Information Based Clustering of Gene Expression Signatures in Primary Breast Carcinoma Patients

Milan Bimali*[1, 2], Michael Brimacombe[3].

Email: mbimali@kumc.edu, mbrimacombe@kumc.edu

1: Office of Research, University of Kansas School of Medicine, Wichita, KS.

2: Department of Preventive Medicine and Public Health, University of Kansas School of Medicine, Wichita, KS

3: Department of Biostatistics, University of Kansas Medical Center, Kansas City, KS.

*Correspondence to Milan Bimali, 1010 N. Kansas Ave., Wichita, KS, 67214. Email: mbimali@kumc.edu. Phone: 316-293-3808. Fax: 316-293-2686.

None of the authors have conflict-of-interest or financial disclosures to declare.

Running title: Information based Clustering of Gene Expression

Word Count: 2120

Abstract Word Count: 115

Number of figures: 3

Number of tables: 2

Abstract

The application of a novel clustering approach is developed that takes into account the structure of gene expression profiles in relation to the distributional assumption as well as information based similarity among gene expressions in the data. It is assumed that the gene expression profile for each subject follows a known distribution and thus a set of relative likelihood functions (likelihood functions rescaled by their mode) can be constructed. The relative likelihood functions thus obtained are further weighted (scaled) by the observed Fisher information to incorporate information related accuracy across the gene expression profiles. The subjects are then eventually clustered based on a distance matrix reflecting the weighted relative likelihoods and applying standard clustering methods.

Keywords: Likelihood functions, k means clustering, distance matrix, Fisher Information.

1. Introduction

Clustering is a grouping procedure focused on identifying subgroups within a dataset (Rencher 2002). While traditional non-parametric clustering methods such as hierarchical clustering and k-means clustering algorithms are commonly used (D'Haeseleer 2005), there has been work dedicated to parametric clustering approaches such as model-based clustering (Bouveyron and Brunet-Saumard 2014). Despite the differences in assumptions and approaches, the objective of most clustering algorithms is to classify subjects or observations into one of a finite set of disjoint clusters while ensuring that subjects within a cluster are more similar than subjects across clusters.

In the context of gene expression data, clustering techniques have been employed to identify sub-groups of patients at the molecular level, to understand gene function and regulation. It has been applied successfully to group similarly expressed genes across a set of subjects as well as to group subjects with similar gene expression profiles (Jiang, Tang et al. 2004). In the context of clustering gene expression data, hierarchical clustering and k means clustering are more commonly used (D'Haeseleer 2005). Other approaches such as fuzzy c means clustering, self-organizing maps, and model-based clustering have also been employed (Toronen, Kolehmainen et al. 1999, Yeung, Fraley et al. 2001, Gasch and Eisen 2002, Nikkila, Toronen et al. 2002, Covell, Wallqvist et al. 2003, Huang, Wei et al. 2006, Arima, Hakamada et al. 2008, Zhang, Adamu et al. 2011, Shahdoust, Hajizadeh et al. 2013, Zhang and Shen 2014).

Our recent work proposes a clustering approach based on the properties of the observed likelihood and Fisher Information for each observation in the dataset (Bimali and Brimacombe 2015). Unlike the traditional non-parametric and model based approach, the proposed method takes into account the structure of data in relation to the distributional assumption as well as information based similarity among observations in the data. In the context of gene expression,

the proposed method assumes that gene expression profile for each subject is follows a known distribution and thus a set of relative likelihood functions (likelihood function scaled by their mode) can be constructed. The relative likelihood functions can be viewed as a transformation of the original gene expression profiles. These relative likelihood functions are then further weighted by the Fisher Information to obtain the weighted relative likelihood function. This is evaluated at different values of the parameter to obtain a data based distance matrix which can be subjected to the clustering algorithms. The proposed clustering approach takes into account the variation in mean expression levels as well as the observed Fisher Information across the patients.

Here we apply the proposed clustering approach to the publicly available dataset by Van De Vijer et al in clustering primary breast carcinomas patients based on a previously recommended set of 70 gene expression profile (van de Vijver, He et al. 2002). The agreement between the proposed clustering approach and authors' classification has been examined. The clusters obtained are also examined in relation to two clinical features – time to overall survival; and time to metastases**.**

2. Data

The dataset has been made available by Van De Vijver et al at http://ccb.nki.nl/data/. The authors describe the study subject as patients having either I or II breast cancer and younger than 53 years. The authors have made available expression profiles for 24496 genes, of which 70 genes formed a subset. Clinical covariates such as time to overall survival, time to distant metastases, death status, and the number of positive nodes were also provided. Van De Vijer et al used 70 gene expression profiles that were identified by Veer et al, to classify 295 patients with primary breast carcinomas into two groups – poor prognosis groups and good prognosis group (van 't Veer, Dai et al. 2002). Among the 295 patients, 180 were classified into poor prognosis groups while 115 were classified into good prognosis groups.

The prognostic classification was based on correlation of these 70 genes with the average profile of these 70 genes in tumors from patients with a good prognosis. The threshold of 0.4, used for correlation coefficients, was determined based on a previous study of 78 tumors which resulted in a false negative rate of 10 percent. The two groups differed significantly with respect to the overall 10-year survival time as well as with respect to time to distant metastases. The authors mentioned that the classification system based on 70 genes outperformed all clinical variables in predicting the risk of distant metastases within 5 years. The dataset provided used by Van De Vijer has been made available publicly. We restrict our attention to the 70 gene expression profiles and examine the subsequent clusters of 295 patients formed based on these 70 gene expressions.

3. Method

Genes in each subject are assumed to follow normal distribution and thus likelihood functions are constructed. The likelihood functions are further scaled by their maxima to transform them into relative likelihood functions. A data matrix is then developed by evaluating the weighted relative likelihood functions at different values in the parameter space, the weights being the Fisher Information matrix evaluated at the mode of the likelihood functions. The proposed approach thus takes into account the structure of data via the distributional assumption as well as information similarity between observations in the data. We assume that the genes for each subject follow a normal distribution.

Let us consider data matrix $\boldsymbol{X} = (\boldsymbol{x_1} \quad \dots \quad \boldsymbol{x_n})'$ where $\boldsymbol{x_i} = (x_{i1} \quad \dots \quad x_{ik})$

$x_{i1}, \dots, x_{ik}$ are iid observations with pdf $f_i(x_{ij}|\theta_i); j = 1, \dots, T_i$ and $\boldsymbol{\theta} = (\theta_1 \quad \dots \quad \theta_n)$

We assume that $\theta_i's$ share the same support. Thus for each θ_i we can construct likelihood functions reflecting assumed pdf giving rise to n likelihood functions based the data matrix $\boldsymbol{X}$.

$$\boldsymbol{L_X}(\boldsymbol{\theta}) = (L_{\boldsymbol{x_1}}(\theta_1) \quad \dots \quad L_{\boldsymbol{x_n}}(\theta_n))'$$

where $L_{\boldsymbol{x_i}}(\theta_i) = \prod_{j=1}^{k} f(x_{ij}|\theta_i)$. Let $\hat{\theta}_i$ be the mle of θ_i. Then the relative likelihood function for each θ_i can be constructed as follows:

$$\boldsymbol{R_X}(\boldsymbol{\theta}) = (R_{\boldsymbol{x_1}}(\theta_1) \quad \dots \quad R_{\boldsymbol{x_n}}(\theta_n))'$$

with $R_{\boldsymbol{x_i}}(\theta_i) = \frac{L_{\boldsymbol{x_i}}(\theta_i)}{L_{\boldsymbol{x_i}}(\hat{\theta}_i)}$. Note that the Fisher information, by definition is the same for both the initial and relatively re-weighted likelihood function.

To improve the assessment of similarity across the set of evaluated likelihoods, the Fisher information matrix for each observation can be used as a weight and we have;

$w_{x_i}(\hat{\theta}) = I(\hat{\theta})$ where $I(\theta) = E\left(\left(\frac{\partial}{\partial\theta} log(L(\theta|\boldsymbol{x_i}))\right)^2\right)$

For exponential families with iid observations, note that the Fisher Information matrix can be simplified to $I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\log l(\theta|\boldsymbol{x}_i)\right)$

The value of the likelihood functions can be evaluated at different values of the $\theta_i's$. For each observation $\boldsymbol{x_i}$, we can compute the value of likelihood functions at k different θ_i values. Thus we can construct a matrix $\boldsymbol{P_X}$ with rows containing the weighted relative likelihood functions evaluated at different values of θ_i .

$$\boldsymbol{P_X} = \begin{pmatrix} w_{\boldsymbol{x_1}}(\hat{\theta}) & \cdots & w_{\boldsymbol{x_1}}(\hat{\theta}) \\ \vdots & \ddots & \vdots \\ w_{\boldsymbol{x_n}}(\hat{\theta}) & \cdots & w_{\boldsymbol{x_n}}(\hat{\theta}) \end{pmatrix} \circ \begin{pmatrix} R_{\boldsymbol{x_1}}(\theta_1) & \cdots & R_{\boldsymbol{x_1}}(\theta_k) \\ \vdots & \ddots & \vdots \\ R_{\boldsymbol{x_n}}(\theta_1) & \cdots & R_{\boldsymbol{x_n}}(\theta_k) \end{pmatrix}$$

where $w_{x_i}(\hat{\theta})$ is the Fisher Information evaluated at the *mle*. $R_{x_i}(\theta_j)$ is the value of the relative likelihood function for x_i evaluated at θ_j and $\circ$ is the Hadamard product operator between the two matrices. The matrix $\boldsymbol{P_X}$ can be subjected to various standard clustering algorithms to explore for patterns and clusters in the data matrix $\boldsymbol{X}$.

Under the assumption of normality of genes for each subject, the weighted relative likelihood function for each subject can be shown to be as follows:

$$w_{\boldsymbol{x_i}}(\theta) = \frac{n}{\hat{\sigma}^2} \times \exp\left(-0.5 \times \frac{n}{\hat{\sigma}^2}\left(\theta - \hat{\theta}\right)^2\right)$$

The above weighted relative likelihood function can be evaluated across different values of θ for each subject to obtain a matrix of weighted relative likelihood functions.

4. Analysis

The assumption of normality of genes for each subject was tested using the Shapiro-Wilk's test of non-normality. Among the 295 subjects, 88 subjects showed significant deviation from the normality assumption based on α −level of 0.01, and were thus excluded from the analysis. Table 1 provides summary statistics on the survival time, time to distant metastases, for good and poor prognosis subjects. The pair-wise correlation of genes across the subjects was examined. The correlations of the gene expression profiles across 207 patients were examined and genes that were moderately to highly correlated with other genes were excluded to be consistent with the iid assumption. The absolute correlation threshold was set at $0.8, 0.7,$ and 0.5 respectively. Thus the data matrix that was analyzed consisted of 207 patients with gene expression profiles whose correlation (absolute value) was below the specified threshold.

For each of the 207 subjects, a weighted relative likelihood function was constructed. The weighted relative likelihood function was then evaluated at 1000 equi-spaced intervals within $(-0.4,0.3)$. For each gene expression profile, the evaluated weighted relative likelihoods were non-zero in this range. The matrix of evaluated weighted relative likelihood function was then subjected to k means clustering with 2 clusters. Choosing two clusters allows us to examine the agreement between the authors classification of poor and good prognosis as the cluster formed based on proposed approach.

4.1 Correlation threshold set at 0.8

The number of gene expression profiles dropped from 70 to 64 $i.e.$ 6 genes were highly correlated (correlation $\geq$ 0.8) with other genes and were dropped from analysis. The data matrix obtained by evaluating the weighted relative likelihood functions was subjected to k means clustering. Fig 1 provides a plot of the weighted relative likelihood functions colored by their cluster assignment. Log-rank test showed that the two clusters differed significantly with respect to overall survival time (p value $= 5.5 \times 10^{-4}$) as well as time to distant metastases (p-value $= 4.79 \times 10^{-3}$) (fig 1). Table 2 provides a summary of the agreement between the authors' classification and the clustering based on weighted relative likelihood function. The summary statistics of the two clusters is provided in Table 1.

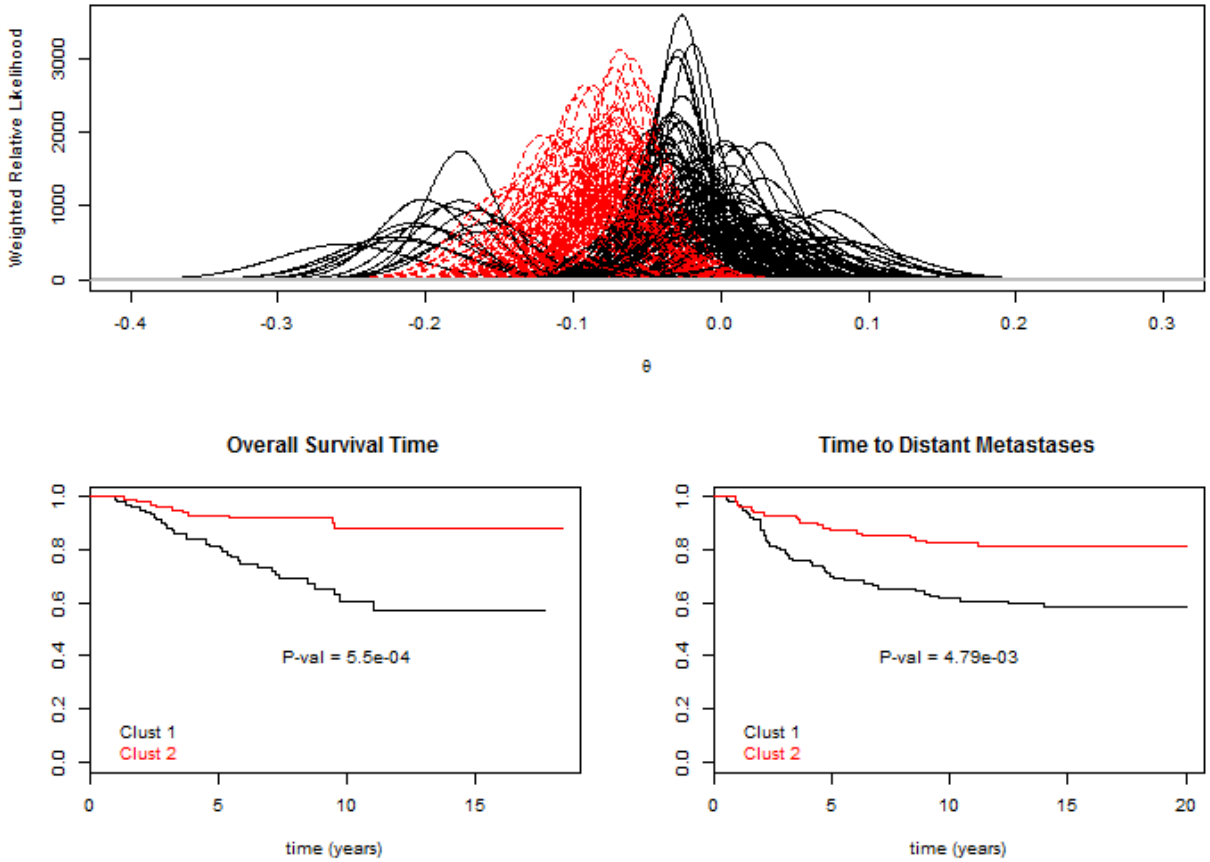

Fig 1: Top – Plot of weighted relative likelihood functions evaluated at 1000 different values of the mean parameter. Bottom-left – KM survival plots for overall survival time between the two clusters. Bottom-right – KM survival plots for time to distant metastases between the two clusters. Correlation threshold set at 0.8.

4.2 Correlation threshold set at 0.7

The number of gene expression profiles dropped from 70 to 58, i.e. 12 genes that are moderately correlated with other genes were dropped from analysis. The data matrix obtained by evaluating the weighted relative observed likelihood function was subjected to k means clustering. Fig 2 provides a plot of the weighted relative likelihood functions colored by their cluster assignment. Log rank test showed that the two clusters differed significantly with respect to overall survival time (p value $\approx 10^{-5}$) as well as time to distant metastases (p-value = 1.28×10^{-3}) (Fig 2). Table 2 provides a summary of the agreement between the authors' classification and the clustering based on weighted relative likelihood function. The summary statistics of the two clusters is provided in Table 1.

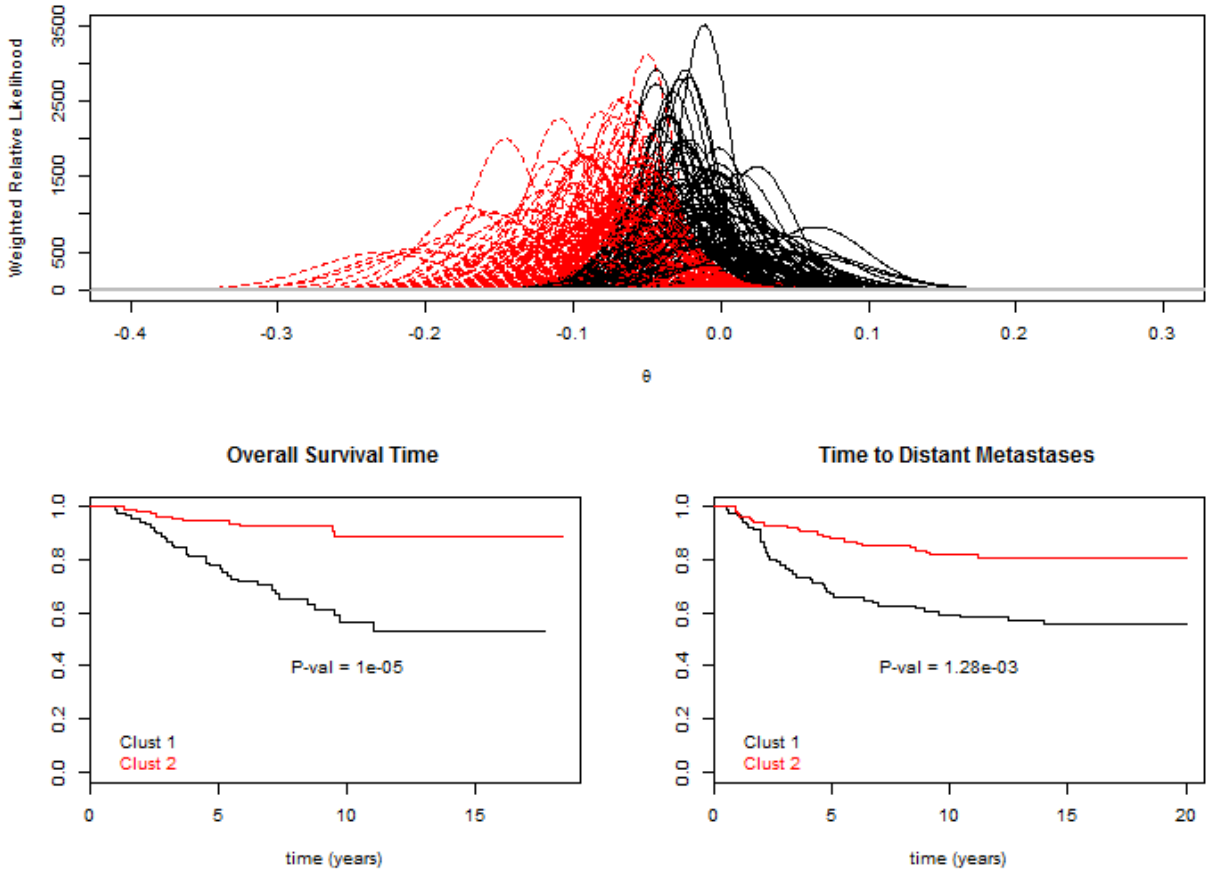

Fig 2: Top – Plot of weighted relative likelihood functions evaluated at 1000 different values of the mean parameter. Bottom-left – KM survival plots for overall survival time between the two clusters. Bottom-right – KM survival plots for time to distant metastases between the two clusters. Correlation threshold set at 0.7.

4.3 Correlation threshold set at 0.5

The number of gene expression profiles dropped from 70 to 45 i.e. there were 25 genes moderately correlated (correlation ≥ 0.5) with other genes and were dropped from analysis. The data matrix obtained by evaluating the weighted relative observed likelihood function was subjected to k means cluster with 2 clusters. Fig 3 provides a plot of the weighted relative observed relative likelihood functions colored by their cluster assignment. Log rank test showed that the two clusters differed significantly with respect to overall survival time (p value ≈ 0.012); however there was no significant difference between time to distant metastases (p-value ≈ 0.055) (Fig 3). Table 2 provides a summary of the agreement between the authors' classification and the clustering based on weighted relative likelihood function. The summary statistics of the two clusters is provided in Table 1.

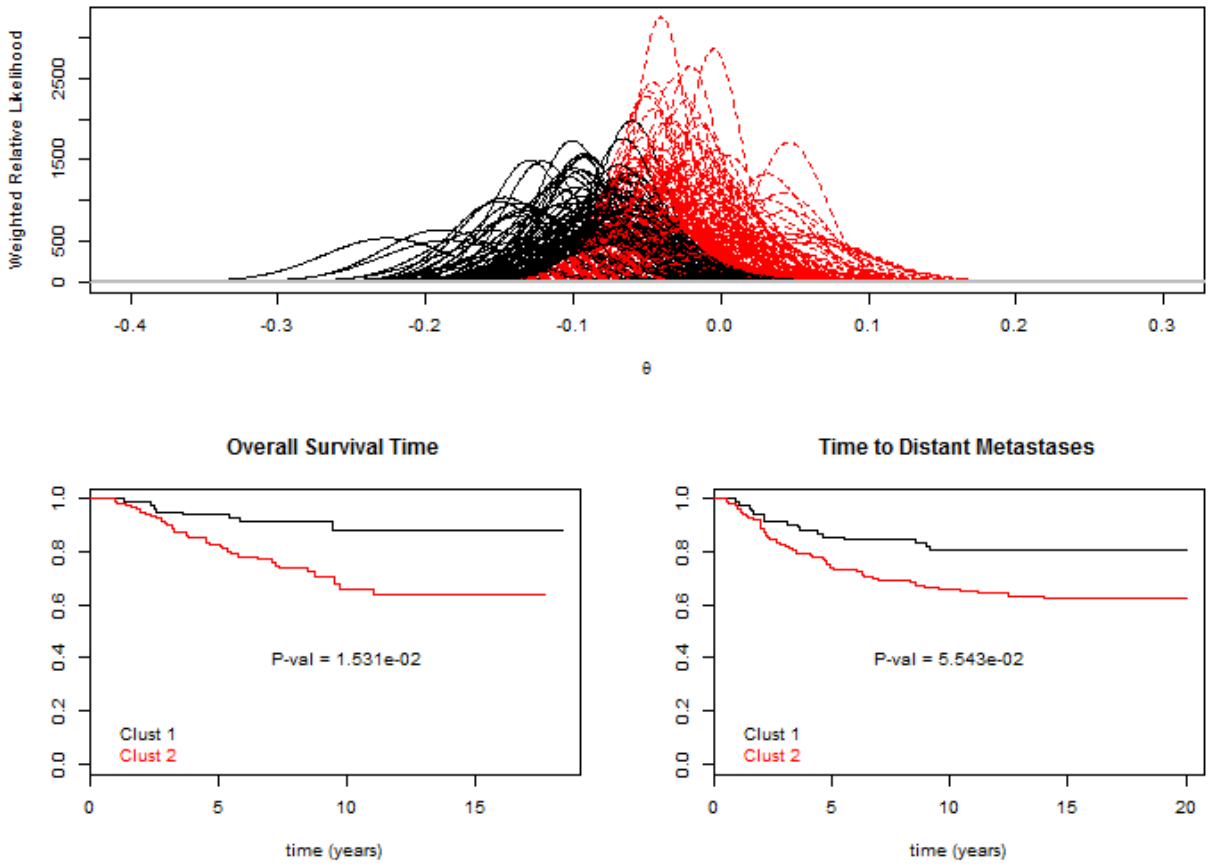

Fig 3: Top – Plot of weighted relative likelihood functions evaluated at 1000 different values of the mean parameter. Bottom-left – KM survival plots for overall survival time between the two clusters. Bottom-right – KM survival plots for time to distant metastases between the two clusters. Correlation threshold set at 0.5.

Prognosis/Cluster	Count	Median Overall Survival	Median Time to Metastases
Good	92	8.91 (4)	4.98 (12)
Poor	115	6.93 (40)	2.94 (50)
Correlation threshold set at 0.8 (n = 64)			
Cluster 1	104	6.88 (34)	3.03 (43)
Cluster 2	103	8.81 (10)	3.66 (19)
Correlation threshold set at 0.7 (n = 58)			
Cluster 1	91	8.77 (10)	4.05 (22)
Cluster 2	116	6.93 (34)	4.14 (40)
Correlation threshold set at 0.5 (n = 45)			
Cluster 1	116	7.27 (35)	13.98 (45)
Cluster 2	91	8.37 (9)	16.96 (17)
Note: Values for $(n\ =)$ represents the number of gene expressions used for clustering. Values in parenthesis in table represent number of subjects experiencing event of interest.			

Table 1: Median Survival time (overall survival and time to metastases) between the two clusters at different correlation thresholds.

Van De Vijver et al's Classification	Correlation threshold set at 0.8 (n = 64)		Correlation threshold set at 0.7 (n = 58)		Correlation threshold set at 0.5 (n = 45)	
	Cluster 1	**Cluster 2**	**Cluster 1**	**Cluster 2**	**Cluster 1**	**Cluster 2**
Good Prognosis (92 subjects)	22	70	19	73	30	62
Poor Prognosis (115 subjects)	82	33	72	43	86	29
Note: Values of n in parenthesis represents the number of gene expressions used for clustering.						

Table 2: Bivariate Table showing agreement between authors classification and clustering results at different correlation thresholds.

5 Discussion

The use of the likelihood function as a summary of the available information in a set of observed data subject to a distributional assumption is well known. Here the likelihood function is used to develop a distance matrix which can be used for clustering. The clusters of patients obtained takes into consideration the variation across the mean expression level of the genes as well as variation across level of the observed Fisher Information. The correlation threshold was set at 0.8,0.7, and 0.5. It is not surprising that as the correlation threshold was relaxed, the number of gene expression profiles decreased gradually from 64 to 58 to 45. The two clusters

differ significantly with respect to overall survival time as well as time to distant metastases for each of the three correlation thresholds.

Unlike previous authors classification methods, our clustering algorithm uses fewer number of gene expression profiles to be consistent with the assumptions in the proposed methodology. Our clustering results show clusters of patients that differed significantly with respect to overall survival time as well as time to distant metastases based on subset of the 70 gene expression profiles.

References

Arima, C., K. Hakamada, M. Okamoto and T. Hanai (2008). "Modified fuzzy gap statistic for estimating preferable number of clusters in fuzzy k-means clustering." J Biosci Bioeng **105**(3): 273-281.

Bimali, M. and M. Brimacombe (2015). "Likelihood Transformation and Information Based Approach to Clustering." Submitted.

Bouveyron, C. and C. Brunet-Saumard (2014). "Model-based clustering of high-dimensional data: A review." Computational Statistics and Data Analysis **72**: 52-78.

Covell, D. G., A. Wallqvist, A. A. Rabow and N. Thanki (2003). "Molecular classification of cancer: unsupervised self-organizing map analysis of gene expression microarray data." Mol Cancer Ther **2**(3): 317-332.

D'Haeseleer, P. (2005). "How does gene expression clustering work?" Nat Biotechnol **23**(12): 1499-1501.

Gasch, A. P. and M. B. Eisen (2002). "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering." Genome Biol **3**(11): RESEARCH0059.

Huang, D., P. Wei and W. Pan (2006). "Combining gene annotations and gene expression data in model-based clustering: weighted method." OMICS **10**(1): 28-39.

Jiang, D., C. Tang and A. Zhang (2004). "Cluster Analysis for Gene Expression Data: A Survey." IEEE Transactions on Knowledge and Data Engineering **16**(11): 1370-1386.

Nikkila, J., P. Toronen, S. Kaski, J. Venna, E. Castren and G. Wong (2002). "Analysis and visualization of gene expression data using self-organizing maps." Neural Netw **15**(8-9): 953-966.

Rencher, A. C. (2002). Methods of multivariate analysis. New York, J. Wiley.

Shahdoust, M., E. Hajizadeh, H. Mozdarani and A. Chehrei (2013). "Finding genes discriminating smokers from non-smokers by applying a growing self-organizing clustering method to large airway epithelium cell microarray data." Asian Pac J Cancer Prev **14**(1): 111-116.

Toronen, P., M. Kolehmainen, G. Wong and E. Castren (1999). "Analysis of gene expression data using self-organizing maps." FEBS Lett **451**(2): 142-146.

van 't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards and S. H. Friend (2002). "Gene expression profiling predicts clinical outcome of breast cancer." Nature **415**(6871): 530-536.

van de Vijver, M. J., Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend and R. Bernards (2002). "A gene-expression signature as a predictor of survival in breast cancer." N Engl J Med **347**(25): 1999-2009.

Yeung, K. Y., C. Fraley, A. Murua, A. E. Raftery and W. L. Ruzzo (2001). "Model-based clustering and data transformations for gene expression data." Bioinformatics **17**(10): 977-987.

Zhang, J. and L. Shen (2014). "An improved fuzzy c-means clustering algorithm based on shadowed sets and PSO." Comput Intell Neurosci **2014**: 368628.

Zhang, M., B. Adamu, C. C. Lin and P. Yang (2011). "Gene expression analysis with integrated fuzzy C-means and pathway analysis." Conf Proc IEEE Eng Med Biol Soc **2011**: 936-939.