Cancer profiles by Affinity Propagation

Federico Ambrogi[2], Elena Raimondi[1], Daniele Soria[3], Patrizia Boracchi[2] and Elia Biganzoli[12]
[1]*Struttura di Statistica e Biometria, Fondazione IRCCS Istituto Nazionale Tumori, Milano, Italy*
[2]*Istituto di Statistica Medica e Biometria "GA Maccacaro", Università degli Studi di Milano, Milano, Italy*
[3]*University of Nottingham, School of Computer Science, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB, UK*
E-mail: federico.ambrogi@unimi.it

Abstract

The affinity propagation algorithm is applied to a problem of breast cancer subtyping using traditional biologic markers. The algorithm provides a procedure to determine the number of profiles to be considered.

A well know breast cancer case series was 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 [5], 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 [18] on clustering of cutaneous melanoma and the works of van't Veer [16] and Perou [13] 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 [13,16]. 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, [3] Krzanowski and Lai, [10], Hartigan, [8] and Tibshirani et al. [15], were considered. It is worth noting that the visual inspection of the dendrogram is an informal method to determine the number of clusters. Such a procedure was criticized in [4] as it can cause difficulty in assessing the validity of the grouping.

According to Getz [6], 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 [17], will be adopted to cluster cancer patients in order to evaluate its performance with respect to the traditional applications. 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

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.

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. [18]. 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. The data and the original analysis are fully described in the book "Design and Analysis of DNA Microarray Investigations" by Simon and colleagues [21].

2.2. Statistical Methods

The clustering technique affinity propagation (AP, [17]) 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 [9] (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 [20].

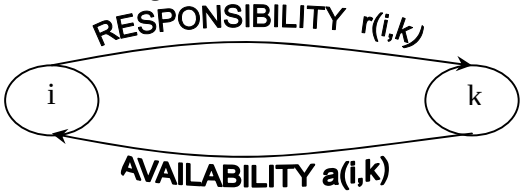
The first step for the algorithm implementation is to choose a measure of similarity, s(i,k), between all pairs of data points. In AP terminology, s(i,k) quantifies how well the data point with index k is suited to be the exemplar for data point i. Generally, as similarity, it is used the negative Euclidean distance. In the case of c-DNA data the Pearson correlation is generally used as similarity measure [18].

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 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:

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

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:

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

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 results. 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 exemplars if k=i or identifies the data point that is the exemplars for point i.
At the end of the message passing, we obtain the number of clusters and the labels for each data point of its exemplars.
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 [19]. The algorithm is feasible even in presence of very large data sets.
Multiple correspondence analysis (MCA), see Greenacre (7) or Lebart et al. (11), 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 [22] were used to visualize the separation of the tumor samples according to the c-DNA microarray data.

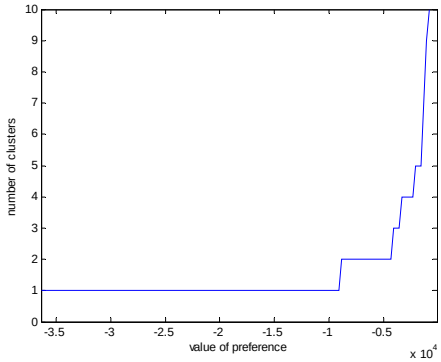
3. Results

3.1. 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 we did the analysis with two clusters (results not shown), by using an input preference value in correspondence with that plateau, we obtained results consistent with expectations tied to data from literature. 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 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, Cluster1 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.

Fig 1: The effect of the value of the input preference on the number of clusters.



The distribution of subject between the classification using K-Medoids and the classification using AP is reported in Table 1.

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

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 Cluster1 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.

Table 1: The distribution of subjects between new and old classification

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

Table 2: The distribution of subjects between new and old classification

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

3.2.Bittner et al. melanoma data

Bittner and colleagues [18] 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 and by cutting the dendrogram by visual inspection [4].

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 cluster 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.

Fig 2: MCA plot of the five discretized biological markers ER, PR, MIB, NEU, P53 (active information) and four clusters (passive information)

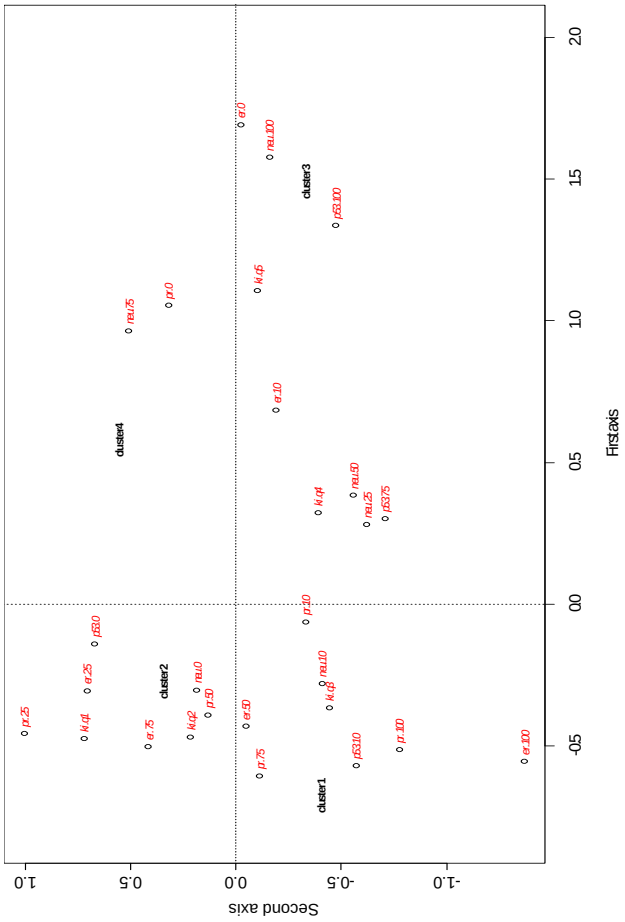


Fig 3: Dendrogram resulting from the application of hierarchical algorithm to Bittner et al. dataset.

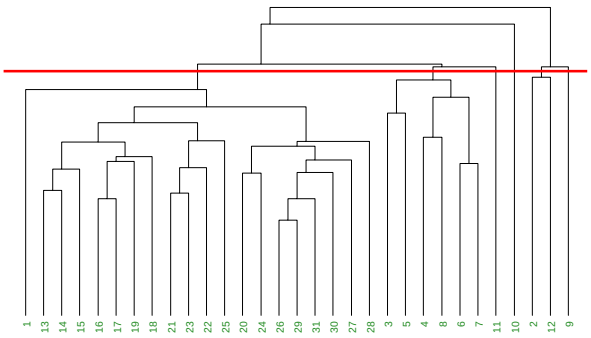


Fig 4: The effect of the value of the input preference on the number of clusters for the melanoma data

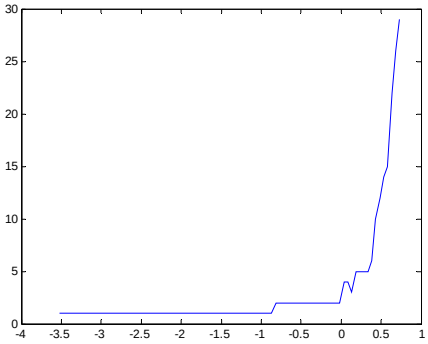
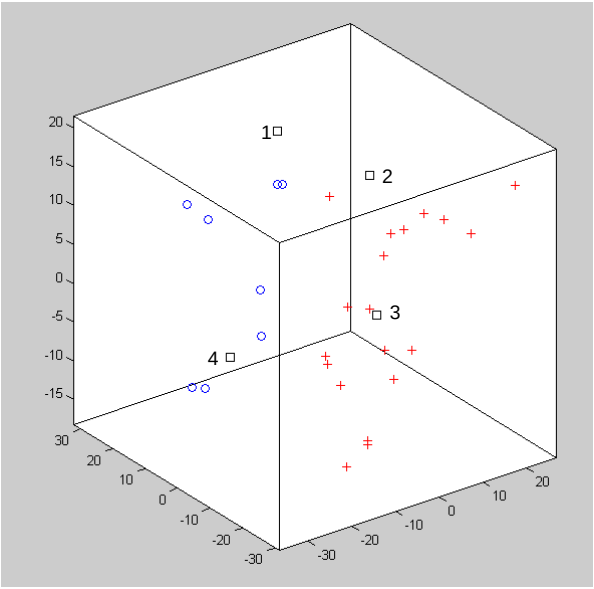


Fig 5: Principal component plot of the gene expression profiles obtained for the 31 melanoma tumors.



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 [5].

At the same time many papers warn against the possible misuse of clustering techniques [4].

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 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 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 [14, 12]. 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.

REFERENCES

1. Ambrogi F, Biganzoli E, Querzoli P, Ferretti S, Boracchi P, Alberti S, Marubini E, Nenci I. Molecular subtyping of breast cancer from traditional tumor marker profiles using parallel clustering methods. Clin Cancer Res. 2006 Feb 1;12(3 Pt 1):781-90.
2. Benzécri JP 1979 Cahiers de l'Analyse des données,4,377-378.
3. Calinski RB, Harabasz J 1974 Communs statist,3,1-27.
4. Goldstein DR, Debashis G and Conlon EM Statistical issues in the clustering of gene expression data Statistica Sinica 12(2002), 219-240
5. Eisen MB, Spellman PT, Brown PO and Botstein D. (1998). Cluster Analysis and Display of Genome-Wide Expression Patterns. Proc Natl Acad Sci U S A 95, 14863-8.
6. Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. Proc Natl Acad Sci U S A. 2000 Oct 24;97(22):12079-84.
7. Greenacre, MJ. 1994 "Theory and Applications of Correspondence Analysis". Academic Press.
8. Hartigan J. 1975 "Clustering Algorithms". Wiley,New York
9. Kaufman L., Rousseeuw P. 1990 "Finding groups in data", Wiley, New York
10. Krzanowski WJ, Lai YT 1985 Biometrics,44,23-34.
11. Lebart L, Morineau A, Piron M. 1995 "Statistique exploratoire multidimensionnelle" Dunod, Paris.
12. Ménard S, Casalini P, Tomasic G, Pilotti S, Cascinelli N, Bufalino R, Perrone F, Longhi C, Rilke F, Colnaghi MI. Pathobiologic identification of two distinct breast carcinoma subsets with diverging clinical behaviors. Breast Cancer Res Treat. 1999 May;55(2):169-77.
13. Perou C.M., Sørlie T., Eisen M.B., van de Rijn M., Jeffrey S., Rees C.A., Pollack J.R., Ross D.T., Johnsen H., Akslen L.A., Fluge Ø., Pergamenschikov A., Williams C., Zhu S.X., Lønning P.E., Børresen-Dale A., Brown P.O., Botstein D. 2000 Nature,406, 747-752.
14. Querzoli P, Ferretti S, Albonico G, Magri E, Scapoli D, Indelli M, Nenci I. Application of quantitative analysis to biologic profile evaluation in breast cancer. Cancer. 1995 Dec 15;76(12):2510-7.
15. Tibshirani R, Walther G, Hastie T 2001 J.R. Statist. Soc. B,63,411-423.
16. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH 2002, Nature,415,530-6.
17. Frey BJ, Dueck D. Clustering by passing messages between data points" Science 315(5814): 972-6
18. Bittner M, Meltzer P, Chen Y et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature,406,536-540.
19. Leone M., Sumedha and Martin Weigt. Clustering by soft-constraint affinity propagation: applications to gene-expression data. Bioinformatics, 23,2708-2615
20. Frey BJ, Dueck D. Mixture modeling by affinity propagation. Freely available at http://books.nips.cc/papers/files/nips18/NIPS2005_0799.pdf
21. Simon, R.M., Korn, E.L., McShane, L.M., Radmacher, M.D., Wright, G.W., Zhao, Y., Design and Analysis of DNA Microarray Investigations, Springer, 2004.
22. Venables WN and Ripley BD, Modern Applied Statistics with S, 2002, Fourth Edition, Springer.