The University of
Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Soria, Daniele and Garibaldi, Jonathan M. and Ambrogi, Federico and Boracchi, Patrizia and Raimondi, E. and Biganzoli, Elia M. (2009) Cancer profiles by Affinity Propagation. International Journal of Knowledge Engineering and Soft Data Paradigms, 1 (3). pp. 195-215. ISSN 1755-3210

Access from the University of Nottingham repository:
http://eprints.nottingham.ac.uk/28138/1/Soria_Ambrogi_AP_extended.pdf

Cancer profiles by Affinity Propagation

D. Soria[1], **F. Ambrogi**[2], **E. Raimondi**[3], **P. Boracchi**[2], **J.M. Garibaldi**[1], **E. Biganzoli**[23]

[1]*School of Computer Science, University of Nottingham, UK*
[2]*Istituto di Statistica Medica e Biometria "G.A. Maccacaro", Università degli Studi di Milano, Milano, Italy*
[3]*Struttura di Statistica e Biometria, Fondazione IRCCS Istituto Nazionale Tumori, Milano, Italy*

Abstract

The Affinity Propagation algorithm is applied to various problems of breast and cutaneous tumours subtyping using traditional biologic markers. The algorithm provides a procedure to determine the number of profiles to be considered.

Well know breast cancer case series and cutaneous melanoma were used to compare the results of the Affinity Propagation with the results obtained with standard algorithms and indexes for the optimal choice of the number of clusters.

Results from Affinity Propagation are consistent with the results already obtained having the advantage of providing an indication about the number of clusters.

1 Introduction

Genomic analysis renewed interest in clustering techniques. After the seminal paper of Eisen and colleagues [6], proposing hierarchical clustering and the visual inspection of the dendrogram to discover unknown pattern of gene associations, the use of clustering has become more and more popular especially for discovering profiles in cancer with respect to high-throughput genomic data. Important applications of the Eisen method are the work of Bittner [3] on clustering of cutaneous melanoma and the works of van't Veer [25] and Perou [19] on breast cancer.

Recently a classification of breast carcinoma using traditional tumor markers was proposed [1]. The classification was in agreement with the classifications obtained with c-DNA microarray data [19, 25]. Different clustering algorithms were used to choose a stable solution across different clustering methods. At last a classification in four clusters was preferred and suggested a possible separation of high risk profiles. One of the main problems connected with cluster analysis is the choice of the number of clusters. In classical cluster analysis it is customary to use indexes to compare one cluster solutions to other cluster solutions and to choose the one suggested as optimal.

In the previous application [1], different indexes were used to select an optimal partition. Namely the indexes proposed by Calinski and Harabasz [4], Krzanowski and Lai [14], Hartigan [12] and Tibshirani et al. [23], were considered. It is worth noting that the visual inspection of the dendrogram suggested by Eisen [6] is an informal method to determine the number of clusters. Such a procedure was criticized in [10] as it can cause difficulty in assessing the validity of the grouping.

According to Getz [9], the number of clusters should be determined internally by the clustering algorithm and should not be externally prescribed.

In this work a new clustering algorithm, the Affinity Propagation [8], will be adopted to cluster cancer patients in order to evaluate its performance with respect to the traditional applications. Several datasets, already published in literature, were considered in our study: the melanoma data of Bittner et al. [3], and four different breast cancer data analyses: namely, the studies of Ambrogi et al. [1], Perou et al. [19], van't Veer et al. [25], and Sotiriou et al. [22]. Although this algorithm does not determine automatically the number of clusters it provides a consistent method to suggest the number of clusters to be created which can be useful to detect different levels of association pattern.

2 Material and methods

2.1 Case data

This section includes a description of the case series adopted in the work.

The information on 633 patients operated on for primary infiltrating breast cancer between 1983 and 1992, archived at the Pathology department of the University of Ferrara, was retrospectively analyzed in the work of Ambrogi et al. [1].

The available data concerned patient age, pathological tumour size, histologic type, pathologic stage, and number of metastatic axillary lymph nodes; as well as immunohistological determinations of oestrogen receptor status (ER),

progesterone receptors status (PR), Ki-67/MIB-1 proliferation index (Ki-67), *c-Erb*B-2/NEU (NEU) and the p53 oncosuppressor gene (p53).

Values of ER, PR and NEU tended to be grouped on the following values: 0%, 10%, 25%, 50%, 75% and 100%; they were consequently discretized on those values. Values of Ki-67 and p53 were used as originally measured.

A second dataset was also analyzed: the melanoma data of Bittner et al. [3]. These data consist of gene expression profiles obtained on a collection of 38 samples, comprised of 31 melanoma tumors and 7 controls. For the analysis described in Section 3, the data from the seven control specimens were excluded and only the ratios for the 3613 genes that were considered "well measured" (that is their intensities were sufficiently high) were used. These ratios were converted to log2 ratios.

Another dataset on breast cancer that was considered for investigation was the one of Perou et al. [19]. Variation in gene expression patterns in a set of 65 surgical specimens of human breast tumours from 42 different individuals have been characterized, using complementary DNA representing 8102 human genes. According to the authors, these patterns provided a distinctive molecular portrait of each tumour [19]. Sets of co-expressed genes were identified for which variation in messenger RNA levels could be related to specific features of physiological variation. The tumours could be classified into subtypes distinguished by differences in their gene expression patterns. In their paper, Perou et al. focused on a set of 1,753 genes in 65 experimental tissue samples. In each sample, the ratio of the abundance of transcripts of each gene to its median abundance across all tissue samples, is reported.

Data and original analyses of Bittner and Perou are fully described in the book "Design and Analysis of DNA Microarray Investigations" by Simon and colleagues [21].

The dataset analysed in van't Veer et al. [25] was also considered in this study: they used DNA microarray analysis on primary breast tumours of 117 young patients, and applied supervised classification to identify a gene expression signature strongly predictive of a short interval to distant metastases in patients without tumour cells in local lymph nodes at diagnosis. All patients were lymph node negative and under 55 years of age at diagnosis. Unsupervised clustering detected two subgroups of breast cancer, which differ in ER status and lymphocytic infiltration.

The last dataset we analysed was taken from Sotiriou et al. [22] where tumour samples from 99 patients with primary local breast cancer were accessed from the John Radcliffe Hospital from January 1993 to December 1994. All of the tumor samples were invasive ductal carcinomas; 46 individuals were node negative and 53 were node positive. Almost all of the patients received adjuvant treatment after surgery, consisting of radiotherapy (80 patients),

chemotherapy (32 patients), and endocrine therapy (78 patients) according to accepted practice guidelines at that time. From each woman belonging to the cohort, a tumour tissue sample, on which all the analyses were run, was extracted using microarray technology. For each spot, Log (base 2) ratios between red signal and green signal were calculated. The log ratios were then normalized within each array by subtracting from each the median log ratio value across the spots on the array.

2.2 Statistical Methods

The clustering technique Affinity Propagation (AP, [8]) will be adopted for grouping tumours with similar biological characteristics.

As other clustering algorithms, this method uses data to find a set of centers such that the sum of squared errors between data points and their nearest center is small.

Like other traditional clustering techniques, the Affinity Propagation algorithm determines the centers from real data points (exemplars). These exemplars correspond, for example, to the medoids in the algorithm Pam [13] (Partitioning Around Medoids, a more robust version of K-means), that is k representative objects among the observations of the dataset that should represent the structure of the data.

As a technical detail, it is worth noting that K-means algorithm does not use exemplars, as the centers are not generally actual data points.

Affinity Propagation combines the properties of different classes of clustering algorithms. On one hand, algorithms like hierarchical clustering are based on grouping pairs of objects with high Affinity. On the other hand model-based clustering uses a probability model based on a mixture of class conditional distributions. Affinity Propagation uses both pairs comparison and a probability model to determine the optimal grouping. According to a more technical point of view, Affinity Propagation can be derived as the sum-product algorithm in a graphical model describing the mixture model [7].

The first step for the algorithm implementation is to choose a measure of similarity, s(i,k), between all pairs of data points. In AP terminology, s(i,k) quantifies how well the data point with index k is suited to be the exemplar for data point i. Generally, as similarity, it is used the negative Euclidean distance. In the case of c-DNA data the Pearson correlation is generally used as similarity measure [3].

This method does not require the number of clusters to be prespecified.

The second step is about the choice of the values of "preferences" which will be indicated, with a little abuse of notation as s(i,i). Please note that this is not a similarity measure. The preferences represent a measure of how much data point i is candidate to be an exemplar. In general, data

RESPONSIBILITY r(i,k)

AVAILABILITY a(i,k)
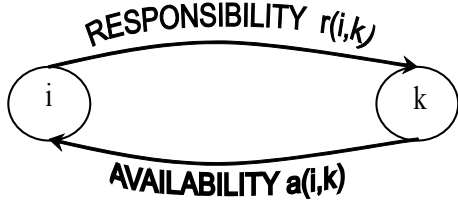
points with larger values of s(i,i) are more likely to be chosen as exemplars. At the beginning, the AP simultaneously considers all data points as potential exemplars (Input the preferences common for all data points).

The number of identified exemplars is influenced by the values of the input preferences, but also emerges as a result of the message passing structure that is illustrated subsequently. For very small value of input s(i,i), for every i, all data points are grouped in one large cluster with a single exemplar; in the opposite case of large s(i,i) for every i, each data point prefers to be its own exemplar. In general, the initial value of the preferences is set equal to the median of all input similarities (resulting in a moderate number of clusters) or to their minimum (resulting in a small number of clusters).

The AP is a method that recursively transmits messages (that will be defined subsequently) between pairs of data points until a good set of exemplars and corresponding clusters emerges.

The algorithm is named Affinity Propagation because at any point in time, each message reflects the current Affinity between one data point and the other that is its exemplars.

In practice, it is adopted a message-passing algorithm in which each data point i furnishes a measure to suggest another data point k to be selected as cluster center, taking into account other potential exemplars for point i.

There are two kinds of message being passed between each pairs of data points that represent the relationship between data points:

- "responsibility": sent from data point i to candidate exemplar k. It is a measure that quantifies how well-suited point k is to be the exemplar for point i, taking into account other potential exemplars for point i. This message is represented by r(i,k) and it is computed using this formula:

$$r(i,k) = s(i,k) - \max_{k' \mathbf{s.t.} k' \neq k} \{a(i,k') + s(i,k')\}$$

where s.t. means "so that";

- "availability": sent from candidate exemplar point k to point i. It is a measure that reflects the evidence for point i to choose point k as its exemplar, considered that other points may have k as an exemplar. This message is represented by a(i,k) and it is computed using

this formula:

$$a(i,k) = \min\{0, r(k,k) + \sum_{i' \mathbf{s.t.} i' \notin \{i,k\}} \max\{0, r(i',k)\}\}$$

A particular measure is the "self-responsibility", that is r(k,k); it reflects accumulated evidence that point k is an exemplar and how it would be unsuitable to be integrated in a group of another cluster center.

At the beginning of the algorithm, the availabilities are initialized to zero, so r(i,k) is set to the input similarity between point i and its potential exemplar k minus the largest of the similarities between point i and other candidate exemplars. After the computation of all the responsibilities, the availabilities are worked out using the previous formula. Only the positive portions of responsibilities between the candidate exemplar k and other data points i' are added because it is only necessary for a good exemplar to explain some data points well ($r(i',k) > 0$) regardless of how poorly it explains other data points ($r(i',k) < 0$). In fact, if $r(i',k) < 0$, k is not suited to be the exemplar for point i'. So in this case, the point i' will not contribute to the message passing from candidate exemplar k to point i.

After that, the messages are recursively updated for a fixed number of iterations or until a stable clustering result. At any stage, the availabilities and responsibilities can be combined to identify exemplars. For point i, the value k that maximizes $a(i,k) + r(i,k)$ identifies point i as exemplar if k=i or identifies the data point that is the exemplars for point i. In other words, as suggested in [5], after exchanging messages, Affinity Propagation identifies a set of exemplars K so as to maximize the net similarity, which is defined by the authors as

$$\sum_{i \notin K} \max_{k \in K} s(i,k) + \sum_{k \in K} p(k)$$

where $p(k)$ is the *a priori* preference that point k be chosen as an exemplar.

At the end of the message passing, we obtain the number of clusters and the labels for each data point of its exemplars.

All the equations shown above are derived and explained in details in the Supporting Online Material for [8].

An advanced characteristic of Affinity Propagation is that it determines the number of clusters on the basis of the message passing architecture and the points that are most representative, given an initial common preference. It is possible to see the effect of the value of the input preference on the number of clusters by a graphic with the value of the common initial preference on the x-axis and the respective number of clusters on the y-axis. In this way, the value to adopt in the analysis can be established in correspondence with plateaus that are observable in this graphic.

Given the initial common preference AP defines a unique solution. One of the strong points of AP is its computational efficiency, as described in [17]. The algorithm is feasible even in presence of very large data sets.

In all our applications of AP, $1 - Pearson\ correlation$ was used as similarity. The only exception was in the first case series where the negative Euclidean distance was used. These choice were in accordance with the ones adopted by the authors of the works considered.

Multiple correspondence analysis (MCA), see Greenacre [11] or Lebart et al. [16], was used as visualization technique to study the composition of the clusters for the breast cancer data due to the discretization of the values of the biomarkers. The five biologic markers (ER, PgR, Ki-67, NEU and p53) were used to create the MCA plot (active information). The cluster classifications were used as passive information. The amount of information explained by the first two axes was calculated following the approach suggested by Benzecrí [2]. In fact, due to a geometric property of MCA, the percentages of the inertia explained by each axis are always a pessimistic indicator of the quality of the representation. Therefore, Benzecrí suggested the following indicator:

$$\varphi(\lambda) = \left(\frac{p}{p-1}\right)^2 \left(\lambda - \frac{1}{p}\right)^2$$

where p is the number of variables and λ is the principal inertia. For the melanoma data principal component plots [26] were used to visualize the separation of the tumor samples according to the c-DNA microarray data.

3 Results

3.1 Ambrogi et al. breast cancer biomarkers data

AP was applied to the breast cancer data analyzed in [1] where the final clustering was obtained using a K-Medoids algorithm to generate four clusters.

A graphical evaluation of the effect of the value of the input preference on the number of clusters for the breast cancer data is reported in Fig. 1.

The presence of three main plateaus in correspondence with two, four and five clusters is shown.

When doing the analysis with two clusters (results not shown), by using an input preference value in correspondence with that plateau, results consistent with expectations tied to data from literature were obtained. Indeed, one cluster was associated with null values of ER and PR, the other with high values of these biological markers.

Then the message-passing algorithm was run with an input preference to obtain 4 clusters. The results are reported in the Multiple correspondence analysis plot (Fig. 2). The
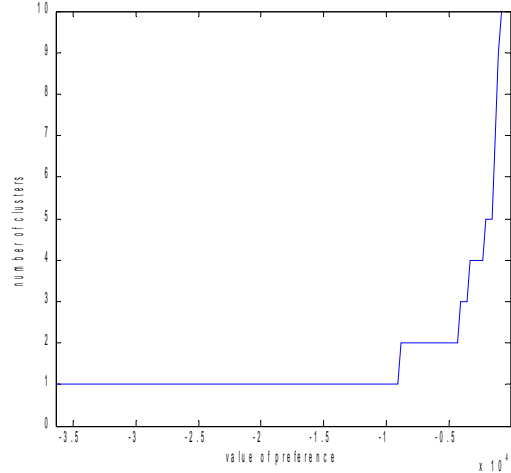

Figure 1. The effect of the value of the input preference on the number of clusters for the Ambrogi et al. data

information explained by the first two axes is near to 89%. Therefore, the two-dimensional plots are expected to be effective representations of the associations displayed.

The MCA plot was generated with the categories of the five biological markers as active information and the AP cluster classification as passive.

As for the contribution of the categories of the biological markers to the construction of the MCA axes, along the first axis there was a separation between high values of PR, ER and the categories of ER and PR absent, high NEU, p53 and Ki-67. The second MCA axis mainly separated the highest values of ER and PR from low PR, Ki-67 and null category of p53.

Null values of PR and high values of NEU were associated with the Cluster 4. Null values of ER, highest values of p53, Neu and Ki-67 were associated with Cluster 3. Therefore, Cluster 3 and Cluster 4 represent groups that are associated with characteristics known to be poor prognostic factors. Whereas, Cluster 1 was associated with highest values of ER and PR, so it seemed to represent subject with characteristics known to be good prognostic factors. Cluster 2 seemed to be associated with intermediated values of PR and ER and null values of Neu; so also this cluster was associated with less aggressive tumour features. As for the triple negative patients, null values of PR, ER and NEU associated with positive values of p53 were grouped in Cluster 3.

The distribution of subject between the classification us-

		AP CLUSTERS			
		1	2	3	4
PREVIOUS	1	253	1	0	2
WORK'S	2	1	122	0	84
CLUSTERS	3	0	1	88	2
	4	1	25	1	52

Table 1. The distribution of subjects between new and old classification (Ambrogi et al. data)
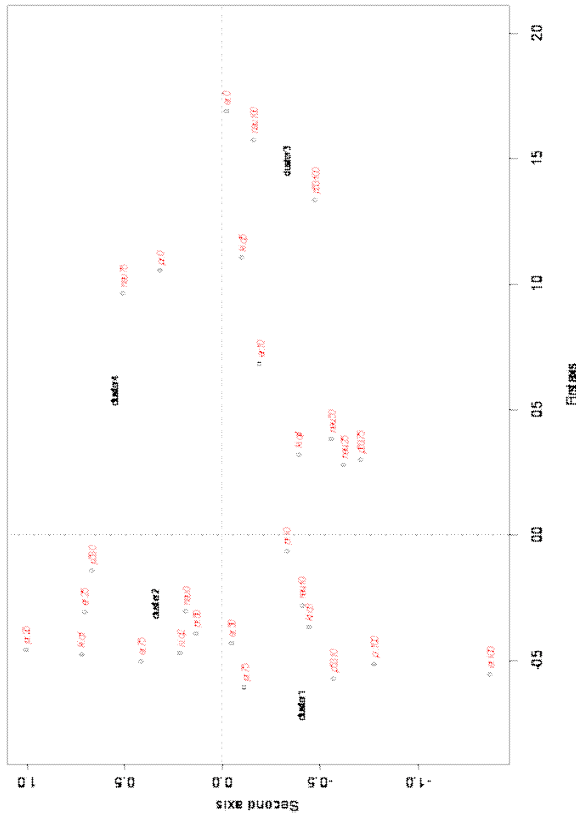

Figure 2. MCA plot of the five discretised biological markers ER, PR, MIB, NEU, p53 (active information) and four clusters (passive information)

ing K-Medoids and the classification using AP is reported in Table 1.

If these results were compared with those of the previous work, null values of PR, ER, NEU and p53 were grouped in Cluster 2, which was the cluster most similar to the characteristics of total sample. Instead, in this new classification null values of PR, ER and NEU associated with null values of p53 lay in Cluster 4, a cluster that is not similar to total sample for the distribution of biological markers and represents groups with poor prognostic factors.

Afterwards, the AP algorithm was applied again to obtain a division of subjects in 5 groups and to compare these results with the four clusters. To do this, a preference value from the plateau in correspondence of five clusters in the first graphic was chosen.

The distribution of subjects between the classification using K-Medoids and the classification using AP is reported in Table 2.

Cluster 4 was more associated with PR absent and high values of NEU. Cluster 2 seemed to be more associated with intermediated values of PR and ER and null values of Neu; it was more associated also with low values of KI-67.

As before Cluster 1 and Cluster 2 were associated with less aggressive tumour features, whereas Cluster 3 and Cluster 4 represents groups with a negative prognosis.

Cluster 5 was mainly characterized by PR absent and high value of NEU. Null values of PR, ER ad NEU associated with positive values of p53 were grouped in Cluster 3. Unlike the previous classification, when we divided subjects in five groups null values of PR, ER and NEU associated with null values of p53 move from Cluster 4 to Cluster 5.

3.2 Bittner et al. melanoma data

Bittner and colleagues [3] attempted to determine if c-DNA microarray data could be used to identify distinct subtypes of cutaneous melanoma, a malignant neoplasm of the skin. In particular they were able to identify two major cancer profiles with different biological characteristics. The result was based on the application of a hierarchical algorithm

		AP CLUSTERS				
		1	2	3	4	5
PREVIOUS	1	211	40	0	1	4
WORK'S	2	0	123	0	0	84
CLUSTERS	3	0	1	87	0	3
	4	1	2	0	74	2

Table 2. The distribution of subjects between new and old classification (Ambrogi et al. data)
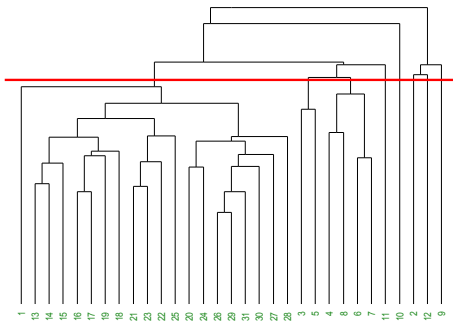
Figure 3. Dendrogram resulting from the application of hierarchical algorithm to Bittner et al. dataset

Figure 4. The effect of the value of input preference on the number of clusters for the melanoma data

and by cutting the dendrogram by visual inspection [10].

In Fig. 3 the dendrogram resulting from the application of a hierarchical algorithm with average linkage and a similarity matrix based on Pearson correlation is reported. The two clusters were obtained by cutting the tree to obtain 5 clusters. In this way the 31 melanomas were divided in a single group comprising 20 melanomas while the remaining 11 (actually grouped in 4 clusters) were considered together.

AP algorithm was applied to the melanoma data using a distance matrix based on correlations. The resulting plot of the cluster number for different preferences levels is reported in Fig. 4. The plot suggests solutions with 2, 3 and 5 clusters. The solution with 5 clusters is the one more similar to the one obtained by Bittner and colleagues. The 3-dimensional principal component plot in Fig. 5 shows the two groups of the 31 melanomas. The red crosses correspond to the "interesting" cluster identified by Bittner and colleagues. The four black squares are tumors classified differently by AP and the hierarchical algorithm. The concordance between the two methods appears satisfying.
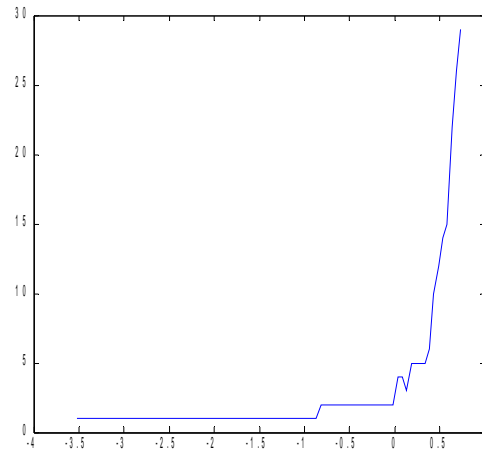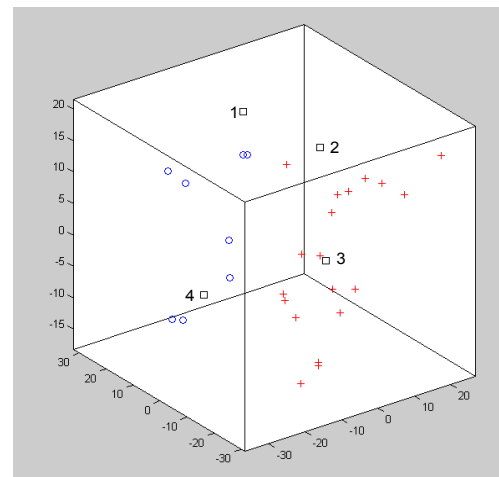

Figure 5. Principal component plot of the gene expression profiles obtained for the 31 melanoma tumors
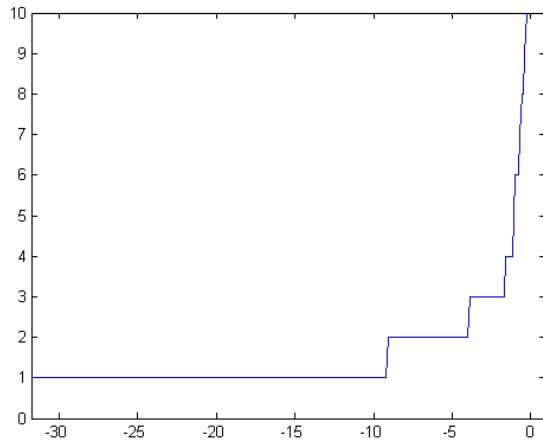
Figure 6. The effect of the value of input preference on the number of clusters for the Perou et al. data

		AP CLUSTERS	
PREVIOUS		1	2
WORK'S	ER pos	26	22
CLUSTERS	ER neg	8	4

Table 3. The distribution of subjects between new and old classification (Perou et al. data)

		AP CLUSTERS			
		1	2	3	4
PREVIOUS	Basal	8	0	0	0
WORK'S	Erb-B2	3	2	1	1
CLUSTERS	Normal	0	11	0	0
	Luminal	7	3	16	10

Table 4. The distribution of subjects between new and old classification (Perou et al. data)

3.3 Perou et al. breast cancer data

In the work of Perou and colleagues [19], an average-linkage hierarchical clustering method, as implemented by [6], was used to group the experimental samples on the basis of similarity in their patterns of expression. In the dendrogram derived from the hierarchical clustering, two large branches were apparent separating the tumour samples into those that were clinically described as ER positive and those that were ER negative. Within these large branches there were smaller branches for which common biological themes could be inferred, namely basal-like, Erb-B2+, normal-like and luminal epithelial/ER+ cancers.

As done in Lama et al. [15], missing values were imputed with the average values of neighbor genes resulted from a k Nearest Neighbors (k-NN) algorithm, setting k=10 (R software package `EMV`, see [24]). The 10 genes 'more similar' to the one with missing value were selected on the basis of the sample correlation: a measure which conforms well to the intuitive biological meaning of 'co-expression'. In this way, all the unavailable data could be recovered and a 1753 x 65 matrix obtained, where rows represented genes and columns represented samples.

The AP algorithm was run and the effect of the value of input preference against the number of clusters is visible in Fig. 6.

By the inspection of plateaus, two, three, four and six may be considered as numbers of clusters determining stable solutions.

When considering two clusters a clear correspondence with the ER+ and ER- groups was not evident. In particular

the distribution shown in Table 3 was obtained (please note that for 5 patients information about their ER status was missing):

Considering three clusters, the ER- group was basically identified by the first cluster, which, on the other hand, contained also several ER+ patients (table not shown).

For four clusters a comparison with the four groups determined by the authors was done: from Table 4 it can be seen that all the basal-like tumours were captured by Cluster 1 and the normal-like tumours were assigned to Cluster 2. Clusters 3 and 4 seem to be characterised by the luminal/epithelial cancers. Instead, the Erb-B2 group was not captured by a single cluster by the AP. It is important to note that in their work, Perou and colleagues assigned three samples to the ER negative group without specifying which of the four subgroups they are part of.

Finally, when our database was divided in six clusters it was found what is reported in Table 5.

Also with this classification in six groups, the Basal-like patients were assigned to Cluster 1. Instead the Normal ones were divided in Clusters 2 and 5. Clusters 3, 4 and 6 seemed to be three Luminal groups. Moreover, the Erb-

		AP CLUSTERS					
		1	2	3	4	5	6
PREVIOUS	Basal	7	0	0	0	1	0
WORK'S	Erb-B2	2	2	0	1	0	2
CLUSTERS	Normal	0	5	0	0	6	0
	Luminal	2	2	15	10	1	6

Table 5. The distribution of subjects between new and old classification (Perou et al. data)
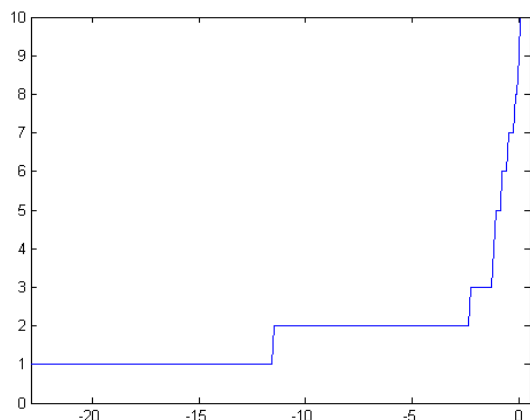
Figure 7. The effect of the value of input preference on the number of clusters for the van't Veer et al. data

B2 group was not captured by a single cluster by the AP.

3.4 van't Veer et al. breast cancer data

In the work of van't Veer et al. [25], an unsupervised hierarchical clustering algorithm allow the authors to cluster 117 tumours on the basis of thousands of genes. For clustering, the 'one minus correlation' distance was used by the authors as described in the Supplementary Information for [25]. In this way two subgroups of breast cancer were detected, differing in ER status (ER-positive and ER-negative) and lymphocytic infiltration (see Fig. 1a in [25]).

293 genes had missing information for all 117 patients. These genes were excluded for the analysis. Other missing values were detected in the remaining data and, ss done for the Perou data, they were imputed with k-NN algorithm. In this way, our data matrix had 117 tumours and 24188 genes of log ratio of the intensities of the red and green channels.

Over this dataset, Affinity Propagation algorithm was applied in order to verify the two ER groups, obtaining the plot in Fig. 7 showing the effect of the value of the common input preference on the number of clusters.

From the plot above, looking at the plateaus, solutions with two, three, five, six and seven clusters were suggested. It is worth noting that AP was not able to indicate a solution with four clusters but one in six groups was evident in accordance with results previously obtained by Perou et al. [19].

When considering two clusters a very good concordance with the two groups obtained separating the ER positive and ER negative patients could be seen. In fact, just 16 over 117 cases did not show such a correspondence (see Table 6).

PREVIOUS WORK'S CLUSTERS		AP CLUSTERS	
		1	2
	ER pos	64	11
	ER neg	5	37

Table 6. The distribution of subjects between new and old classification (van't Veer et al. data)

The split in three clusters is nothing more than a subdivision of the previously determined Cluster 1 into two subgroups (table not shown).

3.5 Sotiriou et al. breast cancer data

Cluster analyses were conducted to search for natural groupings in the profiles. As in the original analysis of Sotiriou and colleagues, before clustering, a screening procedure was applied to eliminate genes showing minimal variation across the set of 99 specimens. Specifically, for each gene, the 5th and 95th percentiles of the ratios were calculated. If the ratio of the 95th to 5th percentile was <3 that gene was not included in the cluster analysis. This process left 706 probe elements for the cluster analysis [22].

Authors applied hierarchical agglomerative clustering to these normalized log ratios by using both compact linkage and average linkage and both Euclidean and one minus Pearson correlation distance metrics. Normalized log ratios were median-centered within each gene for all of the cluster analyses. The clustering results obtained by using compact linkage with one minus Pearson correlation distance applied to the 706 probe elements appeared by visual inspection to yield the most distinctive clusters (remaining blinded to any clinical or outcome variables), and hence this was the clustering algorithm used for the unsupervised cluster analyses based on these probe elements.

Hierarchical clustering results showed that the tumor samples could be confidently separated into two main groups primarily associated with ER status as determined by the ligand-binding assay and confirmed by immunohistochemistry. Eighty-two percent of the tumors in the left main branch are ER-, and 88% of the tumors in the right main branch are ER+ by immunohistochemistry. This finding corroborated the earlier analysis that the major factor discriminating the expression phenotype is ER status. As expected, the ER- cluster had a higher percentage of high-grade tumors than the ER+ cluster. The dendrogram further branched into smaller subgroups within the ER+ and ER-classes.

In the ER- group, tumours characterised by expressions of genes typical of the "basal-like" groups were evident. In particular, based on levels of expressions of specific genes,
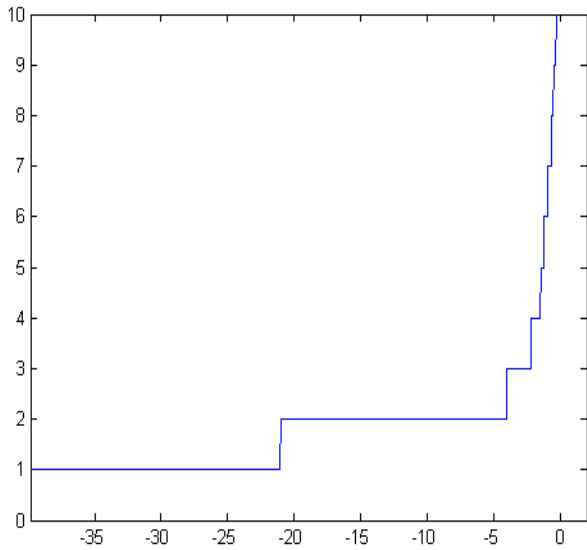
Figure 8. The effect of the value of input preference on the number of clusters for the Sotiriou et al. data

		AP CLUSTERS	
		1	2
PREVIOUS WORK'S CLUSTERS	Grade 1	2	14
	Grade 2	14	24
	Grade 3	31	14
	ER pos	17	48
	ER neg	30	4

Table 7. The distribution of subjects between AP clusters, grade, and ER status (Sotiriou et al. data)

it was possible to distinguish the "basal 1" group from the "basal 2" one. Moreover, in the ER- group, a subgroup characterised by the over-expression of the HER2/NEU oncogen was evident in the basal subgroup.

Concerning the ER positive group, subjects in this cohort were characterised by the expressions of genes specific of the "luminal-like" tumours. This group was further segregated into three smaller subclasses: luminals 1, 2, and 3.

When considering this dataset for the AP analysis, the same technique previously described for dealing with missing values was used.

As for the other datasets considered, the effect of the value of input preference on the number of clusters is reported in Fig. 8.

As it can be seen from the plot, there are three main plateaus in correspondence of solutions with two, three and four clusters and other two smaller ones suggesting six and seven groups. When considering two groups, results consistent with the ones from literature and the two groups of Sotiriou et al. are derived. In fact, classification of subjects affected by breast cancer is associated to the ER status: basing on the immunohistochemistry, the first cluster is characterised by 64% of ER negative, while the second by the 92% of ER positive patients (see Table 7). The distribution of subjects in the two clusters with respect to the histological grade is also reported in Table 7.

From the table above it can be seen that the ER- group (Cluster 1) is basically formed by those patients with histological grade equal to two or three.

After that, in order to obtain a classification of 99 patients in three groups, a preference value in correspondence of the proper plateau was chosen. With this value as an input, the Affinity Propagation algorithm was run again. From the classification in three clusters, a particular group, Cluster 3, is visible: it is mainly formed by patients with negative ER and high histological grade (tables not reported).

The last significant plateau visible in Fig. 8 suggests to consider four groups (tables not shown).

To interpret the four groups obtained with Affinity Propagation on the basis of the genic expression, boxplots of the principal tumour markers were used (plots not shown). This was the only subdivision for which such an analysis could be performed.

Cluster 1 can be characterised as a "luminal-group"; this fact is also confirmed by the overexpression of the ER in this group with respect to the overall population behaviour.

Samples belonging to Cluster 2 present an overexpression of the basal cytokeratins (CKs) and a low expression of the luminal ones. Moreover this group is the only one which have low levels of ER. For these reasons Cluster 2 may be interpreted as a basal-like group even though all Erb-B2 samples were classified in this cluster.

Cluster 3 is characterised by an overexpression of CK8, CK19 (luminal markers), and ER. It can be then considered as a "luminal-like" group.

The values distribution of CKs in Cluster 4 is not different from the one in the overall population. This group is also characterised by the positive estrogen receptor status. It may be concluded that Affinity Propagation is not able to distinguish Cluster 4 from the general population in terms of gene expression.

As in previous examples, a small plateau at six clusters is visible in Fig. 8. The distribution of patients in six clusters from AP compared with the one from the Sotiriou's groups is reported in Table 8.

Sotiriou's basal groups (basal 1 and basal 2) were both captured by AP Cluster 2, while the Erb-B2 was now associated with Cluster 6. The Luminal groups were assigned to different clusters: in particular, luminal 1 was described

		AP CLUSTERS					
		1	2	3	4	5	6
	Basal 1	0	14	0	0	0	2
PREVIOUS	Basal 2	0	8	0	1	1	0
WORK'S	Erb-B2	0	0	0	0	0	7
CLUSTERS	Luminal 1	0	0	1	8	10	0
	Luminal 2	9	0	10	2	2	0
	Luminal 3	10	2	3	1	2	6

Table 8. The distribution of subjects between new and old classification (Sotiriou et al. data)

by samples in Clusters 4 and 5, luminal 2 in Cluster 3 and luminal 3 was mainly associated to Cluster 1 but presents several cases in other clusters as well.

4 Discussion

Cluster analysis is a powerful technique to explore complex diseases and improve prognosis. The recent literature on omic data is rich of new methods of cluster analysis able to deal with huge datasets. Moreover techniques of visualizations are usually adopted to suggest the number of clusters [6].

At the same time many papers warn against the possible misuse of clustering techniques [10].

One of the main problems is the subjectivity of the analysis and the ability of clustering algorithms to create clusters even in absence of real structure.

The choice of the number of clusters is one of the main problems to be faced when applying this kind of analysis. The possibility to use algorithms that incorporate a criterion for the choice of the optimal partition is one of the achievement of the recent developments in this research field. The Affinity Propagation algorithm is characterized by a simple software implementation and it has the ability to suggest the cluster number. In addition, this algorithm has the advantage of taking into account real data points as exemplars; in this way, also categorical data may be analysed using, as similarity, a different distance from the Euclidean one. By looking for a set of exemplars, and not for a set of centroids, the search space is restricted, improving the computational efficiency of the Affinity Propagation. In this work it was demonstrated how the algorithm is in agreement with the solutions obtained with much more effort with traditional algorithms and indexes for the cluster number choice. Moreover the range of the suggested solutions gives insights in the hierarchical structure of the data highlighting different level of information for the treatment of cancer patients well in accordance with previous knowledge. In particular the solution with two clusters for Ferrara breast cancer

data, evidenced in Fig. 1, reflects the well known separation between tumors ER positive and negatives. This is a very important distinction and, in fact, in a number of paper of the pre-genomic era the number of clusters considered was in fact two [20, 18]. The solution with four clusters is in agreement with the solution selected in the previous work and the four clusters obtained are similar to that created by the PAM algorithm. The solution with five clusters suggests a possible more complex pattern to be explored. The clustering obtained by AP on the melanoma data is able to reproduce the interesting findings of Bittner and colleagues having the advantage of avoiding any arbitrary choice due to the visual inspection of the dendrogram. Over Perou et al. the AP could almost reproduce previous classification: the basal-like tumours were captures by Cluster 1, and the normal-like ones were assigned to Cluster 2; Clusters 3 and 4 represent the luminal-epithelial tumours, while the Erb-B2 group was not identified by AP in a single cluster. The solution with two clusters for van't Veer et al. data reflects with a few number of exceptions, the separation between ER+ and ER- groups evidenced in the original work. Applying AP over the dataset of Sotiriou and colleagues, four groups were evident and it was also possible to characterised them on the basis of gene expression of the tumour markers. AP Clusters 1 and 3 were associated with high values of luminal cytokeratins and overexpression of ER, while Cluster 2 presents an overexpression of the basal CKs and low levels of ER. For Cluster 4 was not possible to obtain a distinctive characterisation, as it presented similar features to the overall population.

In conclusion, it is worth noting that, for the last three datasets analysed, a common solution in 6 clusters emerged by the AP algorithm, while one in 4 groups was not evident in van't Veer et al. work.

As a future work, the characterisation of the AP groups on the basis of gene expressions will be carried out and compare with the one obtained in the original papers.

Acknowledgments

References

[1] F. Ambrogi, E. Biganzoli, P. Querzoli, S. Ferretti, P. Boracchi, S. Alberti, E. Marubini, and I. Nenci. Molecular subtyping of breast cancer from traditional tumor marker profiles using parallel clustering methods. *Clinical Cancer Research*, 12(3):781–790, 2006.

[2] J. Benzecrí. *Cahiers de l'Analyse des donnes*, 4:377–378, 1979.

[3] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, N. Hayward, and J. Trent. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406:536–540, 2000.

[4] R. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communs statist*, 3:1–27, 1974.

[5] D. Dueck, B. Frey, N. Jojic, V. Jojic, G. Giaever, A. Emili, G. Musso, and R. Hegele. *Constructing Treatment Portfolios Using Affinity Propagation*, volume 4955 of *Lecture Notes in Computer Science*, pages 360–371. Springer Berlin / Heidelberg, 2008.

[6] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95:14863–8, 1998.

[7] B. Frey and D. Dueck. *Mixture Modeling by Affinity Propagation*. Freely available at http://books.nips.cc/papers/files/nips18/NIPS2005_0799.pdf.

[8] B. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–6, 2007.

[9] G. Getz, E. Levine, and D. E. Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci U S A*, 97(22):12079–84, 2000.

[10] D. Goldstein, G. Debashis, and E. Conlon. Statistical issues in the clustering of gene expression data. *Statistica Sinica*, 12:219–240, 2002.

[11] M. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, 1994.

[12] J. Hartigan. *Clustering Algorithms*. Wiley series in probability and mathematical statistics. Applied Probability and Statistics. New York: Wiley, 1975.

[13] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis*. Wiley series in probability and mathematical statistics. Applied Probability and Statistics. New York: Wiley, 1990.

[14] W. Krzanowski and Y. Lai. A criterion for determining the number of clusters in a data set. *Biometrics*, 44:23–34, 1985.

[15] N. Lama, F. Ambrogi, L. Antolini, P. Boracchi, and E. Biganzoli. Some issues and perspective in microarray data analysis in breast cancer: the need for an integrated research. In *Proceedings of the 1st European Workshop on the Assessment of Diagnostic Performance*, 2004.

[16] L. Lebart, A. Morineau, and M. Piron. *Statistique Exploratoire Multidimensionnelle*. Dunod, Paris, 1995.

[17] M. Leone, Sumedha, and M. Weigt. Clustering by soft-constraint affinity propagation: Applications to gene-expression data. *Bioinformatics*, 23:2708–2615, 2007.

[18] S. Mnard, P. Casalini, G. Tomasic, S. Pilotti, N. Cascinelli, R. Bufalino, F. Perrone, C. Longhi, F. Rilke, and M. Colnaghi. Pathobiologic identification of two distinct breast carcinoma subsets with diverging clinical behaviors. *Breast Cancer Res Treat*, 55(2):169–77, 1999.

[19] C. Perou, T. Sorlie, M. Eisen, M. Van De Rijn, S. Jeffrey, C. Rees, J. Pollack, D. Ross, H. Johnsen, L. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. Zhu, P. Lonning, A. Borresen-Dale, P. Brown, and D. Botstein. Molecular portraits of human breast tumours. *Nature*, 406:747–752, 2000.

[20] P. Querzoli, S. Ferretti, G. Albonico, E. Magri, D. Scapoli, M. Indelli, and I. Nenci. Application of quantitative analysis to biologic profile evaluation in breast cancer. *Cancer*, 76(12):2510–7, 1995.

[21] R. Simon, E. Korn, L. McShane, M. Radmacher, G. Wright, and Y. Zhao. *Design and Analysis of DNA Microarray Investigations*. Springer, 2004.

[22] C. Sotiriou, S.-Y. Neo, L. McShane, E. Korn, P. Long, A. Jazaeri, P. Martiat, S. Fox, A. Harris, and E. Liu. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci USA*, 100:10393–10398, 2003.

[23] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *J.R. Statist. Soc.*, B,63:411–423, 2001.

[24] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520525, 2001.

[25] L. van't Veer, H. Dai, M. van de Vijver, Y. He, A. Hart, M. Mao, H. Peterse, K. van der Kooy, M. Marton, A. Witteveen, G. Schreiber, R. Kerkhoven, C. Roberts, P. Linsley, R. Bernards, and S. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[26] W. Venables and B. Ripley. *Modern Applied Statistics with S*. New York: Springer, 4th edition, 2002.