

# Clustering of protein expression data: a benchmark of statistical and neural approaches

I. H. Jarman · T. A. Etchells · D. Bacciu ·  
J. M. Garibaldi · I. O. Ellis · P. J. G. Lisboa

© Springer-Verlag 2010

**Abstract** Clustering issues are fundamental to exploratory analysis of bioinformatics data. This process may follow algorithms that are reproducible but make assumptions about, for instance, the ability to estimate the global structure by successful local agglomeration or alternatively, they use pattern recognition methods that are sensitive to the initial conditions. This paper reviews two clustering methodologies and highlights the differences that result from the changes in data representation, applied to a protein expression data set for breast cancer ( $n = 1,076$ ). The two clustering methodologies are a reproducible approach to model-free clustering and a probabilistic competitive neural network. The results from the two methods are compared with existing studies of the

same data set, and the preferred clustering solutions are profiled for clinical interpretation.

**Keywords** Clustering · Protein expression · Breast cancer · Neural networks

## 1 Introduction

Structure finding in large databases is increasingly important as a first line analysis to generate hypotheses to guide further, targeted studies. This relies on exploratory analysis, typically with the use of clustering methods, typically hierarchical decision trees,  $k$ -means clustering or model-based approaches. While agglomerative methods such as trees are frequently applied, it has been observed that dynamic cutting of each branch may be necessary in order to capture the true global structure of the data (Langfelder et al. 2008). Conversely, the potential of multivariate partition methods operating directly from the large-scale correlation structure of the data, such as  $k$ -means clustering, has been hampered by their sensitivity to initial conditions. In all cases, it is essential to carry out a robustness and stability analysis. The alternative approach is to focus on the consensus between the cluster allocations from an ensemble of classifiers.

This paper benchmarks two systematic frameworks to identify reproducible clusters on the basis of two plausible optimality requirements. One method relies on mapping the landscape of cluster separation against stability, for many initializations of the  $k$ -means clustering, from which to identify optimal and reproducible cluster partitions; this is termed a separation concordance (SeCo) map (Bacciu et al. 2009). The alternative benchmark method directly optimizes a penalised objective function using a competitive neural

---

I. H. Jarman (✉) · T. A. Etchells · P. J. G. Lisboa  
School of Computing and Mathematical Sciences,  
Liverpool John Moores University, Liverpool, UK  
e-mail: i.h.jarman@ljmu.ac.uk

T. A. Etchells  
e-mail: t.a.etchells@ljmu.ac.uk

P. J. G. Lisboa  
e-mail: p.j.lisboa@ljmu.ac.uk

D. Bacciu  
Department of Computer Science, University of Pisa, Pisa, Italy  
e-mail: bacciu@di.unipi.it

J. M. Garibaldi  
School of Computer Science, University of Nottingham,  
Nottingham, UK  
e-mail: jmg@cs.nott.ac.uk

I. O. Ellis  
Department of Histopathology, School of Molecular Medical  
Sciences, University of Nottingham, Nottingham, UK  
e-mail: ian.ellis@nottingham.ac.uk

network with repetition suppression (CoRe) which dynamically clusters the data into a parsimonious representation that covers the data space (Bacciu and Starita 2008).

The results from the two methods are contrasted and related to those obtained with a consensus analysis obtained with a combination of clustering methods (Green et al. 2007; Soria et al. 2010). We will compare the results from these two plausible methods, one of which is taken as an example of a generic partitioning framework and the other of model-based approaches. Significant differences in cluster allocation are highlighted which relate to the particular requirements of the algorithms. The final interpretation is made with reference to the composition of breast cancer subtypes as highlighted in relevant biomedical literature.

## 2 Materials and methods

This benchmark study is applied to a previously studied data set comprising 1,076 observations of 25 protein expression values scaled to a fixed range. They represent measurements from samples collected at initial excision surgery for breast cancer (Abd El-Rehim et al. 2005). The protein names are listed in Table 1.

The first method applied to these data in this paper is a reproducible framework to map the landscape of the solutions of partition clusters by representing each solution, for a given random initialisation, using two dimensions—one to represent the extent of separation between the clusters, and the other to measure concordance between the clusters, for a given number of clusters,  $c$ . The landscape map of separation against concordance is termed the (SeCo) map. This approach to robustness analysis is summarised elsewhere (Bacciu et al. 2009). In principle, it may be applied to any choice of initialisation-dependent clustering method. Here, we utilise one of the most commonly used clustering methods,  $k$ -means.

The measure separation is obtained from a decomposition of the sum-of-squares matrix  $S$

$$S = S_B + S_W, \quad (1)$$

where given the membership of the  $i$ th cluster,  $C_i$ , its  $n_i$  and mean  $m_i = \frac{1}{n_i} \sum_{\forall x \in C_i} x$ , with overall data mean  $m$ , then the within-cluster scatter matrix is given by

$$S_W = \sum_{i=1}^c \left( \sum_{\forall x \in C_i} (x - m)(x - m)^t \right) \quad (2)$$

and the between-cluster scatter matrix by

$$S_B = \sum_{i=1}^c n_i (m_i - m)(m_i - m)^t. \quad (3)$$

It follows that the sum-of-squares matrix  $S$  generates a separation measure

**Table 1** The proteins, with short names and dilutions, used in this study

Antibody [clone]	Short name	Dilution
Luminal phenotype		
CK 7/8 [clone CAM 5.2]	K7/8	1:2
CK 18 [clone DC10]	CK18	1:50
CK 19 [clone BCK 108]	CK19	1:100
Basal phenotype		
CK 5/6 [cloneD5/16134]	CK5/6	1:100
CK 14 [clone LL002]	CK14	1:100
SMA [clone 1A4]	Actin	1:2,000
p63 ab-1 [clone 4A4]	p63	1:200
Hormone receptors		
ER [clone 1D5]	ER	1:80
PgR [clone PgR 636]	gR	1:100
AR [clone F39.4.1]	AR	1:30
EGFR family members		
EGFR [clone EGFR.113]	EGFR	1:10
HER2/c-erbB-2	HER2	1:250
HER3/c-erbB-3 [clone RTJ1]	HER3	1:20
HER4/c-erbB-4 [clone HFR1]	ER4	6:4
Tumour suppressor genes		
p53 [clone DO7]	53	1:50
nBRCA1 Ab-1 [clone MS110]	nBRCA1	1:150
Anti-FHIT [clone ZR44]	FHIT	1:600
Cell adhesion molecules		
Anti E-cad [clone HECD-1]	E-cad	1:10/20
Anti P-cad [clone 56]	P-cad	1:200
Mucins		
NCL-Muc-1 [clone Ma695]	MUC1	1:300
NCL-Muc-1 core [clone Ma552]	UC1co	1:250
NCL muc2 [clone Ccp58]	MUC2	1:250
Apocrine differentiation		
Anti-GCDFP-15	GCDFP	1:30
Neuroendocrine differentiation		
Chromogranin A [clone DAK-A3]	Chromo	1:100
Synaptophysin [clone SY38]	Synapto	1:30

$$J = \text{trace}(S_W^{-1} \cdot S_B), \quad (4)$$

which has the attractive property of being invariant to affine transformations, which include univariate scaling as well as Mahalanobis transformations, or sphering of the data. This measure has been previously studied (Friedman and Rubin 1967), and may be used to optimise the cluster separation in low-dimensional linear projections (Lisboa et al. 2008) in a manner that is analogous to the calculation of principal components but applied to the sum of squares matrix. In this way, the quality of the visualisation uses information about the cluster structure.

The measure of stability is also not of itself new, involving the calculation of pairwise concordance measures between each pair of clustering for the same number of clusters. The cumulative density function of this measure is an indication of how reproducible the clustering solution is, to repeated initialisations (Ben-Hur et al. 2002). A relevant index in this context is a statistical measure of association, as it indicates the extent to which one cluster solution predicts another. A natural index to use, which is independent of the order of cluster labelling in the cross-matching of one cluster allocation against another, is Crámer’s  $V$  index (Bishop et al. 1975). This returns a value in the range  $[0,1]$ , where values near one show a high association and near zero a low association. It is calculated as a normalisation of Pearson’s chi-squared statistic for a contingency table with  $N$  rows and  $M$  columns,

$$\chi^2 = \sum_{i=1}^N \sum_{j=1}^M \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{5}$$

using

$$C_V = \sqrt{\frac{\chi^2}{n \min(N - 1, M - 1)}}. \tag{6}$$

The final coordinate on the SeCo map is the median of the pairwise  $C_V$  values for each cluster solution compared with other solutions for the same number of clusters.

Note that the number of repeated initialisations is, in principle, arbitrary. Sufficient re-initialisations are required to build a map of the most consistent solutions, i.e. to see groups of cluster solutions forming on the SeCo map. For a given number of clusters, a second observer independently re-sampling from the same data set and clustering with the same algorithm will find a similar structure in the SeCo landscape map with a group of clusters having highest separation and stability which reproduces practically the same cluster solution. In this study, we chose to sample 1,000 times, focusing on only the best 10% separated clusters and discarding the rest prior to building the SeCo plot. Even then, several distinct cluster solutions are evident from the landscape map.

Box plots of the concordance index serve to indicate the most stable cluster numbers, similar to Ben-Hur et al. (2002). Note that neither the separation index  $J$  nor the stability index  $C_V$  are alone sufficient to identify the optimal cluster partitions since distinct cluster groups may appear which are vertically or horizontally aligned in the SeCo plot. Therefore, the more complete multi-stage process incorporating both indices is required. Furthermore, for these data, the high-density group of preferred solutions in the landscape map comprises two slightly different

cluster partitions, which will be described later. This paper demonstrates the benefits of this method in obtaining reproducible and high quality solutions in respect of both clustering performance indices.

Moreover, cluster solutions obtained independently are very consistent across different cluster numbers, forming a hierarchy of partitions with increasing levels of detail. It is this hierarchy that is the focus for the interpretation of the data sub-types.

The second method, known as competitive repetition suppression (CoRe) learning, is a bio-inspired approach defined within a robust Bayesian framework (Bacciu et al. 2007; Bacciu and Starita 2008) that estimates the number of clusters by selectively pruning an initial set of neurons  $U = \{u_1, \dots, u_N\}$  to achieve a parsimonious coding of the stimuli space, i.e. the input dataset, such that each neuron retained at the end of the learning phase identifies a single cluster within the data.

Each CoRe unit  $u_i$  is associated to a prototype vector  $c_i \in \mathbb{R}$ , determining its preferred stimulus, and to an activation function  $\varphi_i(x_k)$  that determines the unit’s response to the sample  $x_k \in \mathbb{R}$ . In particular, we compute a feature activation  $\varphi_{il}(x_{kl})$  for each of the  $l$  components of the  $d$  dimensional input vector  $x_k$ , as well as an overall unit response  $\Phi_i(x_k)$  that is a function of the feature activation and acts as a pooling mechanism determining the amount of interaction between the features. In the remainder of the paper, we assume  $\varphi_{il}$  to be a univariate Gaussian centred on  $c_{il}$  and parametrical with respect to the spread  $\sigma_{il}$ , while  $\Phi_i$  is chosen as the corresponding multivariate Gaussian. This step also requires a univariate standardisation of the data, which will influence the cluster solutions.

Given an input sample  $x_k$ , we compute the response  $\Phi_i$  of each CoRe unit, and the most active neurons are selected to form the *winners pool*

$$\text{win}_k = \{i | \Phi_i(x_k) \geq \theta_{\text{win}}, u_i \in U\} \cup \left\{ i \mid i = \arg \max_{u_i \in U} \Phi_j(x_k) \right\}, \tag{7}$$

while the remainder of the neurons is inserted into the *losers pool*, i.e.  $\text{lose}_k = U \setminus \{\text{win}_k\}$ . The meta-parameter  $\theta_{\text{win}}$  determines the minimum activation level for winner units and, consequently, regulates the target selectivity of the neurons. Units within the pool of winners are allowed positive learning and their response to patterns similar to  $x_k$  is strengthened. Conversely, the response of loser units is reduced by selectively suppressing the feature activation. Such an adaptive strengthening/suppression dynamics is realized by optimizing the instantaneous local error

$$E_{il,k}^t = \delta_{ik}(1 - \varphi_{il}(x_{kl})) + (1 - \delta_{ik})\frac{1}{2}(\varphi_{il}(x_{kl})RS_{kl}^t)^2, \quad (8)$$

with respect to the model parameters  $c_{il}$  and  $\sigma_{il}$ . The term  $\delta_{ik}$  denotes the indicator function for  $\text{win}_k$ , while  $RS_{kl}^t$  determines the amount of repetition suppression (RS) generated in response to the  $l$ th feature of pattern  $x_k$  at time  $t$ , i.e.

$$RS_{kl}^t = \frac{1}{|\text{win}_k|} \sum_{i \in \text{win}_k} \varphi_{il}(x_{kl}) \frac{1}{|\chi_t|} \sum_{x_p \in \chi_t} \min\left(\frac{\varphi_{il}(x_{pl})}{\varphi_{z(U,x_p)l}(x_{pl})}, 1\right), \quad (9)$$

where  $\chi_t$  is the set of patterns  $x_k$  presented to the network up to time  $t$  and  $z(U, x_p)$  returns the index of the most active unit for sample  $x_p$ . Consistent with the biological RS mechanism (Bacciu and Starita 2008), Eq. 9 generates an adaptive feature suppression that is proportional to an estimate of the frequency of the stimuli similar to the current input pattern. Given that such a penalization is applied selectively to the single features, it can be used to suppress the contribution of irrelevant and noisy covariates of the input vector (see Bacciu et al. 2007 for further details). The CoRe model also defines a measure of unit and feature relevance

$$\hat{v}_{il}^t = \frac{1}{\sum_{x_p \in \chi_t} \min\left(\frac{\varphi_{il}(x_{pl})}{\varphi_{z(U,x_p)l}(x_{pl})}, 1\right)} \sum_{x_k \in \text{win}_i^t} \left\{ \frac{\varphi_{il}(x_{kl})}{\varphi_{z(\text{win}_k, x_k)l}(x_{kl})} \right\}_0, \quad (10)$$

where  $\text{win}_i^t$  is the set of patterns  $x_k$  in  $\chi_t$  for which unit  $u_i$  was in the winners pool, while the function  $\{\cdot\}_0$  flattens its

argument to zero if  $v > 1$  and returns  $v$  otherwise. Such relevance measure is used to decide whether a neuron can be pruned from the network. In particular, if, at a certain time  $t$ , the  $\hat{v}_{il}^t$  value falls below a given threshold  $\theta_{pr}$  for all the features  $l$ , then the unit  $u_i$  can be considered irrelevant for the representation of the actual cluster structure, and is thus deleted from the network. Further, the univariate components of the relevance factor identify which prototype dimensions  $c_{il}$  best characterize the input patterns assigned to the  $i$ th unit/cluster, hence providing a significance ranking for the covariates of each discovered cluster profile.

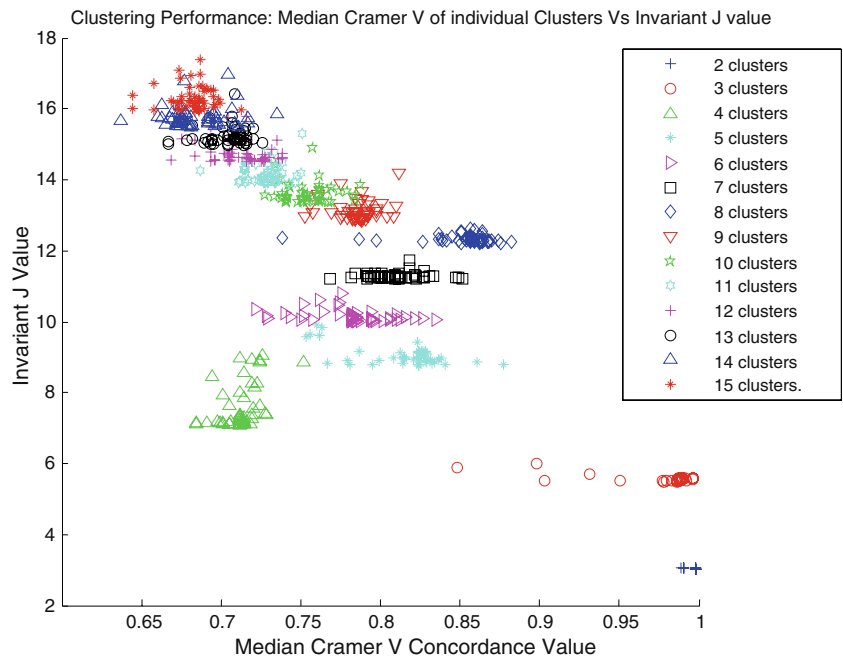
Summarizing, the CoRe model is initialized with a large neural population whose parameters are incrementally updated by minimizing the error in Eq. 8. CoRe iteratively prunes irrelevant units as soon as their relevance value falls under  $\theta_{pr}$ , finally converging to an estimate of the number of clusters that is equal to the number of units retained at the end of the learning phase.

### 3 Results

The first set of results to be described is for landscape mapping, shown in Fig. 1. Cluster assignments for two and three clusters have almost perfect stability. This is expected as the gross structure of the dataset, as represented by its principal components, will serve to ‘anchor’ the  $k$ -means algorithm without necessarily resolving any fine structure.

A consistently low Cramer’s  $V$  of around 0.7, indicating marked instability, is observed among four cluster solutions. Two distinct regions are apparent when five clusters

**Fig. 1** Separation concordance map of  $k$ -means cluster solutions for multiple random initialisations with a given cluster number, shown in the legend. There is a marked increase in stability of the partitions obtained when  $N = 8$ . Nevertheless, there are two overlapping cluster solutions with  $C_V, J$  of approximately 0.86 and 12.1, respectively

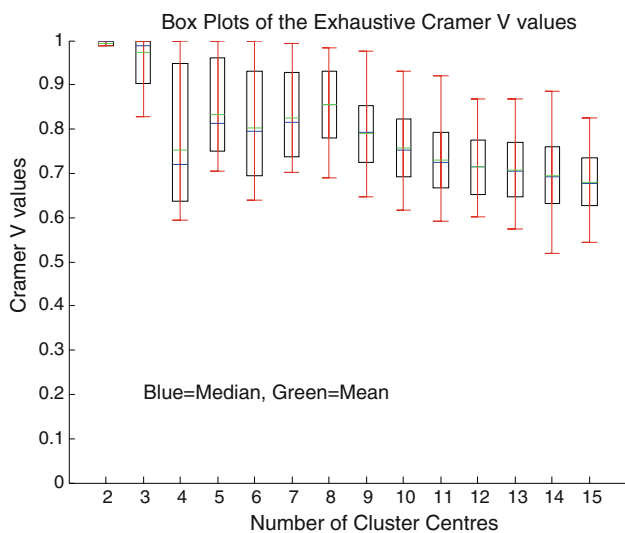


are assumed, the least stable of which has the slightly better separation. A spread of solutions was found for six and seven clusters, with a convergence of stability for eight clusters. This slight increase in stability is also reflected in Fig. 2.

These two observations set our preferred partition for the  $k$ -means algorithm, which is obtained by sampling from the high-density group of eight cluster solutions, where the histogram of the values taken by the index  $C_V$  turns out to have a mode at 0.8565.

The two solutions are similar in six of the eight clusters, the remaining two being partitioned in different directions. This meta-stability is evidenced in two parallel hierarchies of data partitions, shown in Fig. 3. Up to three clusters, the hierarchies agree, but they branch out separately from four clusters onwards, each hierarchy remaining very consistent throughout. This is the case despite the solutions for each assumed cluster number,  $N$ , being obtained by independent re-initialisations, which generates the landscape map in Fig. 1. The reproducible solutions are generated using the following process:

1. Identify high-density groups of points in the SeCo plot, for a given  $N$ .
2. Find the mode of the histogram of concordance index values in the proximity of this high-density region. Calculate the concordance values for each pair of cluster partitions which share the modal value of Cramer  $V$ . In the case of Fig. 1, there are nine solutions with  $C_V = 0.8565$ . These are the preferred cluster solutions.



**Fig. 2** Box plots of the distribution of Cramer  $V$  values in Fig. 1, confirming that the stability is enhanced when eight clusters are assumed

3. Calculate the pairwise values of  $C_V$  for the preferred solutions and cross-tabulate the pair with the least value of this index. In the case of Fig. 1, there are two distinct cluster partitions that share similar locations on the SeCo map, with the remaining seven preferred solutions similar to one of the others.

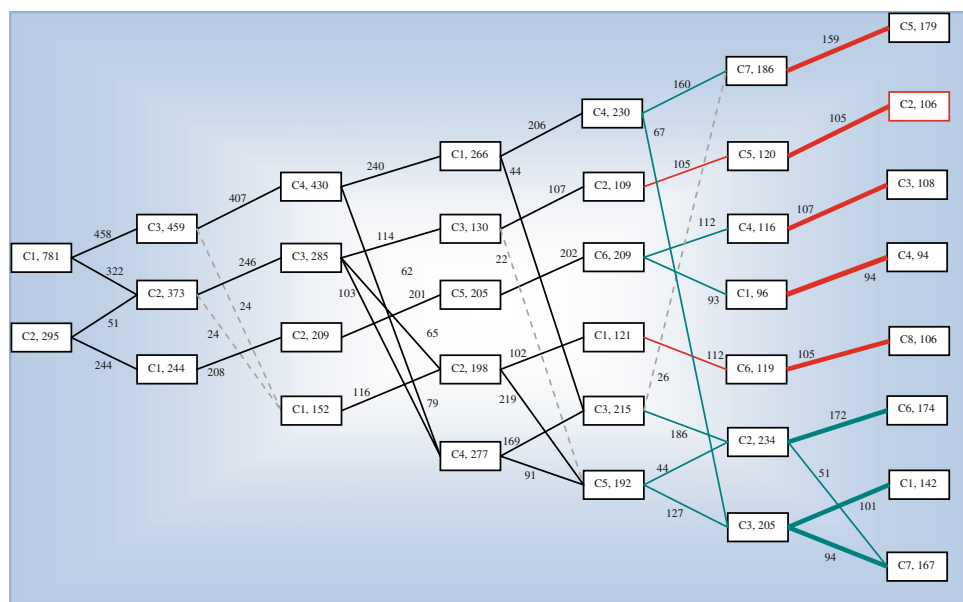
It is interesting to note that the most stable partition in Fig. 1, with  $C_V$  close to 0.9, is neither of these two solutions but a mixture between the two. Similarly, the eight cluster solution with the highest value of  $J$  may be either of the two partitions with similar separation.

We argue that it is the identification of a set of densely packed cluster solutions for a given  $N$  which enables the identification of stable or meta-stable data partitions to interpret. The principle of the landscape mapping method is that a second observer repeating the same process will find the same solution set for a given  $N$ , whether this is a unique partition. For these data, this finding was verified by repeated application of the landscape mapping approach. These clusters are interpreted in the discussion section.

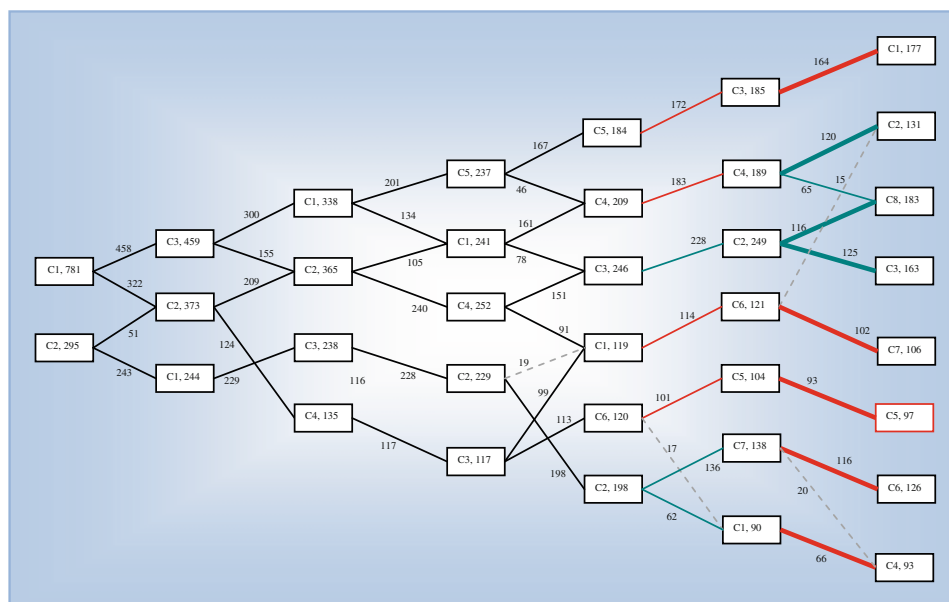
A second set of cluster solutions comes from the competitive neural network with repetition suppression. The results have been obtained starting from three different network sizes, i.e. 20, 50 and 100 units, to assess the effect of initial conditions. The expression levels have been normalized to [0,1] by max–min normalization of the dataset. Cluster number estimation was repeated 50 times for each network size, with varying random initial prototype positions. The choice of meta-parameters and other implementation details follow the standard settings presented in Bacciu and Starita (2008) and Bacciu et al. (2007). The distribution of the estimate of the number of clusters in Fig. 4 shows that all the three networks have a preference for five clusters. In general, a larger initial network size yields to more stable results that are less likely to be dependent on the initial prototype allocation, especially when dealing with medium-dimensional data such as with the Dalia dataset. For instance, with 20 initial units the algorithm tends to identify more 3-cluster solutions that reflect the gross covariance structure, while with 100 initial units there is a stronger concordance on the 5-cluster hypothesis across the independent runs of the algorithm.

The differences between the two clustering methods, shown in Table 2, may result from the effect of the feature-wise suppression in Eq. 9 that modifies the scale at which the single covariates are processed. For instance, covariates that are considered irrelevant tend to be over-smoothed: this on the one hand, allows to reduce the effect of noisy data components; on the other hand, if a significant feature

**Fig. 3** Hierarchical structures of data partitions obtained by mapping the landscape for repeated initialisations of the *k*-means algorithm, with increasing number of clusters. The entries in each *box* are the cases allocated to that cluster. The *lines* list the number of cases switching clusters, where there are more than 10% of box membership and *dashed lines* when fewer than 5%. The final eight partitions are labelled Clusters A and B, respectively



(a)



(b)

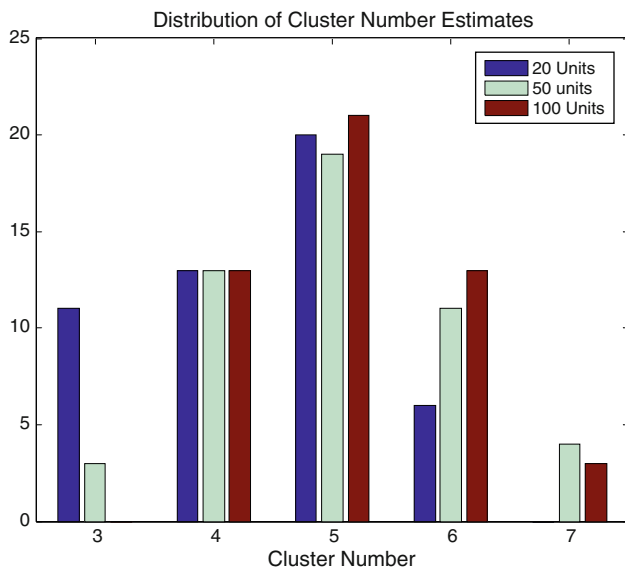
is suppressed, this might influence the correct estimation of the latent structure of the data.

#### 4 Discussion

The method of landscape mapping identified two similar performing clustering solutions for the protein expression data set, which are cross-tabulated in Table 3. Cluster A corresponds very well with class assignments obtained by

a consensus of clustering methods (Green et al. 2007; Soria et al. 2010) as shown in Table 4. The main difference is to pull-out two splinter groups from clusters 1 (Luminal A) and 2 (Luminal N) to form two smaller, but separate clusters that are PgR negative, hence attach to cluster 3 (Luminal B). The box plots of the off-diagonal entries in Table 4 are consistent with their new allocations.

The second difference relates to the allocation of the basal-like tumours. While in Clusters A they are separated



**Fig. 4** Distribution of the most likely cluster number estimated by CoRe over 50 iterations of the algorithm initialized with 20, 50 and 100 units

**Table 2** Contingency table for cross-matching the solutions from the two clustering methods

Count	CoRe5					Total
	1	2	3	5	4	
Cluster A (Crosstab)						
3	91	0	0	0	17	108
4	69	0	0	0	25	94
8	40	29	14	1	22	106
6	0	143	3	24	4	174
5	0	9	145	25	0	179
1	5	0	56	9	72	142
7	0	16	3	125	23	167
2	8	6	12	31	49	106
Total	213	203	233	215	212	1,076
Cluster B (Crosstab)						
6	123	0	0	0	3	126
2	0	127	3	1	0	131
1	3	65	0	49	60	177
8	0	16	122	40	5	183
3	0	1	49	93	20	163
4	35	0	0	0	58	93
5	7	12	4	31	43	97
7	45	12	25	1	23	106
Total	213	233	203	215	212	1,076

Compared with the 5-cluster solution of CoRe, the SeCo clusters A and B score Cramer  $V$  indices of 0.668 and 0.644, respectively

by p53, in Clusters B they split by of mucl1 (and muclco) although with expression levels inversely correlated with those of p53.

**Table 3** Contingency table for cross-matching the two solutions from the SeCo method with  $N = 8$ , which have a concordance index of Cramer  $V = 0.774$

Cluster A	Cluster B								Total
	1	2	3	8	7	5	6	4	
1	118	4	0	1	6	1	0	12	142
5	21	125	0	33	0	0	0	0	179
7	37	0	122	4	0	2	0	2	167
6	0	0	29	145	0	0	0	0	174
8	0	2	6	0	98	0	0	0	106
2	0	0	6	0	0	94	1	5	106
3	1	0	0	0	1	0	64	42	108
4	0	0	0	0	1	0	61	32	94
Total	177	131	163	183	106	97	126	93	1,076

**Table 4** The solution pair found by  $k$ -means using the SeCo method has concordance values of 0.932 and 0.814 compared with the allocated cases in Green et al. (2007)

Count	Cluster in Green et al. (2007)						Total
	2	6	5	4	1	3	
Cluster A (cross-tabulation)							
5	107	0	0	0	6	0	113
1	45	4	9	0	4	2	64
2	0	65	0	0	0	0	65
3	0	0	58	2	0	1	61
4	0	0	2	80	0	0	82
6	0	0	0	0	138	0	138
7	1	8	0	0	54	2	65
8	0	0	0	0	0	75	75
Total	153	77	69	82	202	80	663
Cluster B (cross-tabulation)							
3	67	0	7	0	0	2	76
8	124	7	0	0	0	0	131
2	1	100	0	0	0	1	102
5	0	0	60	0	0	0	60
6	0	0	0	57	43	1	101
4	0	0	6	24	21	0	51
7	0	0	0	1	0	75	76
1	10	46	4	0	5	1	66
Total	202	153	77	82	69	80	663

This interpretation is confirmed by the cluster profiles, shown side by side in Fig. 5 with sub-types mapped in Fig. 6. Both solutions clearly identify the well-known HER2-positive group.

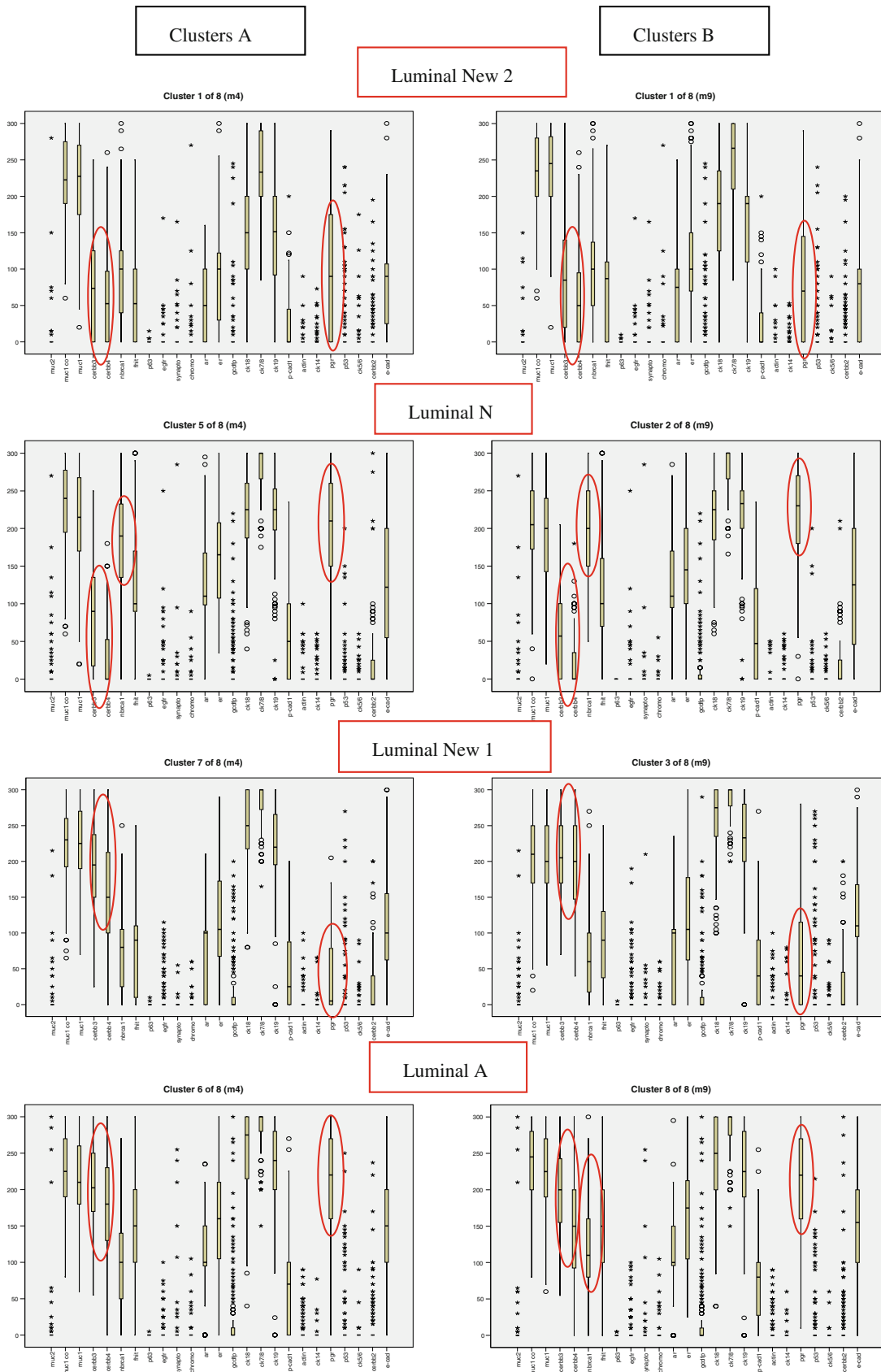


Fig. 5 Box plots of the two solutions identified by the SeCo method for the protein expression data

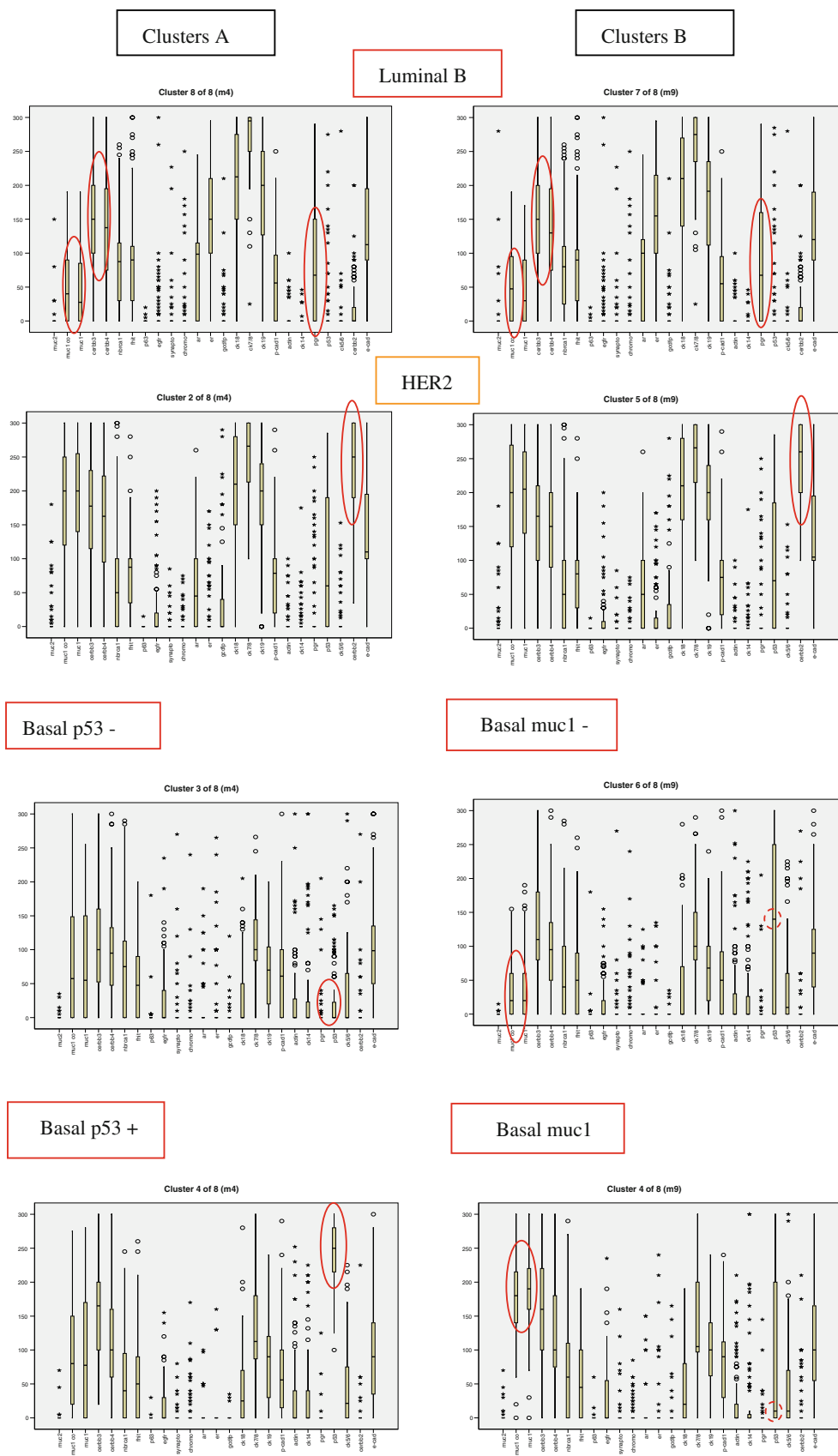
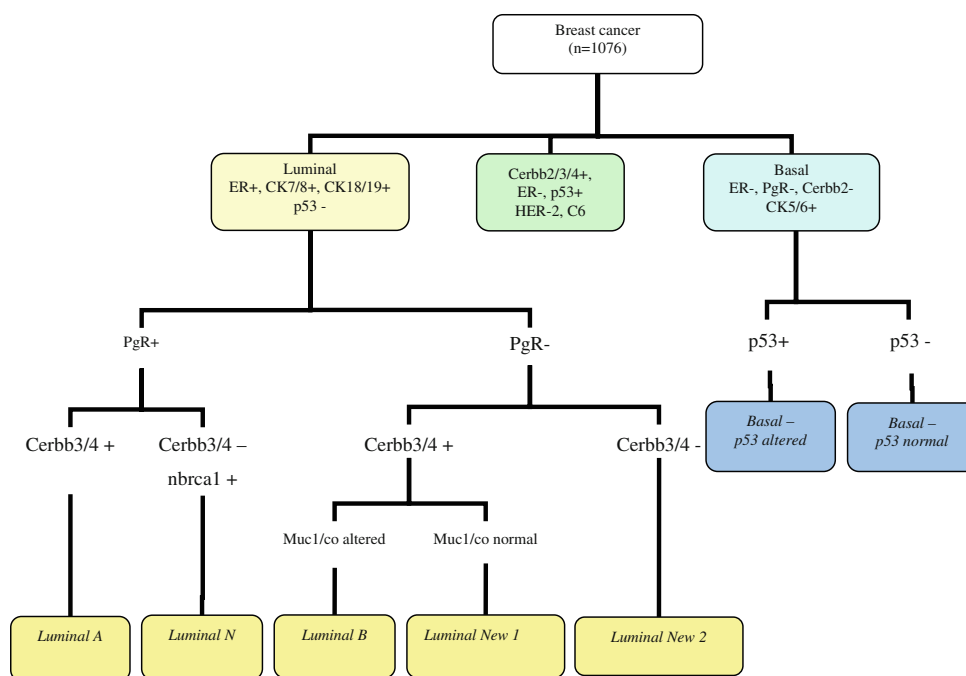


Fig. 5 continued

**Fig. 6** Summary of the interpretation Clusters B found by the SeCo method. It is clear from Fig. 4 that the two cluster groups are consistent except in the case of the basal groups, where Cluster A has better specificity for p53 but lacks specificity for muc1 and muc1co, in contrast with Cluster B

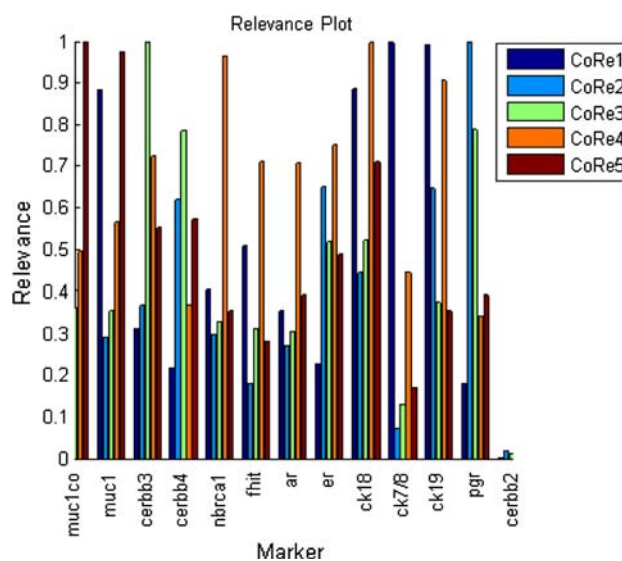


**Table 5** Contingency table for cross-matching the solutions from the CoRe method and a previously obtained cluster partition

Clusters in Green et al. (2007)	CoRe 5-cluster solution				
	2	3	1	4	5
C1	<b>129</b>	4	0	3	<b>66</b>
C2	1	<b>138</b>	0	7	7
C3	14	11	<b>37</b>	16	2
C4	0	0	<b>65</b>	17	0
C5	0	0	<b>56</b>	13	0
C6	1	8	1	<b>37</b>	<b>30</b>
NC	58	72	54	119	110

The bold values indicate a correspondence between the clusters obtained with the CoRe method and those from a previously published study by Green et al.

The correspondence between a representative cluster solution featuring five clusters obtained by the CoRe method and the class assignment solution determined previously is shown in Table 5. It can be seen that CoRe5 clusters 2 and 3 correspond reasonably well to the previously identified Luminal-A (C1) and Luminal-N (C2), although both contain a set of previously unclassified examples. However, in contrast to the SeCo8 cluster assignment, the interpretation of the remaining three clusters (1, 4 and 5) is more problematic. Clusters 1 and 4 seem to be a mix of previously identified basal tumours, with little clear clinical meaning. At the same time, the accepted HER2-positive clinical group is not satisfactorily recovered by this methodology. Such an anomalous clustering seems



**Fig. 7** Plot of CoRe relevance factor  $\hat{v}_{ij}^t$  for the 12 significant covariates plus the HER2-specific marker c-erbB-2

to be due by the inability of the CoRe algorithm to detect the HER2-specific covariate cerb-b-2 as a relevant feature for cluster membership. Figure 7 shows the relevance factor  $\hat{v}_{ij}^t$  for the 12 significant covariates identified by CoRe. The algorithm can detect consolidated prognostic markers while, on the other hand, it neglects the role of cerb-b-2 which prevents the HER2 group from being correctly identified.

In summary, there are two SeCo8 cluster solutions, one of which is highly consistent with previous approaches and

clinical knowledge, while identifying two potentially interesting clinical sub-groups. In contrast, the CoRe5 solution is not similarly compatible with current clinical knowledge and known breast cancer groupings. Such a difference seems to be due to the alteration of the metric space introduced by the suppression of the HER2-specific covariate *cerb-b-2*.

## 5 Conclusions

This study shows that the details of clustering algorithms do matter and reliable exploratory analysis requires a robust methodology. In particular the data allocation must be reproducible, whichever method is used. The results of applying two systematic approaches to a substantial protein expression data set not only showed broad agreement between specific clusters but also significant differences including lack of definition of known subtypes.

Regarding the appropriate number of clusters, the results showed that how a consistent hierarchy of cluster solutions can be defined which, together with the calculation of stability measures, results in a small range of preferred partitions to be interpreted. The final cluster interpretation is consistent with earlier results obtained by a consensus of three clustering methods (Green et al. 2007; Soria et al. 2010).

Future work will involve a further interpretation of the clusters using linear visualisation methods (Lisboa et al. 2008) and rule extraction (Etchells and Lisboa 2006), as well as non-linear visualisation with attention to outliers (Vellido et al. 2006).

**Acknowledgments** The authors acknowledge the support of Dr. A. Green and other members of the Breast Cancer Research Team of the University of Nottingham, UK, and Dr. G. Ball, of Nottingham Trent University, UK, towards the collection of high-throughput protein expression dataset used in this study, and financial support from the European Network of Excellence Biopattern (FP6-2002-IST-1 No. 508803).

## References

Abd El-Rehim D, Ball G, Pinder S, Rakha E, Paish C, Robertson J, Macmillan D, Blamey R, Ellis IO (2005) High-throughput

protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. *Int J Cancer* 116:340–350

- Bacciu D, Starita A (2008) Competitive repetition suppression (CoRe) clustering: a biologically inspired learning model with application to robust clustering. *IEEE Trans Neural Netw* 19(11):1922–1941
- Bacciu D, Micheli A, Starita A (2007) Simultaneous clustering and feature ranking by competitive repetition suppression learning with application to gene data. In: *International conference on computational intelligence in medicine and healthcare (CIMED'07)*
- Bacciu D, Jarman IH, Etchells TA, Lisboa PJG (2009) Patient stratification with competing risks by multivariate fisher distance. In: *International joint conference on neural networks (IJCNN)*, Atlanta, USA
- Ben-Hur A, Elisseeff A, Guyon I (2002) A stability based method for discovering structure in clustered data. In: Altman RB, Lauderdale K (eds) *Pacific symposium on biocomputing 2002*, Kauai, Hawaii, USA, vol 7, pp 6–17
- Bishop YMM, Fienberg SE, Holland DW (1975) *Discrete multivariate analysis: theory and practice*. MIT Press, Cambridge, MA
- Etchells TA, Lisboa PJG (2006) Orthogonal search-based rule extraction (OSRE) from trained neural networks: a practical and efficient approach. *IEEE Trans Neural Netw* 17(2):374–384
- Friedman HP, Rubin J (1967) On some invariant criteria for grouping data. *J Am Stat Assoc* 62(320):1159–1178
- Green AR, Garibaldi JM, Soria D, Ambrogi F, Ball G, Lisboa PJG, Etchells TA, Boracchi P, Biganzoli E, Macmillan RD, Blamey RW, Ellis IO (2007) Identification of sub-classes of breast cancer through consensus derived from automated clustering methods. *Eur J Cancer Suppl* 5(3):18
- Langfelder P, Zhang B, Horvath S (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24(5):719–720
- Lisboa PJG, Ellis IO, Green AR, Ambrogi F, Dias MB (2008) Cluster-based visualisation with scatter matrices. *Pattern Recogn Lett* 29(13):1814–1823
- Soria D, Garibaldi JM, Ambrogi F, Green AR, Powe DG, Rakha EA, Macmillan RD, Blamey RW, Ball G, Lisboa PJG, Etchells TA, Boracchi P, Biganzoli E, Ellis IO (2010) A methodology to identify consensus classes from clustering algorithms applied to immunohistochemical data from breast cancer patients. *Comput Biol Med* 40:318–330
- Vellido A, Lisboa PJG, Vicente D (2006) Robust analysis of MRS brain tumour data using t-GTM. *Neurocomputing* 69(7–9):754–768