

國立交通大學

統計學研究所

碩士論文

藉由交替 K 均值分群程序對潛在群體做預測

Prediction of Underlying Latent Classes via Alternate
K-means Clustering Algorithms



研究生：林弘哲

指導教授：黃冠華 博士

中華民國九十九年六月

藉由交替 K 均值分群程序對潛在群體做預測
Prediction of Underlying Latent Classes via Alternate
K-means Clustering Algorithms

研究生：林弘哲

Student : Hong-Jhe Lin

指導教授：黃冠華

Advisor : Dr. Guan-Hua Huang

國立交通大學



A Thesis

Submitted to Institute of Statistics
College of Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master
in
Statistics
June 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年六月

藉由交替 K 均值分群程序對潛在群體做預測

研究生：林弘哲 指導教授：黃冠華 博士

國立交通大學統計學研究所



摘要

潛在群體模型中的參數估計可以利用群體分析的方法。但是在高維度的資料下，群體分析中的變數選擇問題就更顯重要。這裡，我們發展了群體方法中的交替 k 均值分群，並利用此想法先找出干擾變數以及此模型中最佳的潛在族群體分群，再利用分群後的結果對其他參數做估計。我們並針對潛在群體模型創造一種分類規則。這個程序不但能完整的分析複雜疾病，對於基因資料分析中也有極大的幫助。我們將分析實際的資料來證實這種方法的優點。

關鍵字： 潛在群體分析、高維度資料、變數選擇、交替 k 均值、分類、微陣列

Prediction of Underlying Latent Classes via Alternate K-means Clustering Algorithms

Student: Hong-Jhe Lin

Advisor: Dr. Guan-Hua Huang

Institute of statistics

National Chiao Tung University



Abstract

Parameters in latent class analysis could be estimated by some clustering methods. But in the high-dimensional data, variable selection in cluster analysis is an important problem. Here, we propose an alternate k-means clustering method to first distinguish clustering and noisy surrogates and then estimate the parameters in the latent class model. We also create a classification rule, based on the finite mixture model. This classification procedure can explicitly recognize the heterogeneous nature of the complex disease, which makes it perfect in analyzing high-throughput genomic data. The real data analysis demonstrates the advantages of our proposed methods.

Keywords: latent class analysis; high-dimensional data; variable selection; alternate k-means; classification; microarray

誌謝

在這一年的時間裡，我非常感謝我的指導教授黃冠華老師。老師很有耐心的教導我如何去探討問題，在我有疑問的時候也都很願意幫助我解決。老師真的很厲害，不僅能仔細的教導我很多觀念，而且對於我有疑問的部份也可以詳盡的解答，如果沒有老師的幫助，我很難有今天的研究成果，真的很謝謝老師。

另外，芝賢學姐給我的建議以及告訴我老師帶學生的方法，讓我更能迅速的跟上老師的腳步。我也很感謝我的同學們，在我這段時間可以陪我放鬆心情，不受壓力影響我的研究進度。特別是尚剛在心煩的時候願意陪我聊天以及玩樂、吟玲與我一起唸書奮鬥、書維及亮勳在我心情不好的時候幫我打氣加油以及其他所有的同學對我大大小小的幫助，讓我可以順利的寫好論文以及通過口試。還有所上所有的老師，教導了我許多統計觀念以及工具，讓我在這兩年的所學非常充實。

還要感謝我的口試委員，黃冠華老師、陳鄰安老師、陳君厚老師及鄭又仁老師願意花時間看我的論文，給我許多寶貴的建議，讓我的論文更佳完整。

即將要畢業入伍，很捨不得與大家分開，感覺才剛踏入交大，轉眼就要離開了。非常開心能在這兩年獲得很多做事情的方法以及態度，也認識了很多能交心的朋友。我已經擁有了這麼多的助力，相信一定足夠面對人生的下一個階段。

最後，這兩年家人對我的支持以及幫助非常多，我才能專心的做研究。我衷心的感謝父母、老師以及同學們，也將完成論文的這份喜悅分享給大家，感謝大家。

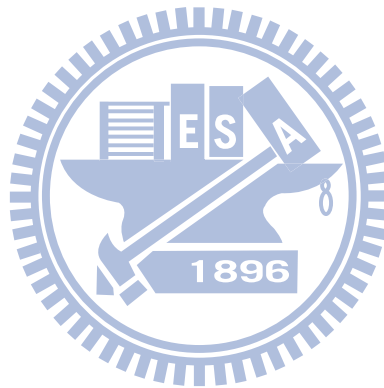


林弘哲 謹誌于
國立交通大學統計研究所
中華民國九十九年六月二十四日

Contents

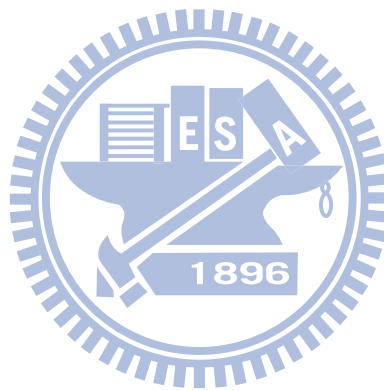
Abstract (in Chinese)	i
Abstract (in English)	ii
Acknowledgements (in Chinese)	iii
Contents	iv
List of Tables	vi
List of Figures	vii
1. Introduction	錯誤! 尚未定義書籤。
2. Literature review	3
2.1 Latent class analysis (LCA)	3
2.2 Regression extension of latent class analysis (RLCA)	4
2.3 Marginalization of the regression extension of latent class model	6
2.3.1 Marginalizing the covariate effects on conditional probabilities	6
2.3.2 Marginalizing the covariate effects on latent prevalences	8
2.4 K-means method	8
2.5 Lasso regression	9
2.6 Panelized Model-Based Clustering	10
2.7 Sparse K-means method	11
3. Models	13
4. Parameter estimations by clustering algorithm	16
4.1 Latent Class Membership Estimation When Not Incorporating Covariate Effects	16
4.1.1 The Measurement For Complete The Assumption A1	17
4.1.2 The Measurement For Complete The Assumption A2	18
4.1.3 The Measurement For Our Alternate K-means Algorithm	20

4.1.4	The alternate K-means Algorithm	21
4.1.5	Estimation of tuning parameter λ	22
4.2	Latent Class Membership Estimation When Incorporating Covariate Effects	22
5.	Classification using finite mixture models	25
6.	Example	27
6.1	Breast cancer data	27
6.2	Schizophrenia Syndrome Scale Data	29
7.	Discussion	32
	References	33



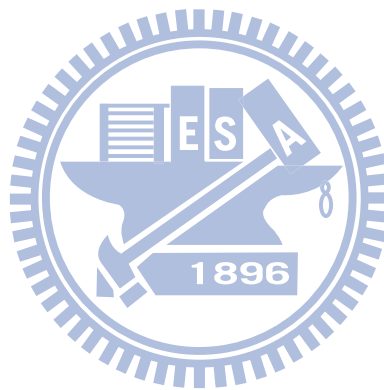
List of Tables

Table 1: Predictions of class membership of 19 tumors by standard k-means clustering method.....	36
Table 2: Predictions of class membership of 19 tumors by alternate k-means clustering method.....	37
Table 3: Predictions of class membership of 10 schizophrenia patient by standard k-means clustering method.....	38
Table 4: Predictions of class membership of 10 schizophrenia patient by alternate k-means clustering method.....	38



List of Figures

Figure 1: Lasso and Ridge regression.....	39
Figure 2: The flow chart of alternate k-means.....	40
Figure 3: The flow chart of standard k-means.....	41
Figure 4: The heatmap for the 200-gene (original).....	42
Figure 5: The heatmap for the 154-gene (selected).....	43
Figure 6: The heatmap for the 30-item (original).....	44
Figure 7: The heatmap for the 18-item (selected).....	45



1 Introduction

Many concepts in medical research are unobservable, hence valid surrogates must be measured in place of these concepts. Models that permit exploration of relationships between unobservable variables and their surrogates are referred to as latent variable models. This unobservable characteristic can be expressed as an univariate continuous score (the latent trait) (Rasch, 1960; Lazarsfeld & Henry, 1968; Moustaki, 1996) or a categorical variable identifying several "classes" that define homogeneous groups of individuals (Goodman, 1974; Titterton, Smith, & Makov, 1985; Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997; Huang & Bandeen-Roche, 2004). In this article, we focus on the cases where an underlying categorical variable (the latent class variable) is used, and thus measured surrogates are independent of one another within any category of the latent variable. These are commonly called as finite mixture models.

Parameters in finite mixture models are typically estimated by the maximum likelihood (ML) for a fixed number of classes. The Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) can be used to compute the ML estimates of parameters. However, the above likelihood function is derived under the assumptions that cannot be checked directly and may critically affect analytic findings (Bandeen-Roche et al., 1997). Also, implementing the EM algorithm to estimate parameters in finite mixture models is typically time-consuming and can be difficult to converge when possible patterns of surrogates are large and the sample size is moderate or small. To avoid the problem of EM algorithm, the finite mixture analysis may legitimately be viewed as the analog of cluster analysis (Huang, Wang & Hsu, manuscript).

Cluster analysis implements a number of different algorithms and methods for grouping objects of similar kind (i.e., with close distance) into respective categories. But in many cases, the number of surrogates is too large, and this may be due to the presence of several "noisy"

noninformative surrogates. The noisy surrogates will influence the clustering process and let the result be fault. In this article, we will focus on how to solve this problem. We develop the alternate k-means clustering methods. This clustering method can select noisy surrogates and calculate the correlation among other clustering surrogates as the distance measure to group objects at the same time. For the selected objects that belong to the same latent class, their clustering surrogates are as close to be independent as possible. By treating the estimated class membership based on the clustering surrogates as a known predictor, it becomes easy to estimate the parameters in the finite mixture model. The proposed method can reduce the dimension of data (i.e., to select surrogates we need), and directly obtain estimated latent classes that best describe the association among surrogates.

This article also used a classification rule for predicting new objects' outcome statuses (e.g., diseased/not diseased, cured/not cured), based on the finite mixture model (Huang, Wang & Hsu, manuscript). The classification rule is especially useful for high-dimensional data (e.g., microarray data). The latent class model explicitly recognizes and hence mitigates errors in measurement, and gives a well summary of measured surrogates (i.e., the latent variable). With our proposed parameter estimating procedure, we can easily perform the finite mixture models on high-dimensional data and thus create classification rules based on the inferred latent classes.

The rest of the article is organized as follows. In section 2, we introduce the latent class analysis (LCA), the regression extension of latent class analysis (RLCA) model, and some clustering methods. Section 3 and 4 provide a model with a new idea and detail the clustering algorithm for estimating the latent classes underlying a finite mixture and the parameters of the model. A classification rule based on the finite mixture model is then described in section 5. Gene expression microarray and schizophrenia syndrome scale data are used to illustrate the proposed methods in section 6. Final, we conclude by discussing the contribution of this article.

2. Literature Review

2.1 Latent class analysis (LCA)

Goodman (1974) provided an excellent overview of the Latent class analysis (LCA) model, including a maximum likelihood strategy for estimating model parameters, conditions to determine local model identifiability, a strategy to test overall model fit, and the use of constraints to identify models. The idea for this model is that all measured surrogates reflect the same unobservable characteristic, and that this characteristic fully explains the associations between observed surrogates. LCA aims to classify objects based on their responses to a set of categorical items. Here, we let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iM})^T$ denote a set of M observable polytomous surrogates for the i th individual in a study sample of N samples Y_{im} , $m = 1, \dots, M$ can take values $\{1, \dots, K_m\}$, where $K_m \geq 2$. The basic model postulates an underlying categorical latent variable $S_i = 1, \dots, J$ for individual i ; within any category of the latent variable, the measured indicators are assumed to be independent of one another. Therefore, the distribution for \mathbf{Y}_i can be expressed as

$$\Pr(Y_{i1} = y_1, \dots, Y_{iM} = y_M) = \sum_{j=1}^J \left\{ \eta_j \prod_{m=1}^M \prod_{k=1}^{K_m} p_{mkj}^{y_{mk}} \right\}, \quad (2.1)$$

where $y_{mk} = I(y_m = k) = 1$ if $y_m = k$; 0 otherwise; and assumes that $\eta_j = \Pr(S_i = j)$ and

$$p_{mkj} = \Pr(Y_{im} = k | S_i = j), \quad (2.2)$$

$$i = 1, \dots, N; m = 1, \dots, M; k = 1, \dots, K_m; j = 1, \dots, J.$$

The model treats class membership probabilities, η_j , and item response probabilities conditional on class membership, p_{mkj} , as homogeneous over individuals. Heuristically, η_j

is the population prevalence of class j , and p_{mkj} is the probability of an individual in class j being at levels k of Y_{im} .

2.2 Regression extension of latent class analysis (RLCA)

Huang and Bandeen-Roche (2004) extend the latent class analysis to allow both the probabilities of latent class membership and the distribution of observed responses given latent class membership to be functionally related to concomitant variables, while preserving model identifiability. By allowing covariate effects on latent class probabilities, we summarize the effect of risk factors on the underlying mechanism. In the case of incorporation covariates into conditional probabilities, we can adjust for characteristics that determine responses other than underlying classes, hence hopefully improving the accuracy of classifying individuals. For instance, in evaluating functional disability, some data have suggested that women tend to rate tasks as “difficult” more readily than men independently of ability (Bandeen-Roche, Huang, Munoz, & Rubin, 1999). Without adjusting for the gender effect, the model might well classify some men and women with identical underlying functioning differently (men as “able”, women as “disabled”). In the literature, they also provided an excellent overview of the RLCA model, including model identification, Expectation-Maximization algorithm for parameter estimation, standard error calculation, convergent properties, and comparison of the RLCA model with models underlying existing latent class modeling software.

In RLCA models, let $(\mathbf{x}_i, \mathbf{z}_i)$ be the concomitant covariates of the i th sample. $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$ are defined as primary covariate hypothesized to be associated with latent class membership, S_i , and $\mathbf{z}_i = (z_{i1}, \dots, z_{im})^T$ with $\mathbf{z}_{im} = (1, z_{im1}, \dots, x_{imL})^T$, $m = 1, \dots, M$, are secondary covariates used to build direct effects on measured surrogates. The sets of covariates can include any combination of continuous and discrete measures, and two sets of

covariates can be mutually exclusive or overlap. The regression extension of LCA may then be stated as follows:

$$\Pr(Y_{i1} = y_1, \dots, Y_{iM} = y_M | \mathbf{x}_i, \mathbf{z}_i) = \sum_{j=1}^J \left\{ \eta_j(\mathbf{x}_i^T \boldsymbol{\beta}) \prod_{m=1}^M \prod_{k=1}^{K_m} p_{mkj}^{y_{mk}}(\boldsymbol{\gamma}_{mj} + \mathbf{z}_{im}^T \boldsymbol{\alpha}_m) \right\} \quad (2.3)$$

with $\eta_j(\mathbf{x}_i^T \boldsymbol{\beta})$ and $p_{mkj}^{y_{mk}}(\boldsymbol{\gamma}_{mj} + \mathbf{z}_{im}^T \boldsymbol{\alpha}_m)$ defined as in the generalized linear framework (McCullagh and Nelder, 1989). Usually, (2.3) is implemented assuming generalized logit (Agresti, 1984) link functions:

$$\log \left[\frac{\eta_j(\mathbf{x}_i^T \boldsymbol{\beta})}{\eta_{j'}(\mathbf{x}_i^T \boldsymbol{\beta})} \right] = \beta_{0j} + \beta_{1j} x_{i1} + \dots + \beta_{pj} x_{ip} \quad \text{for } i = 1, \dots, N; j = 1, \dots, J-1 \quad (2.4)$$

and

$$\log \left[\frac{p_{mkj}(\boldsymbol{\gamma}_{mj} + \mathbf{z}_{im}^T \boldsymbol{\alpha}_m)}{p_{mKj}(\boldsymbol{\gamma}_{mj} + \mathbf{z}_{im}^T \boldsymbol{\alpha}_m)} \right] = \gamma_{mkj} + \alpha_{1mk} z_{im1} + \dots + \alpha_{Lmk} z_{imL} \quad (2.5)$$

for $i = 1, \dots, N; m = 1, \dots, M; k = 1, \dots, (K_m - 1); j = 1, \dots, J$.

The model is also necessary to unambiguously distinguish covariate effects on measured response probabilities from covariate effects on class probabilities. Three assumptions complete (2.3):

$$(C1) \quad \Pr(Y_{i1} = y_1, \dots, Y_{iM} = y_M | S_i, \mathbf{x}_i, \mathbf{z}_i) = \Pr(Y_{i1} = y_1, \dots, Y_{iM} = y_M | S_i, \mathbf{z}_i);$$

$$(C2) \quad \Pr(S_i = j | \mathbf{x}_i, \mathbf{z}_i) = \Pr(S_i = j | \mathbf{x}_i);$$

$$(C3) \quad \Pr(Y_{i1} = y_1, \dots, Y_{iM} = y_M | S_i, \mathbf{z}_i) = \prod_{m=1}^M \Pr(Y_{im} = y_m | S_i, \mathbf{z}_{im}).$$

Notice that, in the conditional probability of model (2.5), we allow unrestricted intercepts and level- and item-specific covariate coefficients, but the coefficients varying across classes

is unallowable (i.e., α_{qmk} is dependent on m, k but independent of j). This constraint is reasonable if the primary purpose of modeling conditional probabilities is to prevent possible misclassification by adjusting for characteristics associated with item measurements.

2.3 Marginalization of the regression extension of latent class model

We introduce a process to “eliminate” the covariates effect, hence “marginalize” the RLCA model (2.3). The marginalization process (Huang 2005) includes two stages. Stage 1 aims to eliminate \mathbf{z}_i effect. At stage 2, we apply the marginalization property; proposed by Bandeen-Roche et al. (1997), to average \mathbf{x}_i effect out of the latent prevalence.

2.3.1 Marginalizing the covariate effects on conditional probabilities

For achieving the RLCA model assumption (C3), we need to eliminate the covariate effect. The key to marginalizing over \mathbf{z}_i is that the process must yield random variables that follow a finite mixture distribution that is both independent of \mathbf{z}_i and has J mixing components. A method for achieving such marginalization can be motivated by the properties of added variable plots for linear regression models.

Consider the linear model

$$\mathbf{Y} = \mathbf{x}_1^T \boldsymbol{\beta}_1 + \mathbf{x}_2^T \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \quad (2.6)$$

where $\boldsymbol{\varepsilon}$ with mean $\mathbf{0}$ and variance matrix \mathbf{V} . Let $\tilde{\mathbf{Y}}$ denote the residuals of regressing \mathbf{Y} on \mathbf{x}_2 , and $\mathbf{W} = \mathbf{V}^{-1}$ be the weight matrix. Then, it is well known that if \mathbf{x}_1 and \mathbf{x}_2 are orthogonal (i.e., $\mathbf{x}_1^T \mathbf{W} \mathbf{x}_2 = 0$), $\tilde{\mathbf{Y}}$ has mean $\mathbf{x}_1^T \boldsymbol{\beta}_1$ and variance \mathbf{V} . That is, the simple linear regression of $\tilde{\mathbf{Y}}$ on \mathbf{x}_1 yields exactly the same inferences about $\boldsymbol{\beta}_1$ as if we performed the analysis on the more complicated model (2.6) (Cook and Weisberg, 1982). Now, viewing the just-described stability of $\boldsymbol{\beta}_1$ as analogous to the desired stability of latent

class dimension, J , the added variable property can be applied to model (2.6) to obtain the marginalized conditional probabilities.

To present the key ideas more clearly, we let the measured indicators (Y_{i1}, \dots, Y_{iM}) are assumed to be binary (i.e., $K_1 = \dots = K_M = 2$). Notice that (2.5) can be viewed as fitting a logistic regression of Y_{im} on S_i adjusting for \mathbf{z}_{im} , separately for each m . Now, we make the analogy to (2.6), let $S_{ij} = I(S_i = j)$ for $i = 1, \dots, N$; $j = 1, \dots, J - 1$. We can reparameterize (2.5) as

$$\text{logit} [E(Y_{im} | \mathbf{S}_i, \mathbf{Z}_{im}^c)] = \mathbf{S}_i^T \boldsymbol{\gamma}_m + (\mathbf{Z}_{im}^c)^T \boldsymbol{\alpha}_m \quad \text{for } i = 1, \dots, N ; m = 1, \dots, M \quad (2.7)$$

where $\mathbf{S}_i = [1, S_{i1}, \dots, S_{i(J-1)}]^T$;

$\mathbf{Z}_{im}^c = [(z_{im1} - \bar{z}_{m1}), \dots, (z_{imL} - \bar{z}_{mL})]^T$, (“centered” covariate vector);

$\bar{z}_{mp} = (1/N) \sum_{i=1}^N z_{imp}$;

$\boldsymbol{\gamma}_m = [\gamma_{m0}, \gamma_{m1}, \dots, \gamma_{m(J-1)}]^T$; and $\boldsymbol{\alpha}_m = [\alpha_{1m}, \alpha_{2m}, \dots, \alpha_{Lm}]^T$.

Therefore, for any realization of \mathbf{S}_i , (2.7) is a logistic regression with dependent variable: Y_{im} and predictors: $\mathbf{S}_i, \mathbf{Z}_{im}^c$.

Next, the problem becomes how to calculate residuals form the generalized linear model

$$\text{logit} [E(Y_{im} | \mathbf{S}_i, \mathbf{Z}_{im}^c)] = (\mathbf{Z}_{im}^c)^T \boldsymbol{\alpha}_m^* \quad \text{for } i = 1, \dots, N ; m = 1, \dots, M \quad (2.8)$$

The “pseudo-residuals” are given by

$$\mathbf{R}_m = [R_{1m}, \dots, R_{Nm}]^T = \hat{\mathbf{V}}_m^{-1} (\mathbf{Y}_m - \hat{\boldsymbol{\mu}}_m). \quad (2.9)$$

Here “hat” represents the estimated values;

$$\mathbf{Y}_m = [Y_{1m}, \dots, Y_{Nm}]^T; \mathbf{V}_m = \text{diag}(V_{1m}, \dots, V_{Nm}); V_{im} = \text{Var}(Y_{im}); \mathbf{Z}_m^c = [\mathbf{Z}_{1m}^c, \dots, \mathbf{Z}_{Nm}^c]$$

If \mathbf{x}_i and \mathbf{z}_{im} are independent, we can extract the \mathbf{Z}_{im}^c from conditional probabilities by treating the residuals from the model (2.8) as new response variables and regressing them on \mathbf{S}_i . We substitute the estimate of γ_m^* in the linear model

$$R_{im} = \mathbf{S}_i^T \gamma_m^* + \varepsilon_{im}, \quad i = 1, \dots, N; m = 1, \dots, M. \quad (2.10)$$

For the estimate of γ_m in the model (2.7), a formal justification shows that γ_m^* and γ_m can be very close under reasonable regularities. The above results can be extended to the cases where (Y_{i1}, \dots, Y_{iM}) is polytomous as in (2.1) and (2.3).

2.3.2 Marginalizing the covariate effects on latent prevalences

For the marginalization of model (2.3) over \mathbf{x}_i , we use the nice property of the RLCA model that the covariates associated with latent class prevalences, \mathbf{x}_i , can be ignored.

2.4 K-means method

MacQueen (1967) suggests the term K-means for describing an algorithm that assigns each item to the cluster having the nearest centroid (mean). In its simplest version, the process is composed of these three steps:

1. Partition the items into K initial clusters.
2. Proceed through the list of items, assigning an item to the cluster whose centroid (mean) is nearest. (Distance is usually computed using the Euclidean distance with either standardized or unstandardized observations.) Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.
3. Repeat Step 2 until no more reassignments take place.

Rather than starting with a partition of all items into K preliminary groups in Step 1, we could specify K initial centroids (seed points) and then proceed to Step 2.

The final assignment of items to clusters will be, to some extent, dependent upon the initial partition or the initial selection of seed points. Experience suggests that most major changes in assignment occur with the first reallocation step.

2.5 Lasso regression

When the data have high dimension or high correlation, variable selection is very important. Tibshirani (1996) proposed Lasso regression. It is popular to solve this problem. Unlike the original least square method, Lasso uses the penalized least square, which can avoid high correlation problem and estimate some parameter to 0 at the same time. Frank and Friedman (1993) propose the Bridge regression, and Lasso is the special case. The Bridge regression is based on least square, and limit the parameter by $\sum |\beta_j|^r \leq t (t > 0)$. The parameter estimated can be

$$\hat{\beta}^{Bridge}(\lambda) = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^p |\beta_j|^r \right\} \quad (2.11)$$

where $r \geq 0$ (if $r < 0$, the penalize function will be concave function, that is, $\hat{\beta}^{Bridge}$ has no minimum value), and λ is the tuning parameter. When $r=1$, we called this Lasso regression. And $r=2$, we called Ridge regression.

We compare the difference of Lasso regression and Ridge regression. Assume there are two parameters in the model. Note that,

$$\hat{\beta}^{Lasso}(\lambda) = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (2.12)$$

$$\hat{\beta}^{\text{Ridge}}(\lambda) = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^p |\beta_j|^2 \right\} \quad (2.13)$$

In the figure 1, the limited region of Lasso and Ridge regression will be blue square and blue circle. The least square method with no constraint will be red ellipse, and the center is the solution of least square estimate. The advantage of Lasso regression is the limited region has corners. If ellipse touches the square at the corner, it means Lasso regression will estimate some parameters to 0. That is, Lasso can select variable and estimate parameters at the same time.

2.6 Penalized model-based clustering

Variable selection in clustering analysis is both challenging and important. Pan and Shen (2007) propose a penalized likelihood approach with an L_1 penalty function, automatically realizing variable selection via thresholding and delivering a sparse solution. For the original model-based clustering method, given the observation x is drawn from a finite mixture distribution $f(x; \Theta) = \sum_{k=1}^K \pi_k f_k(x; \theta_k)$, with the proportion π_k , component-specific distribution f_k and its parameters θ_k . The log-likelihood is

$$\log L(\Theta) = \sum_{j=1}^N \log \left[\sum_{k=1}^K \pi_k f_k(x_j; \theta_k) \right]. \quad (2.11)$$

With the same motivation as in penalize regression, they propose a penalized likelihood:

$$\log L(\Theta) = \sum_{j=1}^N \log \left[\sum_{k=1}^K \pi_k f_k(x_j; \theta_k) \right] - h_{\lambda}(\Theta), \quad (2.12)$$

where $h_{\lambda}(\cdot)$ is a penalty function with penalization parameter λ . The choice of $h_{\lambda}(\cdot)$ depend on the goal of the analysis. The EM algorithm can be applied to obtain the maximum likelihood estimator of Θ . The K-means algorithm can be used in this process, and find the

variable cluster. More importantly, this process can select variable automatic.

2.7 Sparse k-means method

The standard k-means method can assign each item to the cluster with the features, and there usually are large set of features. We might expect that the true underlying clusters present in the data differ only with respect to a small fraction of the features, and will be missed if one clusters the observations using the full set of features. Witten and Tibshirani (2010) propose a method for sparse clustering, which allows us to group the observations using only an adaptively-chosen subset of the features. Suppose we want to cluster n observations on p dimensions. Let $X_j \in \mathfrak{R}^n$ denote feature j . Many clustering methods can be expressed as optimizing criteria of the form

$$\underset{\Theta \in D}{\text{maximize}} \left\{ \sum_{j=1}^p f_j(X_j, \Theta) \right\} \quad (2.13)$$

where $f_j(X_j, \Theta)$ is some function that involves only the j th feature of the data, and Θ is a parameter restricted to lie in a set D . Then, they define sparse clustering as the solution to the problem

$$\underset{w: \Theta \in D}{\text{maximize}} \left\{ \sum_{j=1}^p w_j f_j(X_j, \Theta) \right\} \text{ subject to } \|w\|^2 \leq 1, \|w\|_1 \leq s, w_j \geq 0 \forall j \quad (2.14)$$

where w_j is the weight corresponding to feature j .

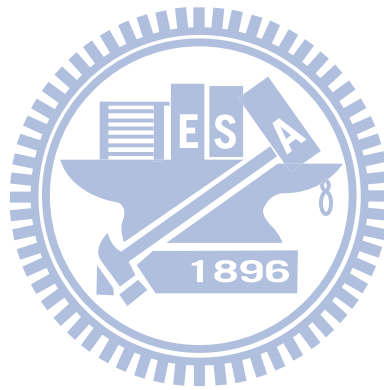
We optimize (2.14) using an alternate algorithm: holding w fixed, we optimize (2.14) with respect to Θ , and holding Θ fixed, we (2.14) with respect to w . In general, we do not achieve a global optimum of (2.14) using this alternate approach; however, we are guaranteed that each iteration increases the objective function. Notice that, to optimize (2.14) with respect

to w with Θ held fixed, we note that the problem can be re-written as

$$\underset{w: \Theta \in D}{\text{maximize}} \{w^T a\} \text{ subject to } \|w\|^2 \leq 1, \|w\|_1 \leq s, w_j \geq 0 \forall j$$

(2.15)

where $a_j = f_j(X_j, \Theta)$. Details are referred to Tibshirani (2010).



3 Model

Let (Y_{i1}, \dots, Y_{iM}) denote a set of M observable surrogates and S_i denote the unobservable class membership, for the i th individual in a study sample of N samples. Unlike traditional LCA model, we think that some surrogates have no difference among unobservable latent classes. We call these surrogates as “noisy surrogates”. The other surrogates that have different distributions in different latent classes are called “clustering surrogates”. We hope to find the noisy surrogates and exclude their influences in estimating latent classes. So, under this idea, we let

$\underline{Y} = (Y_{i1}, \dots, Y_{iM}) = (\underline{Y}_i^{(1)}, \underline{Y}_i^{(2)}) = (Y_{i1}^{(1)}, \dots, Y_{iM_1}^{(1)}, Y_{i1}^{(2)}, \dots, Y_{iM_2}^{(2)})$, where $\underline{Y}_i^{(1)}$ denote the noisy surrogates, $\underline{Y}_i^{(2)}$ denote the clustering surrogates, and $M_1 + M_2 = M$. Y_{im} can be either continuous, ordinal or categorical, for $m=1, \dots, M$, and S_i can take values $\{1, \dots, J\}$. The distribution for (Y_{i1}, \dots, Y_{iM}) can be expressed as the finite mixture density:

$$\begin{aligned}
 \Pr(\underline{Y}) &= \sum_{j=1}^J \Pr(S_i = j) \Pr(\underline{Y} | S_i = j) \\
 &= \sum_{j=1}^J \Pr(S_i = j) \Pr(\underline{Y}^{(1)} | \underline{Y}^{(2)}, S_i = j) \Pr(\underline{Y}^{(2)} | S_i = j) \\
 &\stackrel{\text{assumption 1}}{=} \Pr(\underline{Y}^{(1)} | \underline{Y}^{(2)}) \sum_{j=1}^J \Pr(S_i = j) \Pr(\underline{Y}^{(2)} | S_i = j) \\
 &\stackrel{\text{assumption 2}}{=} \Pr(\underline{Y}^{(1)} | \underline{Y}^{(2)}) \times \sum_{j=1}^J \left\{ \eta_j \prod_{m=1}^{M_2} \prod_{k=1}^{K_m} p_{mkj}^{y_m^{(2)}} \right\}
 \end{aligned}$$

(3.1)

where $p_{mkj} = \Pr(Y_{im}^{(2)} = k | S_i = j)$ are the “conditional probabilities” of the measured responses given the underlying variable category, $\eta_j = \Pr(S_i = j)$ are the “latent class probabilities” of each underlying variable category and $y_{mk}^{(2)} = 1$ if $y_m^{(2)} = k$; 0 otherwise. This

finite mixture model will be completed by two assumptions:

$$(A1) \Pr(Y_{i1}^{(2)}, \dots, Y_{iM_2}^{(2)} | S) = \prod_{m=1}^{M_2} \Pr(Y_{im}^{(2)} | S);$$

$$(A2) \Pr(Y_i^{(1)} | Y_i^{(2)}, S_i) = \Pr(Y_i^{(1)} | Y_i^{(2)});$$

Heuristically, η_j is the population prevalence of class j , and p_{mkj} is the probability of an individual in class j being at levels k of $Y_{im}^{(2)}$, and we do not explore the influence of $Y_{im}^{(1)}$ in the following article.

Some authors have extended the finite mixture model to describe the effects of measured covariates on the underlying mechanism and/or on measured surrogate distributions within latent levels. One can summarize the effect of risk factors on the underlying mechanism by allowing covariates $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$ to be functionally related to latent class S_i (Dayton & Macready, 1998; Bandeen-Roche et al., 1997; Huang & Bandeen-Roche, 2004). And we implement the generalized linear framework (McCullagh & Nelder, 1989) to incorporate covariate effects into S_i :

$$\log \left[\frac{\eta_j(\mathbf{x}_i)}{\eta_J(\mathbf{x}_i)} \right] = \beta_{0j} + \beta_{1j}x_{i1} + \dots + \beta_{pj}x_{ip} \quad \text{for } i = 1, \dots, N; j = 1, \dots, J-1, \quad (3.2)$$

To adjust for characteristics associated with surrogates, hence prevent possible misclassification of underlying variable categories, we can incorporate individual-level independent variables into the within-class distributions of measured surrogates (Melton, Liang, & Pulver, 1994; Huang, & Bandeen-Roche, 2004; Muthen, & Muthen, 2007).

Let $\mathbf{z}_i = (z_{i1}, \dots, z_{iM})^T$ with $\mathbf{z}_{im} = (1, z_{im1}, \dots, x_{imL})^T$, $m = 1, \dots, M$ be covariates used to build direct effects on measured surrogates within latent classes for the i th individual. When

surrogates are ordinal or categorical variables, we assume that

$(Y_{im}^{(2)} | S_i = j, z_{im}) \sim \text{Multinomial}(1; p_{m1j}(z_{im}), \dots, p_{mK_m j}(z_{im}))$, and

$$\log \left(\frac{p_{mkj}(z_{im})}{p_{mK_m j}(z_{im})} \right) = \gamma_{mkj} + \alpha_{1mk} z_{im1} + \dots + \alpha_{Lmk} z_{imL} \quad , \quad (3.3)$$

where $k = 1, \dots, (K_m - 1)$, $K_m \geq 2$, and $m = 1, \dots, M$

If surrogates are continuous variables, we assume that

$(Y_{im}^{(2)} | S_i = j, z_{im}) \sim \text{Normal}(\mu_{mj}(z_{im}), \sigma_m^2)$, and

$$\mu_{mj}(z_{im}) = \theta_{mj} + \tau_{1m} z_{im1} + \dots + \tau_{Lm} z_{imL} . \quad (3.4)$$

By incorporating (x_i, z_i) , we relax the homogeneous probability (i.e., p_{mkj} in the model (3.1)) in the sense that the probabilities vary with some individual characteristics. In the conditional distribution models (3.3) and (3.4), we allow unrestricted intercepts, but we do not allow the covariate coefficients to vary across classes (i.e., σ_m^2 , τ_{lm} , and α_{lmk} , $l = 1, \dots, L$ are independent of j). This constraint is logical if the primary purpose of modeling conditional probabilities is to prevent possible misclassification by adjusting for characteristics associated with surrogates. In addition, after adjusting covariate effects, the conditional independence assumption is also conditioning on z_i , that is

$$\Pr(Y_{i1}^{(2)}, \dots, Y_{iM_2}^{(2)} | S_i, z_i) = \prod_{m=1}^{M_2} \Pr(Y_{im}^{(2)} | S_i, z_{im}) . \quad (3.5)$$

4 Parameter Estimation by Clustering Algorithm

Parameters in RLCA model are typically estimated using the EM algorithm (Goodman, 1974; Bandeen-Roche et al., 1997; Huang & Bandeen-Roche, 2004). Although this method has notable advantages (e.g., obtaining consistent and asymptotically normally distributed estimations, and directly providing standard error estimates for parameters), it can be vulnerable to the violation of model assumptions and be difficult to converge when fitting models with large numbers of surrogates and/or latent classes. Here, we propose an alternative strategy for estimating parameters. The proposed method consists of two stages: first, the alternate k-means method used in cluster analysis can find some noisy surrogates and implemented to estimate the underlying latent class membership. Second, the estimated class membership is treated as a known variable and other parameters are then estimated.

4.1 Latent class membership estimation when not incorporating covariate effects

Finite mixture analysis is a useful tool to classify objects based on their responses to a set of surrogates. The basic model postulates an underlying categorical latent variable $S_i \in \{1, \dots, J\}$, and, within any category of the latent variable, measured clustering surrogates are assumed to be independent of one another, and noisy surrogates are assumed to be no difference in each class when given clustering surrogates. But when we want to control more than one assumption, the traditional k-means algorithm will fail to work. So, we proposed the alternate k-means clustering method and to estimate S_i by applying this method to find noisy surrogates, and to group the objects into J subgroups such that objects in one subgroup will have a set of statistically independent clustering surrogates. Unlike the traditional EM approach that intends to derive the grouping of objects under the assumption, the proposed method tries to find the “optimal” grouping that is the most satisfying of the assumption.

4.1.1 The measurement for complete the assumption A1

The assumption (A1) means the clustering surrogates should be independent when they in the same latent class. So, we just to use clustering surrogates $\underline{Y}^{(2)}$ to calculate the sample covariance matrix. For continuous surrogates, the sample covariance matrix is a $M_2 \times M_2$ matrix with component (m, t), being the sample variance between $\underline{Y}_{im}^{(2)}$ and $\underline{Y}_{it}^{(2)}$. For polytomous categorical surrogates, each component of $(\underline{Y}_{i1}^{(2)}, \dots, \underline{Y}_{iM_2}^{(2)})$ is represented as a vector with elements being the indicators of each category:

$$\tilde{\underline{Y}}_i^{(2)} = (\tilde{Y}_{i1}^{(2)}, \dots, \tilde{Y}_{iM_2}^{(2)}) = (Y_{i11}^{(2)}, \dots, Y_{i1(K_1-1)}^{(2)}, \dots, Y_{iM_21}^{(2)}, \dots, Y_{iM_2(K_{M_2}-1)}^{(2)}) \quad (4.1)$$

with $Y_{imk}^{(2)} = I(Y_{im}^{(2)} = k) = 1$ if $Y_{im}^{(2)} = k$, 0 otherwise; $m = 1, \dots, M_2$; $k = 1, \dots, (K_m - 1)$. Then,

$$Cov(\tilde{\underline{Y}}_i^{(2)}) = \left\{ Cov(Y_{imk}^{(2)}, Y_{its}^{(2)}) \right\} = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1M_2} \\ B_{21} & B_{22} & \cdots & B_{2M_2} \\ \vdots & \vdots & \ddots & \vdots \\ B_{M_21} & B_{M_22} & \cdots & B_{M_2M_2} \end{bmatrix}, \quad (4.2)$$

where $B_{mt} = Cov(\tilde{Y}_{im}^{(2)}, \tilde{Y}_{it}^{(2)})$ is a $(K_m - 1) \times (K_t - 1)$ block matrix. Various component of the above covariance matrix are

$$Cov(Y_{imk}^{(2)}, Y_{its}^{(2)}) = \begin{cases} \Pr(Y_{imk}^{(2)} = 1) - \Pr(Y_{imk}^{(2)} = 1)\Pr(Y_{its}^{(2)} = 1) & \text{if } m = t, k = s \\ -\Pr(Y_{imk}^{(2)} = 1)\Pr(Y_{its}^{(2)} = 1) & \text{if } m = t, k \neq s \\ \Pr(Y_{imk}^{(2)} = 1, Y_{its}^{(2)} = 1) - \Pr(Y_{imk}^{(2)} = 1)\Pr(Y_{its}^{(2)} = 1) & \text{if } m \neq t, k \neq s \end{cases} \quad (4.3)$$

The sample covariance matrix is obtained by replacing the probabilities with the sample

averages. Let $ACov_j$ be the average of absolute values of entries in off diagonal elements (continuous surrogates) / blocks (polytomous surrogates) of the sample covariance matrix using objects in class j . Then, we define the “loss of independence” as

$$\text{LoI} = \sum_{j=1}^J w_j ACov_j \text{ with } w_j = \frac{\text{the number of objects in class } j}{N}. \quad (4.4)$$

Notice that, the LoI is the weighted average of $ACov_j$ over all classes with weights proportional to numbers of objects in each class. LoI can be used as the measure for evaluating assumption (A1). The smaller the value of LoI, the more satisfying the assumption (A1).

4.1.2 The measurement for complete the assumption A2

The assumption (A2) means the conditional expectations in any group should be equal, that is, $E(Y_i^{(1)} | Y_i^{(2)}, S_i = 1) = \dots = E(Y_i^{(1)} | Y_i^{(2)}, S_i = J) = E(Y_i^{(1)} | Y_i^{(2)})$, and we use a non-parametric method to evaluate the conditional expectation. In order to complete our algorithm, we need to create a measurement, which is called the between class variation.

For the continuous surrogates, using the “nearest neighbor” approach, we define

$(Y_{im}^{(1)} | Y_i^{(2)}) = Y_{im}^*$ and estimate Y_{im}^* by

$$Y_{im}^* = \frac{1}{r} \sum_{k \in C_i} Y_{km}^{(1)} \quad (4.5)$$

where $i = 1, \dots, N$, $m = 1, \dots, M_{K_1}$, and C_i be the set of indices of the r nearest neighbors of $Y_i^{(2)}$ among $\{Y_1^{(2)}, \dots, Y_N^{(2)}\}$. Here we define the “distance” between $Y_s^{(2)}$ and $Y_r^{(2)}$ by

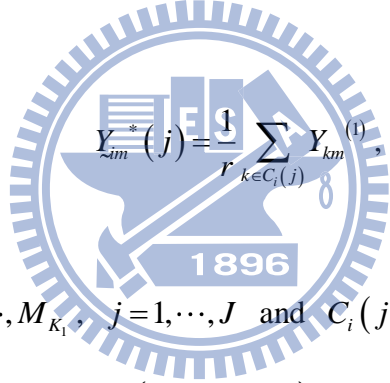
$\|Y_{\tilde{s}}^{(2)} - Y_{\tilde{t}}^{(2)}\|_2^2$. Then, we fixed r and take

$$\begin{aligned}
C_i = \{ & (i_1, \dots, i_r) \subset (1, 2, \dots, N) | \\
& i_1 = \arg \min_{k \in \{1, \dots, N\} \setminus \{i\}} \|Y_{\tilde{k}}^{(2)} - Y_{\tilde{i}}^{(2)}\|_2^2, \\
& i_2 = \arg \min_{k \in \{1, \dots, N\} \setminus \{i, i_1\}} \|Y_{\tilde{k}}^{(2)} - Y_{\tilde{i}}^{(2)}\|_2^2, \\
& \vdots \\
& i_r = \arg \min_{k \in \{1, \dots, N\} \setminus \{i, i_1, \dots, i_{r-1}\}} \|Y_{\tilde{k}}^{(2)} - Y_{\tilde{i}}^{(2)}\|_2^2 \}
\end{aligned} \tag{4.6}$$

Let $\bar{Y}_m^* = \frac{1}{N} \sum_{i=1}^N Y_{im}^*$, we can have the over all conditional expectation mean

$\bar{Y}_{\tilde{z}}^* = (\bar{Y}_1^*, \dots, \bar{Y}_{M_1}^*)^t$. On the other hand, we define $(Y_{\tilde{im}}^{(1)} | Y_{\tilde{i}}^{(2)}, S_i = j) = Y_{im}^*(j)$ and estimate

$Y_{im}^*(j)$ by

$$Y_{im}^*(j) = \frac{1}{r} \sum_{k \in C_i(j)} Y_{km}^{(1)}, \tag{4.7}$$


where $i=1, \dots, N$, $m=1, \dots, M_{K_1}$, $j=1, \dots, J$ and $C_i(j)$ be the set of indices of the r

nearest neighbors of $Y_{\tilde{i}}^{(2)}$ among $\{Y_{\tilde{j}_1}^{(2)}, \dots, Y_{\tilde{j}_{N_j}}^{(2)}\}$, where $S_{j_1} = S_{j_2} = \dots = S_{j_{N_j}} = j$. Let

$\bar{Y}_m^*(j) = \frac{1}{N_j} \sum_{i=1}^{N_j} Y_{im}^*(j)$, and we can also have the conditional expectation mean

$\bar{Y}_{\tilde{z}}^*(j) = (\bar{Y}_1^*(j), \dots, \bar{Y}_{M_1}^*(j))^t$ for the j th class, $j=1, \dots, J$. Now, we create the between class

variation matrix by

$$\begin{aligned}
B^* &= \sum_{j=1}^J (\bar{Y}_m^*(j) - \bar{Y}_m^*) (\bar{Y}_t^*(j) - \bar{Y}_t^*)^t \\
&= \sum_{j=1}^J BCov_j,
\end{aligned} \tag{4.8}$$

Then, we define the “between class variation”, the distance measure used when

performing alternate k-means clustering. The “between class variation” is then defined as

$$BCV = \sum_{j=1}^J w_j BCov_j \text{ with } w_j = \frac{\text{the number of objects in class } j}{N} \quad (4.9)$$

The BCV is the weighted average of $BCov_j$ over all classes with weights proportional to numbers of objects in each class. BCV can be used as the measure for evaluating assumption (A2). The smaller the value of BVC, the more satisfying the assumption (A2).

When the surrogates are polytomous categorical, each component of $\underline{Y}_i^{(1)}$ and $\underline{Y}_i^{(2)}$ should represent as a vector with elements being the indicators of each category:

$$\tilde{\underline{Y}}_i^{(1)} = \left(\tilde{Y}_{i1}^{(1)}, \dots, \tilde{Y}_{iM_1}^{(1)} \right) = \left(Y_{i11}^{(1)}, \dots, Y_{i1(K_1-1)}^{(1)}, \dots, Y_{iM_11}^{(1)}, \dots, Y_{iM_1(K_m-1)}^{(1)} \right) \quad (4.10)$$

with $Y_{imk}^{(1)} = I(Y_{im}^{(1)} = k) = 1$ if $Y_{im}^{(1)} = k$, 0 otherwise; $m = 1, \dots, M_1$; $k = 1, \dots, (K_m - 1)$, and define the length of $Y_i^{(1)} = M_{K_1}$. And

$$\tilde{\underline{Y}}_i^{(2)} = \left(\tilde{Y}_{i1}^{(2)}, \dots, \tilde{Y}_{iM}^{(2)} \right) = \left(Y_{i11}^{(2)}, \dots, Y_{i1(K_1-1)}^{(2)}, \dots, Y_{iM_21}^{(2)}, \dots, Y_{iM_2(K_m-1)}^{(2)} \right) \quad (4.11)$$

with $Y_{imk}^{(2)} = I(Y_{im}^{(2)} = k) = 1$ if $Y_{im}^{(2)} = k$, 0 otherwise; $m = 1, \dots, M_2$; $k = 1, \dots, (K_m - 1)$, and define the length of $Y_i^{(2)} = M_{K_2}$. Then, we can do above process with these new $\tilde{\underline{Y}}_i^{(1)}$ and $\tilde{\underline{Y}}_i^{(2)}$.

4.1.3 The measurement for our alternate k-means algorithm

We have to create a criterion for our alternate k-means algorithm. The idea is from the sparse k-means clustering method (Daniela M. Witten, & Robert Tibshirani, 2009). Let

$pLoI = LoI_{Y^{(2)}} + \lambda \times BCV_{Y^{(1)}|Y^{(2)}}$ where $LoI_{Y^{(2)}}$ is the loss of independence calculated on $Y^{(2)}$, $BCV_{Y^{(1)}|Y^{(2)}}$ is the between class variation calculated on $Y^{(1)}$ given $Y^{(2)}$, and λ is the tuning parameter.

4.1.4 The alternate k-means algorithm

The alternate k-means algorithm is carried out through following steps to obtain the estimated class membership for individuals and surrogates:

- IK1. Randomly partition the objects into j initial classes.
- IK2. Let all the surrogates be clustering surrogates. Proceed through the list of objects, assigning objects to latent classes with the "loss of independence" as the distance measure.
- IK3. Randomly assign the surrogates to the clustering group with probability 0.8 and to the noisy group with probability 0.2.
- IK4. Fix the object class obtained from IK2. Proceed through the list of surrogates, assigning surrogates to clustering or noisy group with the $pLoI$ as the distance measure.
- IK5. Fix the surrogate group obtained from IK4. Proceed through the list of objects, assigning objects to latent classes with the $pLoI$ as the distance measure.
- IK6. Iterate IK4 and IK5, until the surrogate group assignment convergence. (i.e., there is no surrogate changing group)

In the algorithm IK4 and IK5, we use an standard k-means clustering method to assigning an object to the class and assigning surrogates to clustering/noisy group with the $pLoI$ as the distance measure. The following algorithm describes how the standard K-means clustering method work:

- K1. First, all objects (or surrogates) are partitioned into K initial clusters.
- K2. Proceed through the list of objects (or surrogates), assigning an object (or surrogates) to

the cluster where the minimum $pLoI$ is reached.

K3. Repeat step 2 until no more reassignments take place.

The flow chart for alternate k-means and standard k-means algorithm are showing in Figure 2 and Figure 3.

4.1.5 Estimation of tuning parameter λ

Our alternate k-means algorithm is sensitive to the tuning parameter λ . We have to choose an appropriate value of λ . Here, we propose an idea to select this parameter. First, we calculate the loss of independence $LoI_{\tilde{Y}^{(2)}}$ on $\tilde{Y}^{(2)}$ and the between class variation $BCV_{\tilde{Y}^{(1)}|\tilde{Y}^{(2)}}$ on $\tilde{Y}^{(1)}$ given $\tilde{Y}^{(2)}$ after the algorithm step IK1 and IK2. Then, we set $\lambda \approx \frac{LoI}{BCV}$. This setting can reduce the effect resulting from the difference of these two values.

We believe the large difference between $LoI_{\tilde{Y}^{(2)}}$ and $BCV_{\tilde{Y}^{(1)}|\tilde{Y}^{(2)}}$ will make the algorithm failed, and we find the appropriate λ not only shrinks the difference of two values, but also makes a good prediction result.

4.2 Latent class membership estimation when incorporating covariate effects

The alternate k-means clustering algorithms are based on the assumption (A1) and (A2). If covariates \mathbf{z}_{im} are incorporated into the conditional distributions as in model (3.3) and (3.4), the conditional independence assumption is also conditioning on incorporated covariates (i.e., the assumption (3.5)). To apply these algorithms to model (3.3) and (3.4), one would need to “eliminate” the covariate effects, hence “marginalize” model (3.3) and (3.4).

Here, we adopt the marginalization process develop in 3.3.1 of (Huang, 2005). To present the process, we first reparameterize models (3.3) and (3.4) as

$$\log \left(\frac{\Pr(Y_{im} = k | S_{i1}, \dots, S_{i(J-1)})}{\Pr(Y_{im} = K | S_{i1}, \dots, S_{i(J-1)})} \right) = \gamma_{mk0} + \gamma_{mk1} S_{i1} + \dots + \gamma_{mk(J-1)} S_{i(J-1)} + \alpha_{1mk} z_{im1} + \dots + \alpha_{Lmk} z_{imL} \text{ for } k = 1, \dots, (K_m - 1) \quad (4.12)$$

and

$$E(Y_{im} | S_{i1}, \dots, S_{i(J-1)}, z_{im}) = \theta_{m0} + \theta_{m1} S_{i1} + \dots + \theta_{m(J-1)} S_{i(J-1)} + \tau_{1m} z_{im1} + \dots + \tau_{Lm} z_{imL} \quad (4.13)$$

where $S_{ij} = I(S_i = j)$, $j = 1, \dots, (J-1)$. In brief, the process assumes that the incorporated covariates \mathbf{z}_{im} and the class membership S_i are orthogonal, and calculates the residual of regressing Y_{im} on \mathbf{z}_{im} separately for each $m \in \{1, \dots, M\}$. One can then extract \mathbf{z}_{im} from conditional distributions by treating these residuals as new response variables and regressing them on S_i . Therefore, the conditional independent assumption (3.5) is considered satisfied if objects belonging to the same latent class have a set of M_2 statistically independent residuals.

Now, we consider $\underline{Y} = (Y_{i1}, \dots, Y_{iM}) = (\underline{Y}_i^{(1)}, \underline{Y}_i^{(2)})$. When Y_{im} 's are continuous, the typical residuals of linear regressions R_{im} (i.e., the differences between observed responses and their modeled predictors) are computed. When Y_{im} 's are categorical, the problem becomes how to calculate residual from the generalized linear model

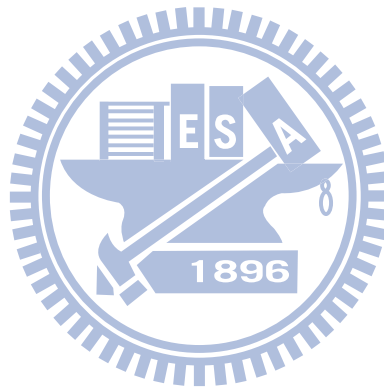
$$\log \left(\frac{\Pr(Y_{im} = k | z_{im})}{\Pr(Y_{im} = K | z_{im})} \right) = \alpha_{1mk} z_{im1} + \dots + \alpha_{Lmk} z_{imL}, \text{ for } k = 1, \dots, (K_m - 1). \quad (4.14)$$

We propose to use the ‘‘pseudo-residual’’

$$\tilde{R}_{im} = \left(Cov(\tilde{Y}_{im}) \right)^{-1} (\tilde{Y}_{im} - \hat{p}_{im}),$$

(4.15)

where \tilde{Y}_{im} is as defined in section, $p_{im} = E(\tilde{Y}_{im} | z_{im})$, and “hat” denotes the estimated values based on (4.13). The pseudo-residual (4.14) is defined by analogizing the alternately reweighted least-squares of generalized linear models with the least-square estimates of linear regressions (Landwehr, Pregibon, & Shoemaker, 1984; Huang, 2005). We then classify objects based on new response variables R_{im} (continuous surrogates) or \tilde{R}_{im} (categorical surrogates) as done in the previous subsection.



5 Classification Using Finite Mixture Models

In many researches, it is major interest to predict new observations' unknown disease statuses based on their measurements on surrogates. Some literature develops the method to create the classification rules (Huang, Wang & Hsu, manuscript), and we use their ideas to create the classification rules base on our model. Consider a set of N objects with known disease statuses D_i and measured surrogates Y_i plus incorporated covariate x_i, z_i if existing, for $i = 1, \dots, N$, where D_i takes values $\{1, \dots, C\}$. We use these to fit finite mixture models (3.1), (3.2), (3.3) and/or (3.4). Then. We can obtain estimations $\hat{S}_i, \hat{\beta}_{pj}, (\hat{\theta}_{mj}, \hat{t}_{lm}, \hat{\sigma}_m^2)$ and $(\hat{\gamma}_{mkj}, \hat{\alpha}_{lmk})$, for all i, j, m, p, l, k . For a new object with measurements on surrogates $Y^* = (Y_1^*, \dots, Y_M^*)$ and covariates $x^*, z^* = (z_1^*, \dots, z_M^*)$, we assume that

$$\Pr(D^* = c | S^* = j, Y^*, x^*, z^*) = \Pr(D^* = c | S^* = j), \quad (5.1)$$

then the posterior probability of classifying him/her as the disease status $D^* = c$ is

$$\Pr(D^* = c | Y^*, x^*, z^*) = \sum_{j=1}^J \left\{ \Pr(D^* = c | S^* = j) \Pr(S^* = j | Y^*, x^*, z^*) \right\} \quad (5.2)$$

where S^* is the presumed latent class membership of the new object. In other word, the latent class can fully capture the association between the disease status and observed surrogates, which is reasonable when viewing the latent class variable as well summary of measured surrogates. We can estimate the right hand side of (5.2) by

$$\hat{\Pr}(D^* = c | S^* = j) = \frac{\sum_{i=1}^N I(\hat{S}_i = j) I(D_i = c)}{\sum_{i=1}^N I(\hat{S}_i = j)} \quad (5.3)$$

and

$$\hat{\Pr}(S^* = j | Y^* = (y_1^*, \dots, y_M^*), x^*, z^*) = \frac{\hat{\eta}_j(x^*) \prod_{m=1}^{M_2} \hat{f}_t(y_m^{(2)*} | z_m^*)}{\sum_{t=1}^J \left\{ \hat{\eta}_t(x^*) \prod_{m=1}^{M_2} \hat{f}_t(y_m^{(2)*} | z_m^*) \right\}}, \quad (5.4)$$

where $\hat{\eta}_j(x^*)$ is the estimated latent prevalence of the j th class for new observation x^* , evaluated at estimator $\hat{\beta}_{pj}$. Notice that, only surrogates in the clustering group $\{y_1^{(2)*}, \dots, y_{M_2}^{(2)*}\}$ are used in the prediction. $\hat{f}_j(y_m^{(2)*} | z_m^*)$ is the estimated conditional distribution of the m th surrogate given the j th class for the new observation $(y_m^{(2)*}, z_m^*)$, evaluated at estimators $(\hat{\theta}_{mj}, \hat{\tau}_{lm}, \hat{\sigma}_m^2)$ and $(\hat{\gamma}_{mkj}, \hat{\alpha}_{lmk})$. We propose to choose c for the maximum estimated posterior probability is reached, i.e.,

$$c^* = \arg \max_{c \in \{1, \dots, C\}} \hat{\Pr}(D^* = c | Y^*, x^*, z^*). \quad (5.5)$$

In this classification rule, a new object's disease status is predicted through his/her inferred latent class variable S^* , and it can be viewed as the summary of the new object's measured surrogates through the training set $\{Y_1, \dots, Y_N\}$.

6 Example

In this section, we consider the Breast cancer data (continuous) and Schizophrenia syndrome scale data (categorical) examples, and use standard k-means and alternate k-means clustering method to estimate the parameters in original LCA model and our model. Furthermore, we use the proposed classification rule (Huang, &Wang, &Hsu) for prediction. Here, we introduced a useful tool for clustering. Heatmap has the notion of rearranging the columns and rows to show structure in the data. A heatmap is a two-dimensional, rectangular, colored grid, and shows data that themselves come in the form of a rectangular matrix. The color of each rectangle is determined by the value of the corresponding entry in the data matrix. The rows and columns of the matrix can be rearranged independently. Usually they are using clustering methods for reorder such that similar rows are placed next to each other, and the same for columns. Among the orderings that are widely used are those derived from a hierarchical clustering, but many other orderings are possible. If hierarchical clustering is used, then it is customary that the dendrograms are provided as well. Here, we use non-hierarchical clustering methods (i.e., k-means and alternate k-means clustering methods) to find some subgroup for individuals and plot the heatmap by these groups. On the other hand, we use agglomerative hierarchical clustering methods to grouping the surrogates with distance measurement using one minus correlation. We will use the heatmap figures to show our result.

6.1 Breast cancer data

The data come from a study of using gene expression profiling to predict breast cancer outcome (Veer et al., 2002). The 78 sporadic lymph-node-negative patients under 55 years of age were selected specifically to search for a prognostic signature in their gene expression profiles. Forty-four patients remained free of disease after their initial diagnosis for an interval

of at least 5 years (good prognosis group, mean follow-up of 8.7 years), and 34 patients had developed distant metastases within 5 years (poor prognosis group, mean time to metastases 2.5 years). From each patient, total RNA was isolated from tumor material and used to drive cRNA. A reference cRNA pool was made by pooling equal amounts of cRNA from each of the sporadic carcinomas. Fluorescence intensities were quantified, normalized and corrected to yield the transcript abundance of a gene as an intensity ratio with respect to that of the signal of the reference pool (Hughes et al., 2001).

Here, we aim to predict good and poor prognostic patients through gene expression profiling. We use a two-step selection process was performed to retain genes in the analysis. Firstly, 4741 genes selected from 24481 genes with the intensity ratio > 2 or < 0.5 (i.e., more than two-fold difference) and the significance of regulation p-value < 0.01 in more than 3 patients. This was used in the original paper and focused the attention to the most informative genes. In the second step, we applied a selection of genes based on the ratio of their between-group to within-group sums of squares, as suggested by (Dudoit, Fridlyand, & Speed, 2002). For a gene m , that ratio is

$$BW(m) = \frac{\sum_i \sum_c I(d_i = c) (\bar{y}_{cm} - \bar{y}_{.m})^2}{\sum_i \sum_c I(d_i = c) (y_{im} - \bar{y}_{cm})^2}, \quad (6.1)$$

where y_{im} denotes the intensity ratio of gene m in the patient i , d_i is the indicator of good (=1) or poor (=0) prognosis group of patient i , and \bar{y}_{cm} and $\bar{y}_{.m}$ are the average intensity ratio of gene m across samples belonging to prognosis group c only and across all patients, respectively. We use (6.1) to compute BW ratio for each gene and selected top 200 genes with the largest BW ratios for finite mixture analysis.

Using 200 selected expression ratios as observed surrogates, a finite mixture model (3.1), (3.2), (3.4) was fitted. In the fitted model, age at diagnosis (year) was chosen to be associated with conditional probabilities, and latent prevalence was also modeled as depending on age at

diagnosis. We used the standard k-means clustering approach to group patients and resulted in 3 classes of size 40, 21, and 17. In the alternate k-means clustering method, only needs to decide the tuning parameter λ ; and the nearest neighbors r . We choose $\lambda=130$ (the initial $\text{LoI}_{\chi^{(2)}} = 0.058633$ and $\text{BCV}_{\chi^{(1)}|\chi^{(2)}} = 0.154481$) and $r=3$. The alternate k-means clustering method selected 154 out of 200 genes as the clustering surrogates. This approach resulted in 3 classes of size 33, 22, and 23. The heatmap for the 200-gene (original) and 154-gene (selected) expression profile are displayed in Figures 4 and 5.

An additional independent set of primary tumors from 19 young, lymph-node-negative breast cancer patients was used to validate the above 154-gene prognosis classifier. This group consisted of 7 patients who remained free of disease for at least five years, and 12 patients who developed distant metastases within five years. Table 1 and 2 shows the result of prediction from the standard k-means and alternate k-means. Consequently, the standard k-means approaches had 4 out of 19 incorrect classifications, but the alternate k-means approaches had 3 out of 19 incorrect classifications.

6.2 Schizophrenia syndrome scale data

The data were collected from a series of projects, aiming at investigating the clinical manifestations of schizophrenia and searching for neuropsychological, environmental and genetic factors underlying schizophrenia. Details of study design and eligibility criteria were described previously (Liu, Hwu, & Chen, 1997; Chen et al., 1998; Chang et al., 2002). The analyzed data include 164 acute-state patients of schizophrenia who were recruited within one week of index admission and 155 subsided stage patients who were living with community and under family care.

In this study, schizophrenia symptoms were assessed by the Positive and Negative Syndrome Scale (PANSS) (Cheng, Ho, Chang, Lane, & Hwu, 1996). The PANSS has 30

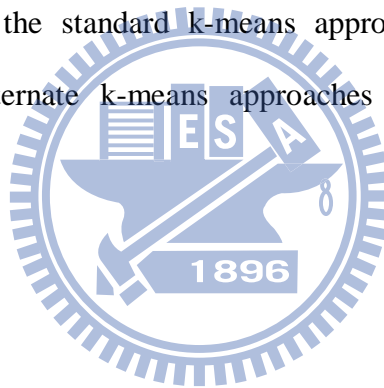
items and consists of three subscales: positive (seven symptoms: P1-P7), negative (seven symptoms: N1-N7) and general psychopathology (sixteen symptoms: G1-G16). Each item was originally rated on a 7-point scale (1=absent, 7=extreme), but we reduced the 7-point scale by merging the points that had the response percentages less than 10%. This study considered external covariates including demographic variables and environmental / neuropsychological factors. Demographic variables included gender, age at recruitment, years of education, and occupation (having versus no occupation). The category of no occupation included housewives, students, unemployed and retired people. The environmental factors were related to obstetric complications, prenatal growth retardation, special personal behavior and psychological adjustment problems. And the neuropsychological batteries assessed reaction time, attention, speed of information processing, and active problem solving. Specifically, the test batteries included several standard neuropsychological instruments with demonstrated reliability and validity, and we concentrated on the Continuous Performance Test (CPT), which had been widely used to measure sustained attention deficits in psychotic disorders (Chen et al., 1998).

The analysis aims to explore the subtype (groups) of schizophrenia patients based on PANSS measurement. In our application, the latent class model of (3.1), (3.2), and (3.3) was applied to 30 PANSS items. We let the covariates associated with conditional probabilities include variables of sex, age (year), years of education (year), and occupation (with versus without occupation), and the covariates associated with latent prevalence include variables of age of onset (year), $envir_{11}$, $envir_{21}$, $envir_{22}$, $envir_{31}$, $envir_{32}$, and $dprime$. We used the standard k-means clustering approach to group patients and resulted in 4 classes of size 231, 31, 52, and 5. We choose the tuning parameter $\lambda = 30$ (the initial $LoI_{Y^{(2)}} = 1.174653$ and $BCV_{Y^{(1)}|Y^{(2)}} = 1.898806$) and $r = 3$ in alternate k-means process. Eighteen (3 positive, 2 negative, and 13 general psychopathology) out of 30 items were selected as clustering

surrogates. This approach grouped patients in 4 classes of size 221, 41, 47, and 10. The heatmap for the 30-item (original) and 19item (selected) are showed in Figures 6 and 7.

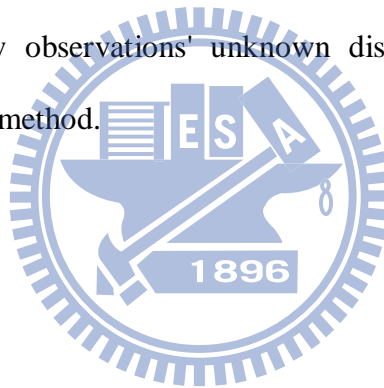
In general, class 1 appeared to represent a group who had severe/extreme positive symptoms and moderate negative symptoms; class 2 was a group who had moderate positive symptoms but mild negative symptoms; class 3 represented a group who had widespread whole syndrome of severe positive and negative symptoms; and class 4 was a remitted group who rarely had any symptom.

Then, we are interested in using the PANSS ratings to predict patients' phases of chronicity of disease (acute versus subsided). There has 10 patients in the prediction group which is consisted of 5 acute patients and 5 subsided patients. Table 3 and 4 shows the result of predicts. Consequently, the standard k-means approaches had 3 out of 10 incorrect classifications, and the alternate k-means approaches had just 1 out of 10 incorrect classifications.



7 Discussion

We have proposed to use the alternate k-means clustering methods to search for the optimal class allocation that can make clustering surrogates as independent as possible for objects belonging the same class, select the surrogates for estimating parameters in the model and create classification rule. By treating the identified class allocation as a known predictor, the parameters underlying a finite mixture model can then be estimated. We further use a classification rule, based on the finite mixture model. From the real data analysis, we demonstrate the ability in surrogate selection and handling the high-dimensional data and the accuracy of the classification rule in predicting new observations' unknown disease statuses. Here, we can see that the alternate k-means clustering method can reduce the size of surrogates and predict new observations' unknown disease statuses more accurate than original K-means clustering method.



Reference

- Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple outcomes. *Journal of the American Statistical Association*, 92 , 1375-1386.
- Brusco, M. J., & Cradit, J. D. (2001). A variable selection heuristic for k-means clustering. *Psychometrika*, 66 , 249-270.
- Chang, C. J., Chen, W. J., Liu, S. K., Cheng, J. J., Ou Yang, W. C., Chang, H. J., et al. (2002). Morbidity risk of psychiatric disorders among the first degree relatives of schizophrenia patients in taiwan. *Schizophrenia Bulletin*, 28 , 379-392.
- Chen, W. J., Liu, S. K., Chang, C. J., Lien, Y. J., Chang, Y. H., & Hwu, H. G. (1998). Sustained attention deficit and schizotypal personality features in nonpsychotic relatives of schizophrenic patients. *American Journal of Psychiatry*, 155 , 1214-1220.
- Cheng, J. J., Ho, H., Chang, C. J., Lane, S. Y., & Hwu, H. G. (1996). Positive and negative syndrome scale (panss): Establishment and reliability study of a mandarin chinese language version. *Taiwanese Journal Psychiatry*, 10 , 251-258.
- Dayton, C. M., & Macready, G. B. (1998). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83 , 173-178.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, 39 , 1-38.
- Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Latent variable regression for multiple outcomes. Comparison of discrimination methods for the classification of tumors using gene expression data, 97 , 77-87.
- Friedman, J. H., & Meulman, J. J. (2004). Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society. Series B*, 66 , 815-849.
- Frank, I. E. & Friedman, J. H. (1993). A statistical view of some chemometrics regression

- tools. *Technometrics*, 35, 109-148.
- Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61 , 215-231.
- Huang, G. H. (2005). Selecting the number of classes under latent class regression: a factor analytic analogue. *Psychometrika*, 70 , 325-345.
- Huang, G. H., & Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, 69 , 5-32.
- Huang, G. H., Wang, S.M., & Hsu, C.C. Prediction of Underlying Latent Classes via K-means and Hierarchical Clustering Algorithms. Manuscript.
- Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., et al. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology*, 19 , 342-347.
- Landwehr, J. M., Pregibon, D., & Shoemaker, C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, 79 , 61-71.
- Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. New York: Houghton-Mifflin.
- Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88 , 365-411.
- Liu, S. K., Hwu, H. G., & Chen, W. J. (1997). Clinical symptom dimensions and deficits on the continuous performance test in schizophrenia. *Schizophrenia Research*, 25 , 211-219.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*, second edition. London: Chapman and Hall.
- Melton, B., Liang, K. Y., & Pulver, A. E. (1994). Extended latent class approach to the study of familial/sporadic forms of a disease: its application to the study of the heterogeneity of schizophrenia. *Genetic Epidemiology*, 11 , 311-327.
- Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables.

- British Journal of Mathematical and Statistical Psychology, 49 , 313-334.
- Muthen, L. K., & Muthen, B. O. (2007). Mplus user's guide. fifth edition. Los Angeles: Muthen & Muthen.
- Pan, W. & Shen, X. (2007). Penalized model-based clustering with application to variable selection. Journal of Machine Learning Research 8, 1145-1164.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen & Lydiche.
- Rosvold, H. E., Mirsk, A. F., Sarason, I., Bransome Jr, D. D., & Bech, L. H. (1956). A continuous performance test of brain damage. Journal of Consulting Psychology, 20 , 343-350.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, 58, 267-288.
- Titterington, D. M., Smith, A. F., & Makov, U. E. (1985). Statistical analysis of finite mixture distributions. New York: John Wiley & Sons.
- Veer, L. J. van't, Dai, H., Vijver, M. J. van de, He, Y. D., Hart, A. A., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415 , 530-536.
- Witten, D.M. & Tibshirani, R. (2010). A framework for feature selection in clustering. Journal of the American Statistical Association. To appear.

Table 1: Predictions of class membership of 19 tumors by standard k-means clustering method

Individual membership*	Prediction class	Posterior Probability		True class membership
		class 1	class 2	
1	1	0.550	0.450	1
2	1	0.706	0.294	1
3	1	0.550	0.450	1
4	1	0.550	0.450	1
5	1	0.706	0.294	1
6	1	0.550	0.450	1
7	1	0.550	0.450	1
8	1	0.550	0.450	1
9	1	0.550	0.450	1
10	1	0.550	0.450	1
11	1	0.706	0.294	1
12	1	0.550	0.450	1
13	2	0.001	0.999	2
14	1	0.550	0.450	2
15	2	0.001	0.999	2
16	1	0.550	0.450	2
17	1	0.550	0.450	2
18	1	0.706	0.294	2
19	2	0.001	0.999	2

*Values in bold are misclassification

Table 2: Predictions of class membership of 19 tumors by alternate k-means clustering method

Individual	Prediction class membership*	Posterior Probability		True class membership
		class 1	class 2	
1	1	0.576	0.424	1
2	1	0.565	0.435	1
3	1	0.572	0.428	1
4	1	0.576	0.424	1
5	1	0.565	0.435	1
6	1	0.573	0.427	1
7	1	0.575	0.425	1
8	1	0.565	0.435	1
9	1	0.570	0.430	1
10	1	0.565	0.435	1
11	1	0.565	0.435	1
12	1	0.576	0.424	1
13	2	0.091	0.909	2
14	1	0.574	0.426	2
15	2	0.091	0.909	2
16	1	0.576	0.424	2
17	2	0.469	0.531	2
18	1	0.565	0.435	2
19	2	0.091	0.909	2

*Values in bold are misclassification

Table 3: Predictions of class membership of 10 schizophrenia patient by standard k-means clustering method

Individual membership*	Prediction class	Posterior Probability		True class membership
		class 1	class 2	
1	2	0.494	0.506	2
2	2	0.115	0.885	2
3	2	0.494	0.506	2
4	2	0.115	0.885	2
5	2	0.494	0.506	2
6	1	0.968	0.032	1
7	1	0.958	0.042	1
8	2	0.494	0.506	1
9	2	0.497	0.503	1
10	2	0.494	0.506	1

*Values in bold are misclassification

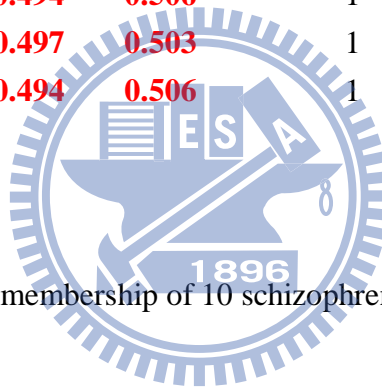


Table 4: Predictions of class membership of 10 schizophrenia patient by alternate k-means clustering method

Individual membership*	Prediction class	Posterior Probability		True class membership
		class 1	class 2	
1	2	0.489	0.511	2
2	2	0.128	0.872	2
3	2	0.488	0.512	2
4	2	0.170	0.830	2
5	2	0.471	0.529	2
6	1	0.826	0.174	1
7	1	0.699	0.301	1
8	1	0.519	0.481	1
9	1	0.803	0.197	1
10	2	0.488	0.512	1

*Values in bold are misclassification

Figure 1: Lasso and Ridge regression

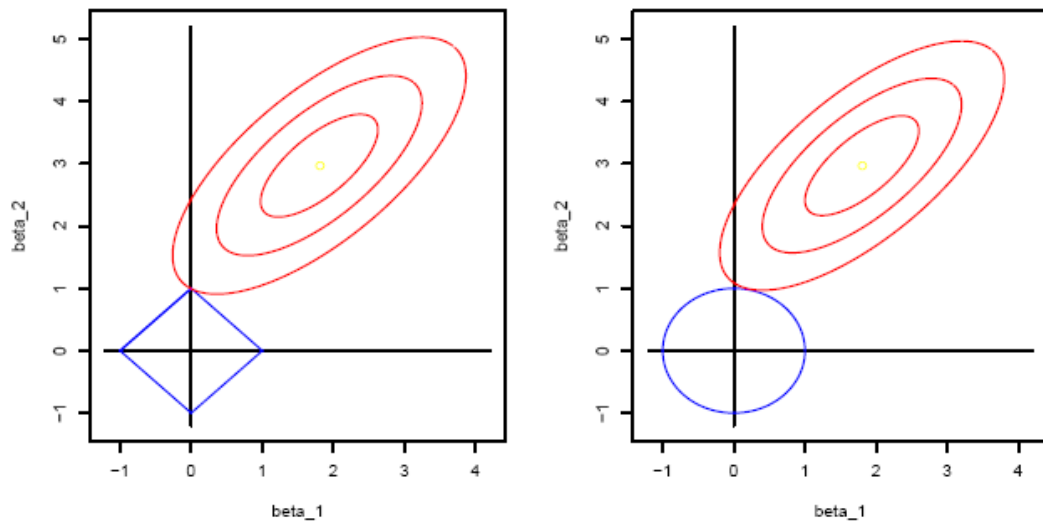


Figure 1:Lasso and Ridge

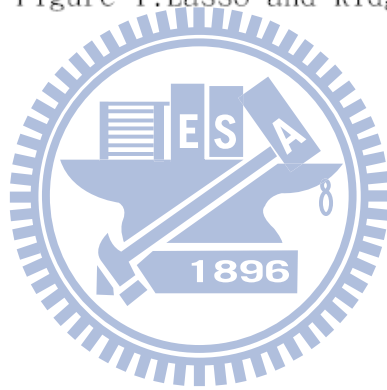


Figure 2: The flow chart of alternate k-means

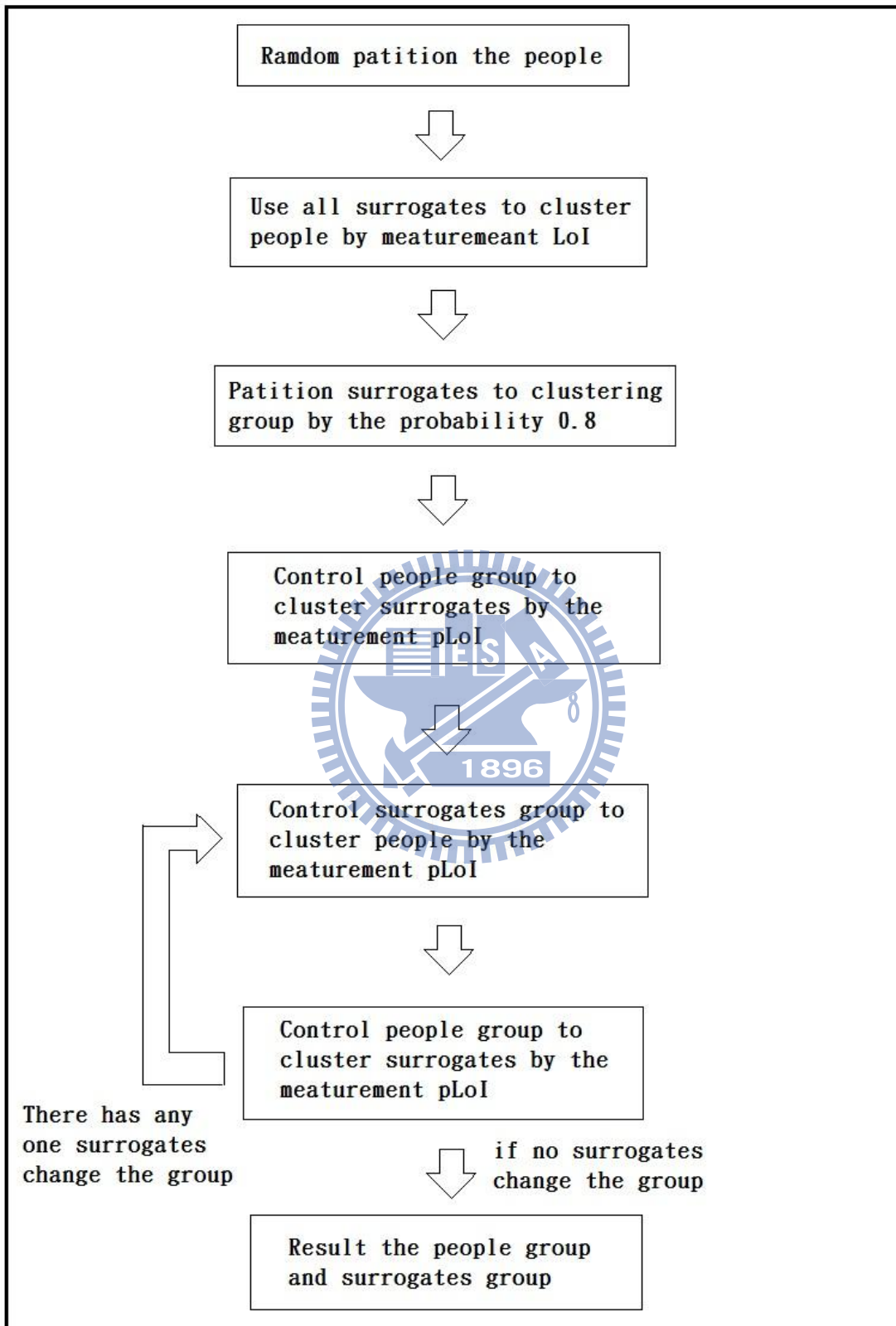


Figure 3: The flow chart of standard k-means

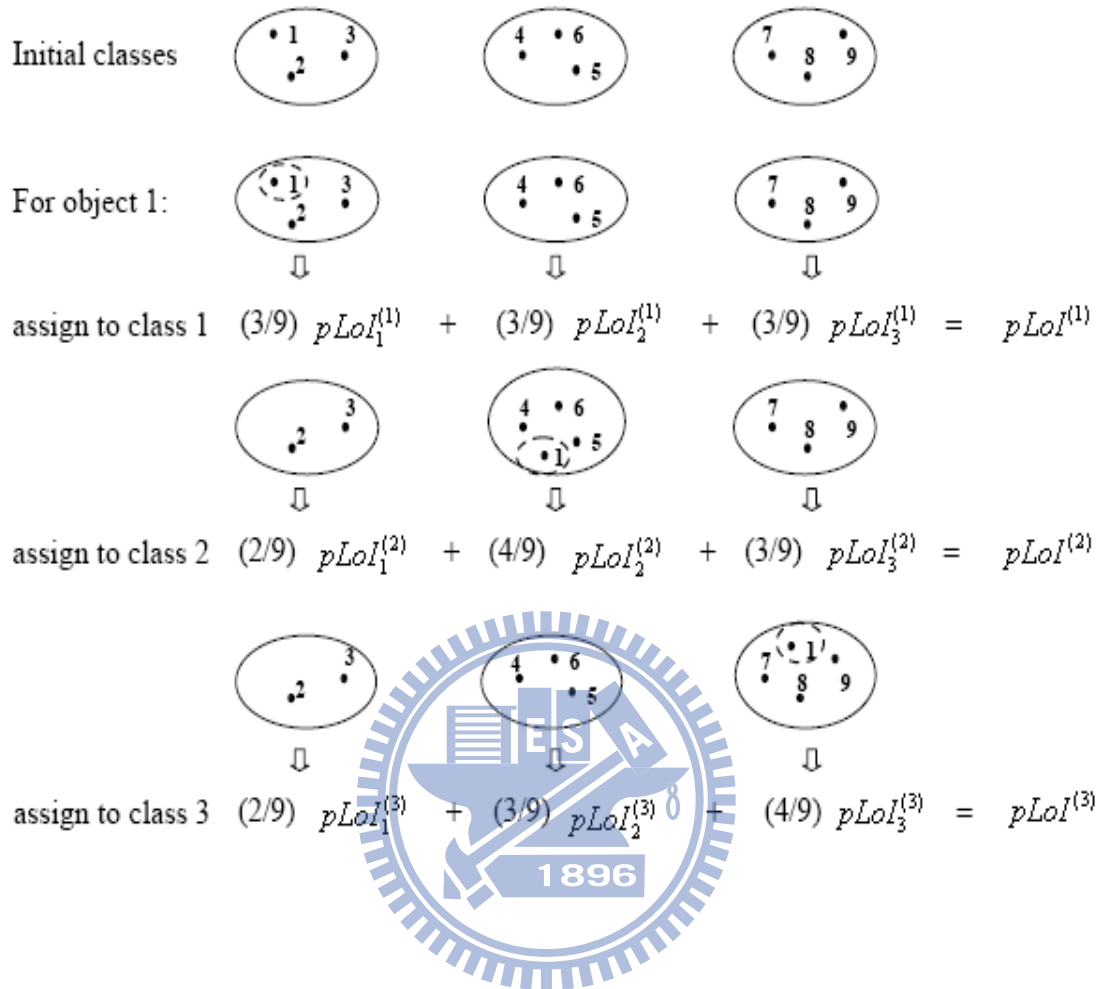


Figure 4: The heatmap for the 200-gene (original)

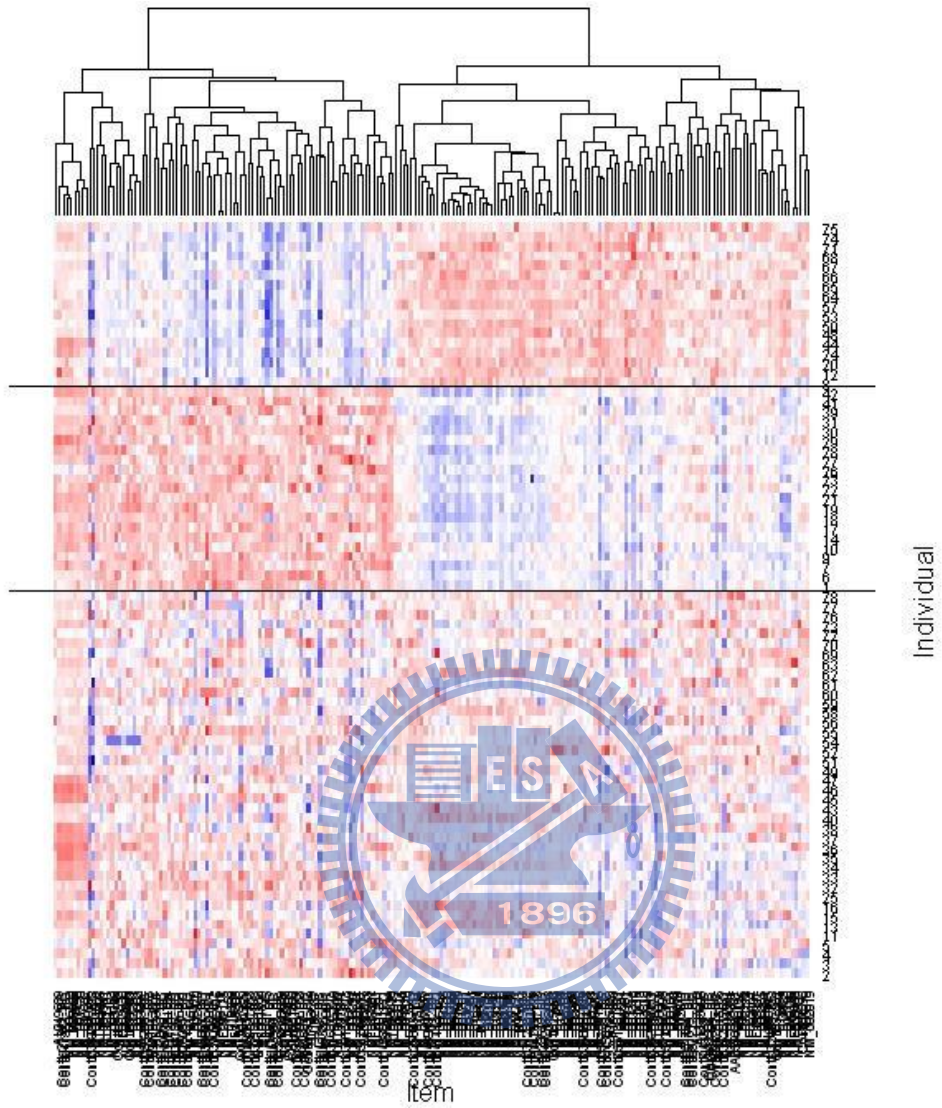


Figure 5: The heatmap for the 154-gene (selected)

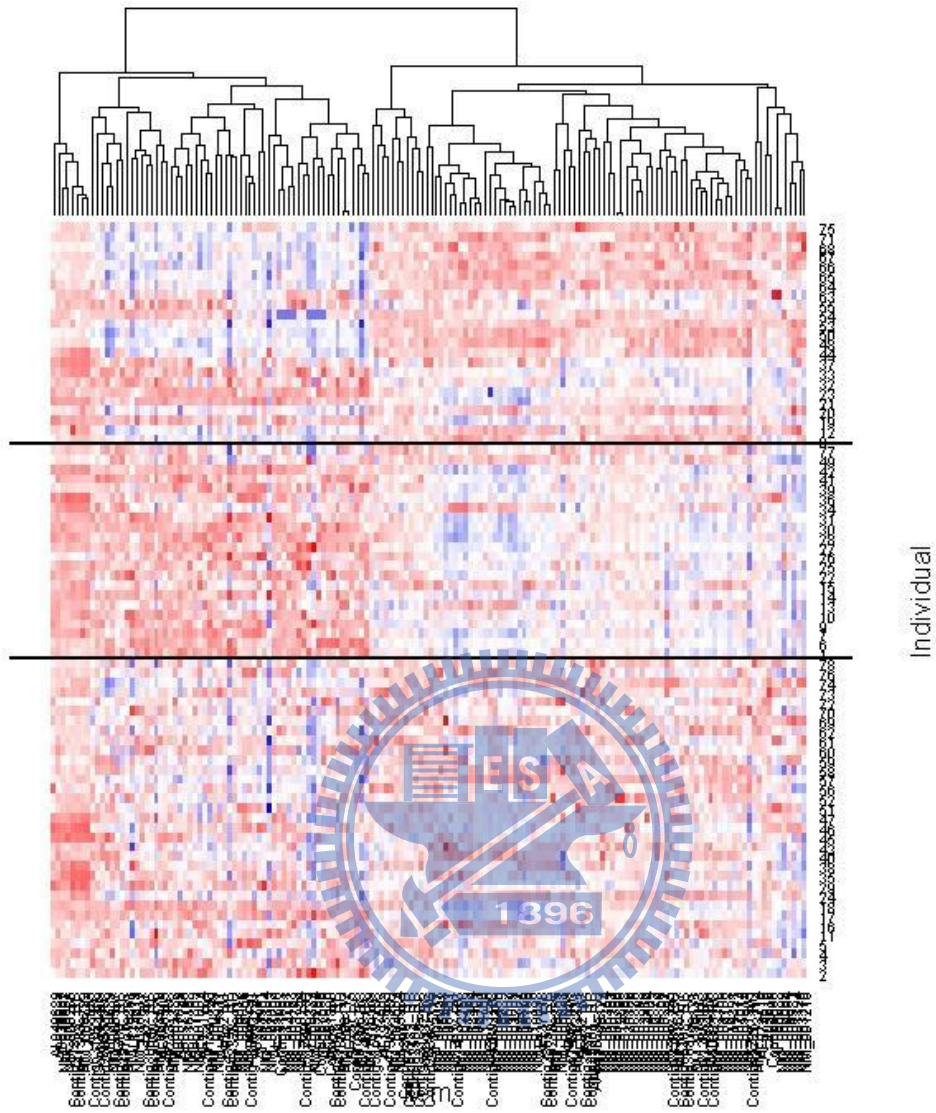


Figure 6: The heatmap for the 30-item (original)

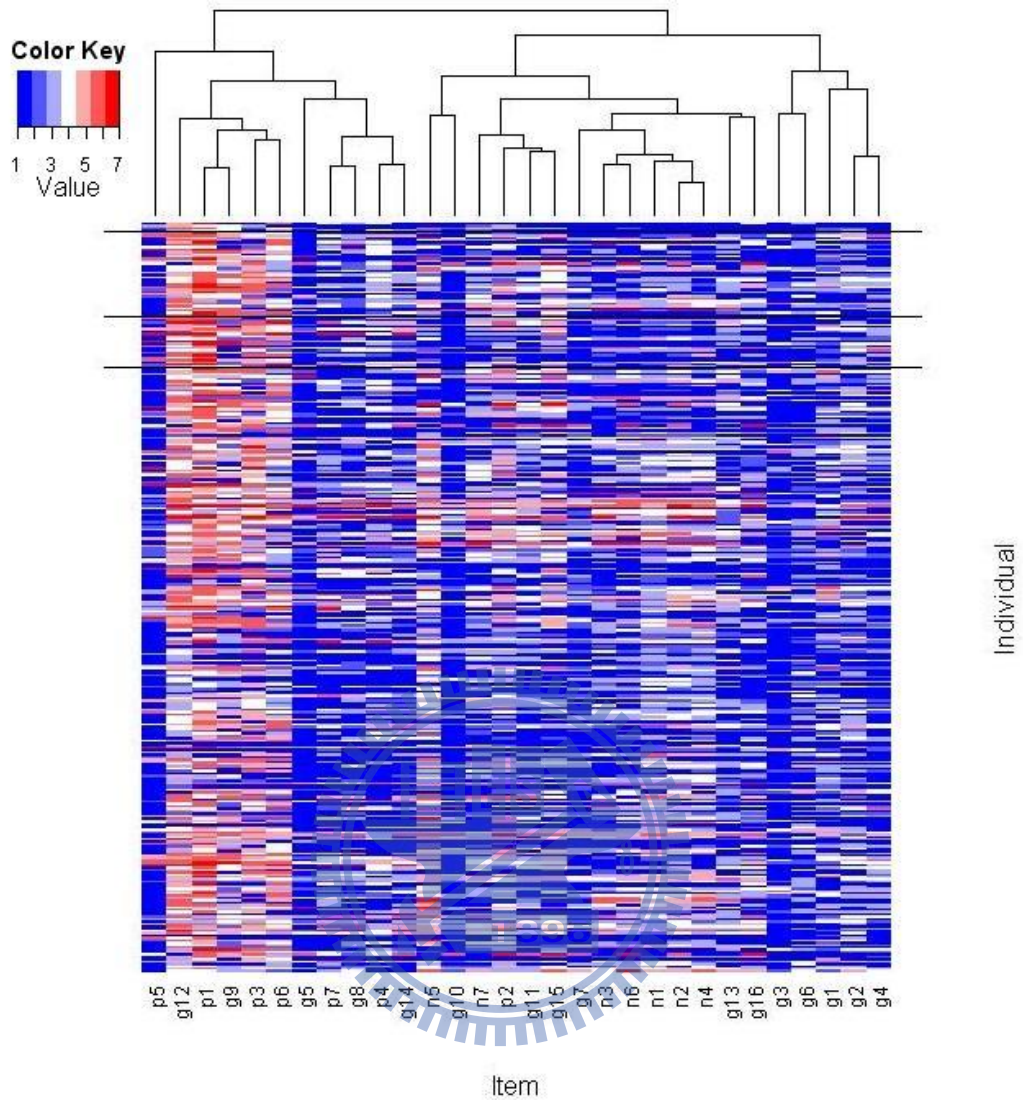


Figure 7: The heatmap for the 18-item (selected)

