among surrogates as the distance measure.

Finite mixture model

$\mathbf{Y}_i = (Y_{i1}, \cdots, Y_{iM})^T$: M observable surrogates

$$f(y_{i1}, \ldots, y_{iM}) = \sum_{j=1}^{J} \{\Pr(S_i = j) f(y_{i1}, \cdots, y_{iM} \mid S_i = j)\}$$

$$= \sum_{j=1}^{J} \left\{ \Pr(S_i = j) \prod_{m=1}^{M} f(y_{im} \mid S_i = j) \right\}$$

Latent Class Membership Estimation

Background

- The key is to estimate the latent class membership.
- Use K-means and hierarchical clustering methods to group the objects such that observed variables are statistically independent within latent classes.
- Use sample covariance matrix as the independence measurement.

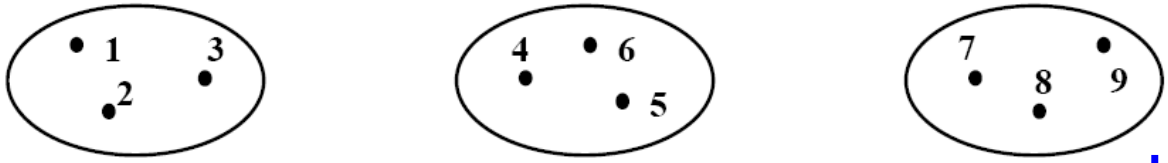
Independence measurement

- Supposed $\tilde{\mathbf{Y}}_i = (Y_{i1}, Y_{i2}, \cdots, Y_{iM})$

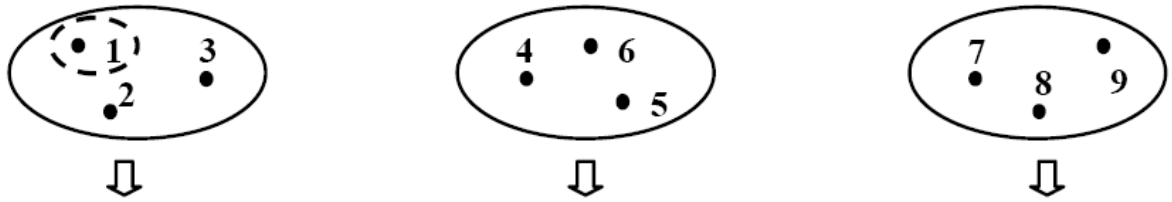
  Then,

$$\mathrm{Cov}(\tilde{\mathbf{Y}}_i) = \begin{bmatrix} \mathrm{cov}(Y_{i1}, Y_{i1}) & \mathrm{cov}(Y_{i1}, Y_{i2}) & \cdots & \mathrm{cov}(Y_{i1}, Y_{iM}) \\ \mathrm{cov}(Y_{i2}, Y_{i1}) & \mathrm{cov}(Y_{i2}, Y_{i2}) & \cdots & \mathrm{cov}(Y_{i2}, Y_{iM}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(Y_{iM}, Y_{i1}) & \mathrm{cov}(Y_{iM}, Y_{i2}) & \cdots & \mathrm{cov}(Y_{iM}, Y_{iM}) \end{bmatrix}$$

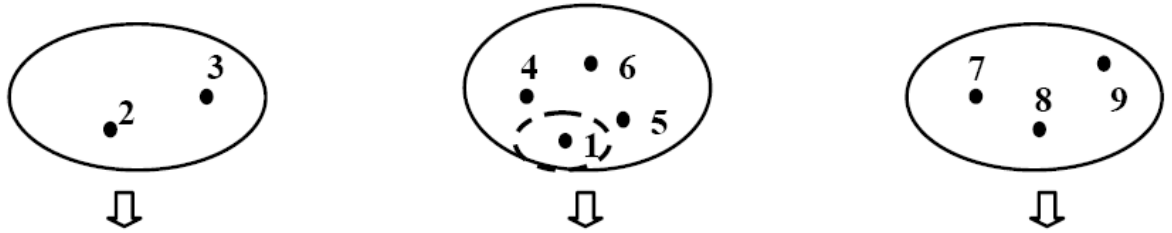
- $\mathrm{ACov} = mean(|\, entries \text{ in } non-diagonal-block\,|)$

Initial classes 

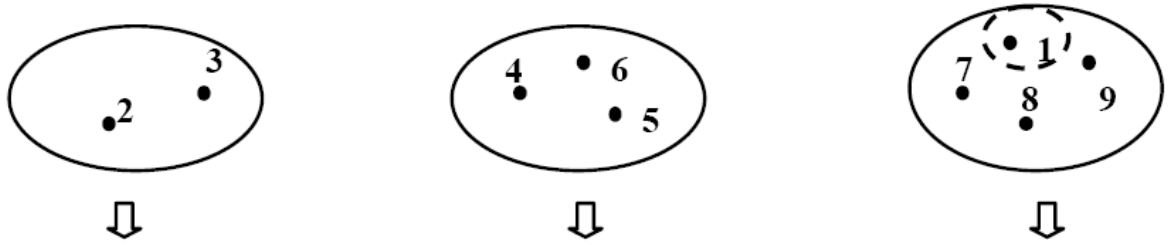
K-means

For object 1: 

assign to class 1 $(3/9)\ ACov_1^{(1)} + (3/9)\ ACov_2^{(1)} + (3/9)\ ACov_3^{(1)} = LoI^{(1)}$



assign to class 2 $(2/9)\ ACov_1^{(2)} + (4/9)\ ACov_2^{(2)} + (3/9)\ ACov_3^{(2)} = LoI^{(2)}$

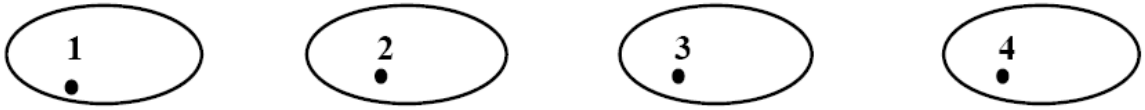


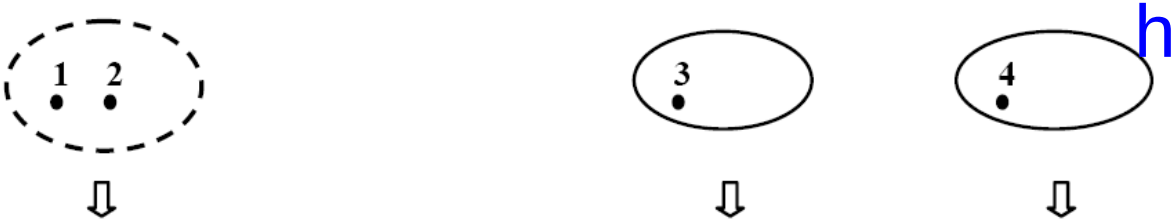
assign to class 3 $(2/9)\ ACov_1^{(3)} + (3/9)\ ACov_2^{(3)} + (4/9)\ ACov_3^{(3)} = LoI^{(3)}$

=> Assign object 1 to the class corresponding to minimum LoI

Agglomerative hierarchical

Initial





merge 1, 2 $(2/4)\mathrm{ACov}_{(12)} + (1/4)\mathrm{ACov}_3 + (1/4)\mathrm{ACov}_4 = \mathrm{LoI}^{(12)}$

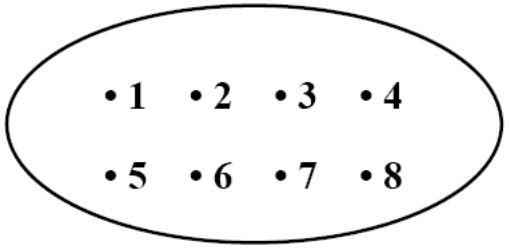
merge 1, 3 $(2/4)\mathrm{ACov}_{(13)} + (1/4)\mathrm{ACov}_2 + (1/4)\mathrm{ACov}_4 = \mathrm{LoI}^{(13)}$

merge 3, 4 $(1/4)\mathrm{ACov}_1 + (1/4)\mathrm{ACov}_2 + (2/4)\mathrm{ACov}_{(34)} = \mathrm{LoI}^{(34)}$

=> Merge the pair of classes whose combination results in the minimum LoI

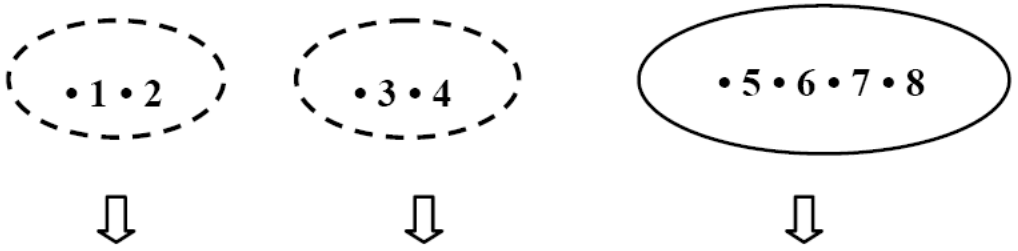
Divisive hierarchical

Preliminary class



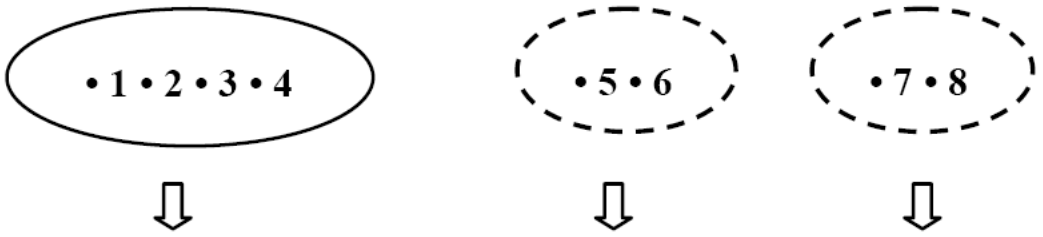
Split by 2-means

Split class 1 by 2-means



$$(2/8)\,\mathrm{ACov}_{(1)_1} + (2/8)\,\mathrm{ACov}_{(1)_2} + (4/8)\,\mathrm{ACov}_2 = \mathrm{LoI}^{(1)}$$

Split class 2 by 2-means



$$(4/8)\,\mathrm{ACov}_1 + (2/8)\,\mathrm{ACov}_{(2)_1} + (2/8)\mathrm{ACov}_{(2)_2} = \mathrm{LoI}^{(2)}$$

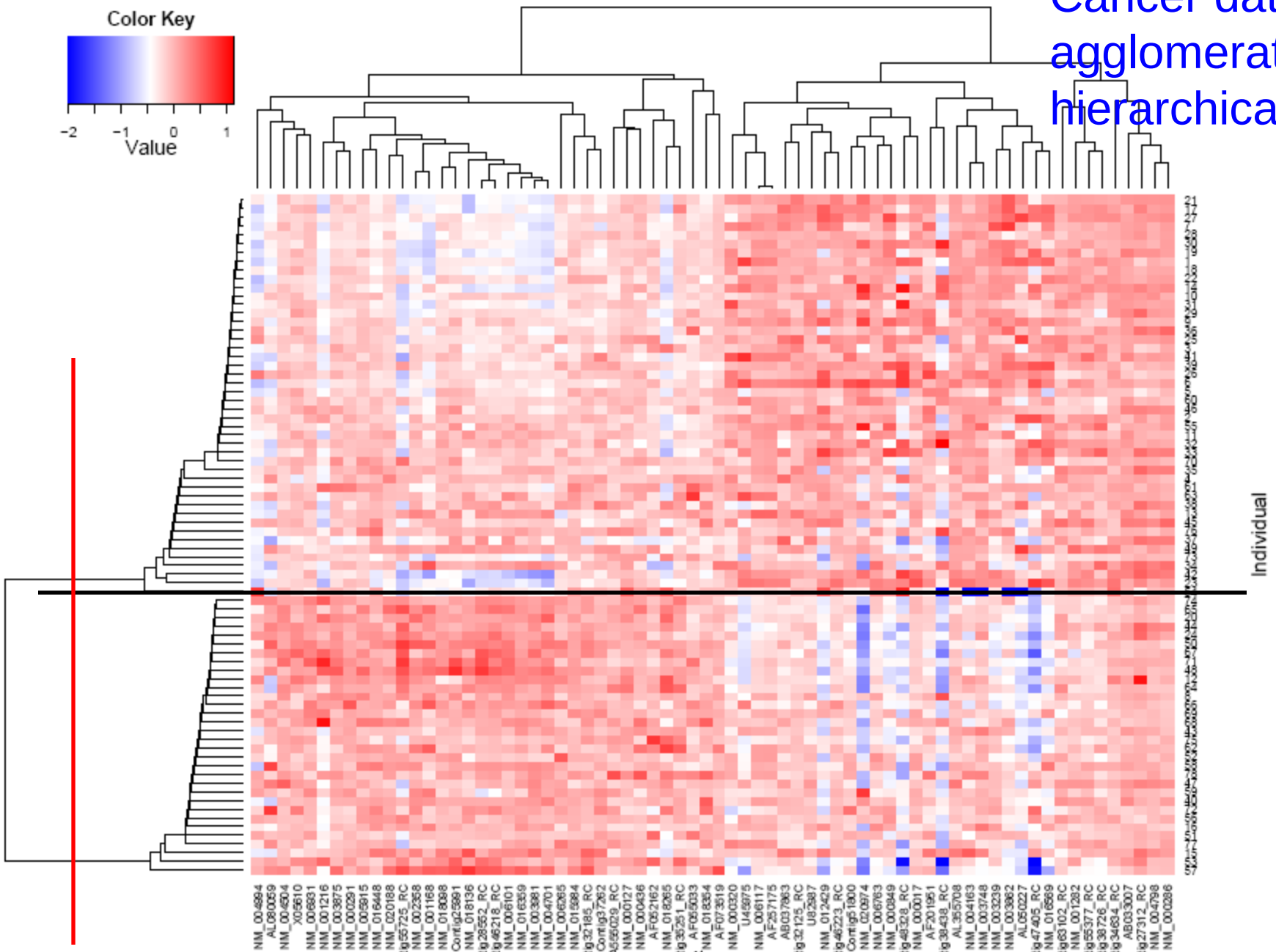
=> Split the class whose division results in the minimum LoI

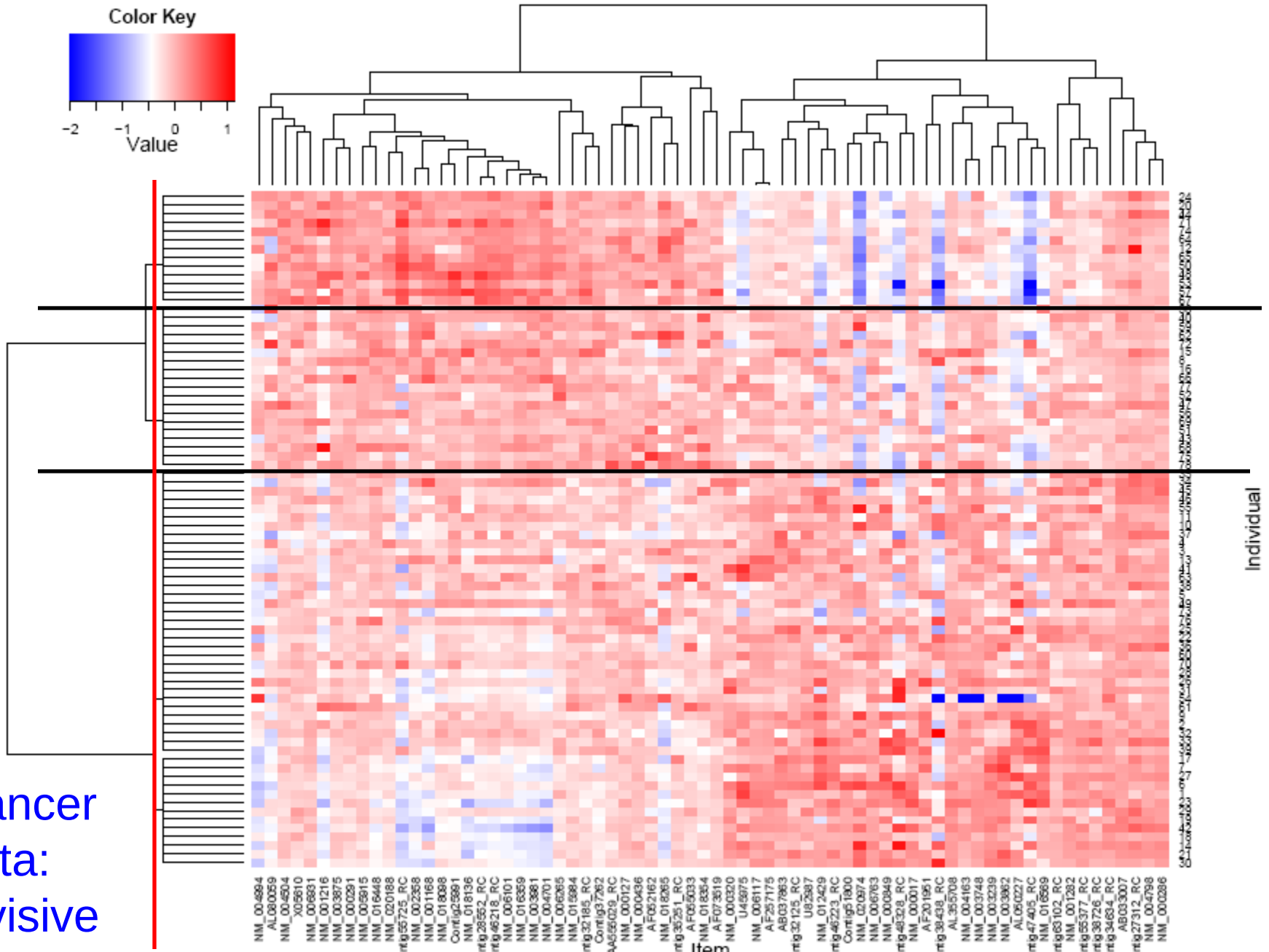
Classification using finite mixture models

- For a new object $Y^* = (Y_1^*, \cdots, Y_M^*)$ with the disease status D^*

$$\Pr(D^* = c \mid Y^*) = \sum_{j=1}^{J} \left\{ \Pr(D^* = c \mid S^* = j, Y^*) \times \Pr(S^* = j \mid Y^*) \right\}$$

- Allocate Y^* to $D^*=c^*$ at which the maximum estimated posterior probability is reached


Color Key
−2
−1
0
1
Value
Cancer data: agglomerative hierarchical
Individual
Item



Cancer data: divisive hierarchical

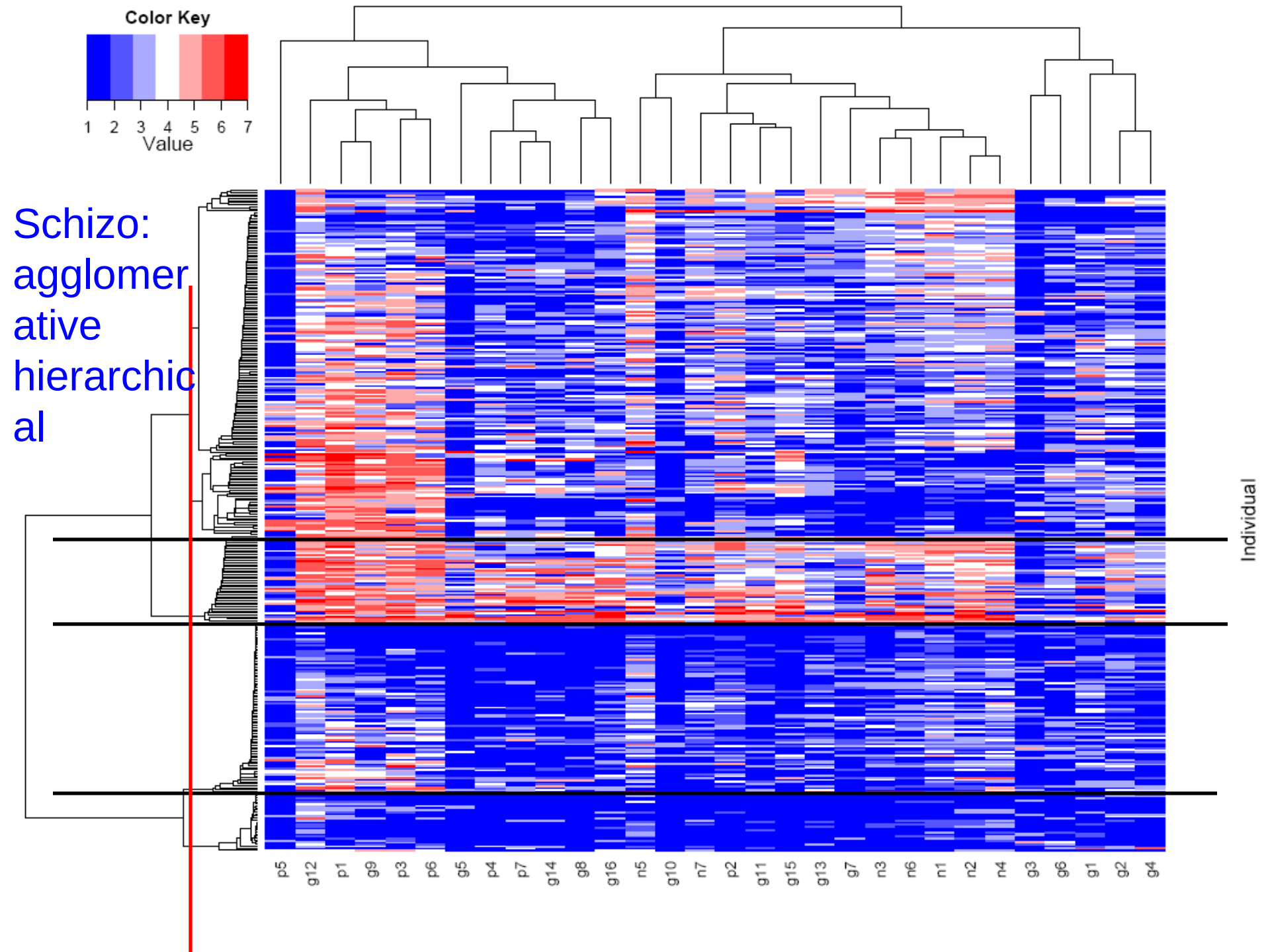
Leave-one-out cross-validation

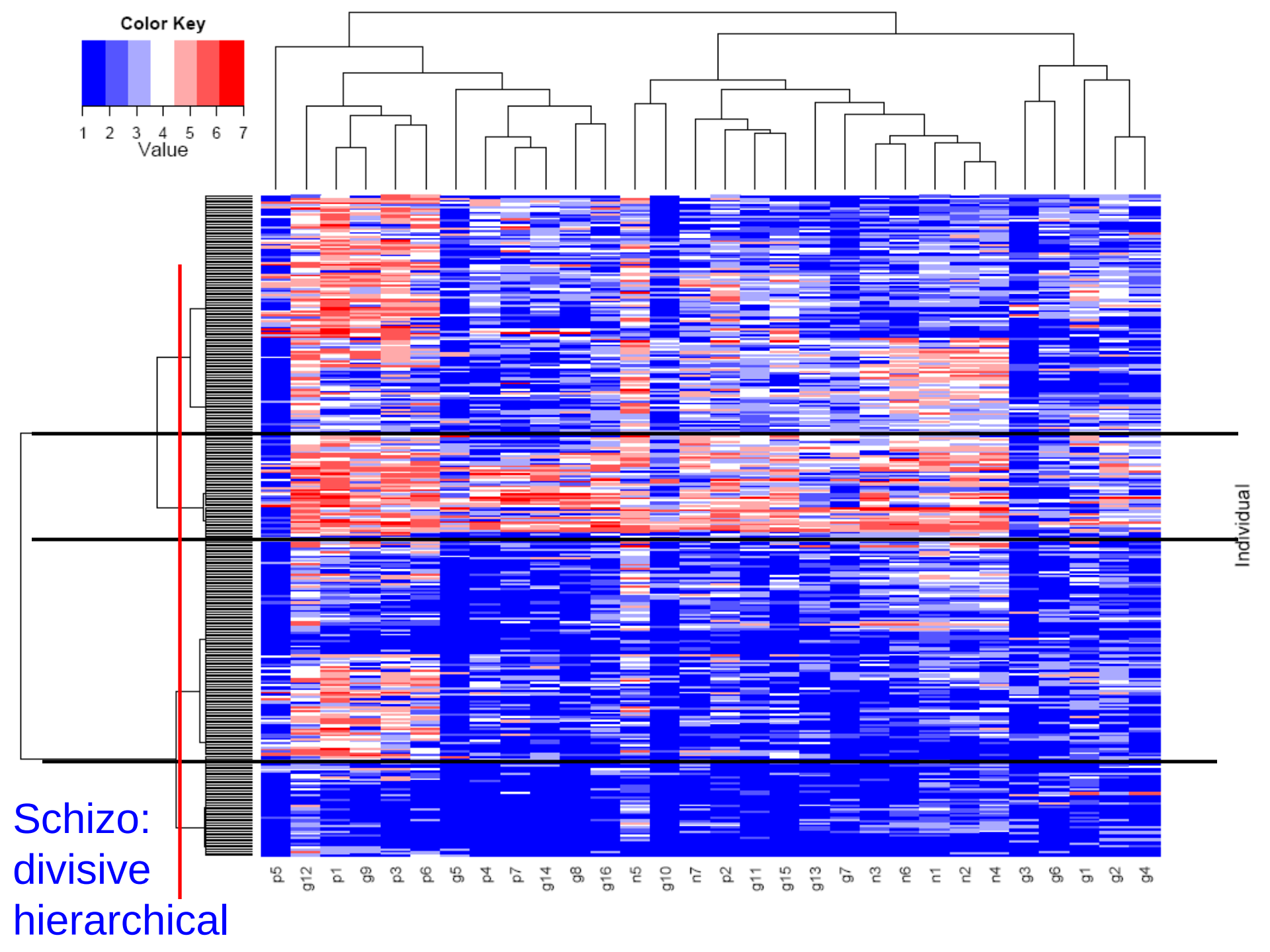
- Misclassification rates in predicting poor vs. good prognosis
  - k-means: 24.36%
  - agglomerative hierarchical: 26.92%
  - divisive hierarchical: 29.49%

Additional independent test set

- Independent 19 young, lymph-node-negative breast cancer patients:
  - 12 poor prognosis
  - 7 good prognosis

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
True	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
KM	0	0	0	0	0	0	0	0	0	0	0	0	1	**0**	1	**0**	1	**0**	1
AH	0	0	0	**1**	0	0	0	0	0	0	0	**1**	1	1	1	1	1	**0**	1
DH	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	**0**	1	**0**	1


Color Key
1 2 3 4 5 6 7
Value
Schizo: agglomerative hierarchical
Individual
p5 g12 p1 g9 p3 p6 g5 p4 p7 g14 g8 g16 n5 g10 n7 p2 g11 g15 g13 g7 n3 n6 n1 n2 n4 g3 g6 g1 g2 g4



Schizo: divisive hierarchical

Leave-one-out cross-validation

- Misclassification rates in predicting acute vs. subsided schizophrenia
  - k-means: 23.10%
  - agglomerative hierarchical: 24.01%
  - divisive hierarchical: 28.27%