Classification method for heterogeneity in monoclonal cell population

S Aburatani[1], K Tashiro[2] and S Kuhara[2]

[1] Senior Research Scientist, BRIDD, National Institute of AIST, AIST Tokyo Waterfront Bio-IT Research Building, 2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan
[2] Grad. Sch. of Biores. Bioenv. Sci, Department of Systems Life Sciences, Kyushu Univ., 6-10-1, Hakozaki, Higashi-ku, Fukuoka-city, Fukuoka, 812-8581, Japan

E-mail: s.aburatani@aist.go.jp

Abstract. Monoclonal cell populations are known to be composed of heterogeneous sub-populations, thus complicating the data analysis. To gain clear insights into the mechanisms of cellular systems, biological data from a homogeneous cell population should be obtained. In this study, we developed a method based on Latent Profile Analysis (LPA) combined with Confirmatory Factor Analysis (CFA) to divide mixed data into classes, depending on their heterogeneity. In general cluster analysis, the number of measured points is a constraint, and thereby the data must be classified into fewer groups than the number of samples. By our newly developed method, the measured data can be divided into groups depending on their latent effects, without constraints. Our method is useful to clarify all types of omics data, including transcriptome, proteome and metabolic information.

1. Introduction

According to progress in science and technology, we can obtain information about the intracellular behaviors of cell components as omics data, such as expression profiles, proteome data and metabolic data [1,2,3]. To analyze these data, it was assumed that the cell populations were basically homogeneous. However, recent studies have shown that cell populations, even monoclonal cell populations, are heterogeneous [4,5]. Furthermore, the differences in cell kinetics dependent on the heterogeneity affect organogenesis and cell differentiation controls in pluripotent stem cells [6], but the mechanisms have remained unclear. To clarify the mechanisms in living cells, homogeneous data should be analyzed.

In order to obtain homogeneous data from mixed heterogeneous data, we developed a method for dividing heterogeneous data into subclasses, depending on their heterogeneity. In our previous investigations, we utilized an improved Structural Equation Modelling (SEM) approach to infer the latent effect for gene regulatory networks [7], and latent variable models were employed to infer the unobserved effects for perturbations of cellular components and metabolism.

[1] Sachiyo Aburatani, s.aburatani@aist.go.jp.

In this study, we combined Latent Profile Analysis (LPA) with Confirmatory Factor Analysis (CFA) to estimate the effects of unknown sub-populations. According to the estimated effects, the measured data can be divided into subclasses. We then reapplied our developed method to each data subclass. Although the clustering method is restricted by the number of samples in general, our method allows classification into more subclasses than the measured data sample number.

2. Methodology

2.1. Latent profile analysis

In the latent mixture model, a population is composed of unobserved subclasses that depend on an independent distribution. The measured data in a monoclonal cell population are considered to be composed of different data populations. To detect the heterogeneity within the data, we applied Latent Profile Analysis (LPA). Latent profile analysis is a cluster analysis technique based on latent variable models, and is applied when the latent variables are categorical data and the measured data are numerical data [8,9]. The measured variables in LPA are assumed to be normal, and their conditional distribution given the latent variables is assumed to be normal [9]. The advantages of latent profile analysis over cluster analysis are that they are model-based, and thus generate probabilities for group membership.

In this analysis, the LPA model for a single item can be written as follows:

$$x_{vi} = \sum_{g=1}^{G} \pi_g f_{ig} \ . \tag{1}$$

Here, x_{vi} denotes the standardized data for a randomly selected individual sample v obtained on item i (i=1, …, I), and π_g is the class size parameter, which indicates the unconditional probability of belonging to latent class g (g=1, …, G). In the LPA model, each individual v belongs to one latent class. Thus, the sum of all class size parameters equals 1:

$$\sum_{g=1}^{G} \pi_g = 1 . \tag{2}$$

The parameter f_{ig} is the effect score on item i given membership in class g, and f_{ig} is calculated for each x_{vi}. This x_{vi} is often referred to as a conditional response, and it depends on both the latent class prevalence π_g and the class-specific effect.

2.2. Confirmatory factor analysis

To assess the effect of each latent variable, we applied Confirmatory Factor Analysis (CFA). Factor analysis is a statistical method for describing the variability among observed variables by assumed latent variables. The number of latent variables is usually smaller than the number of observed samples [10]. In the CFA model, it is assumed that each observed sample may be related to all latent variables. Let us assume that there are p latent variables F_1, F_2, … , F_p and q observed samples x_1, x_2, … , x_q. Since all of the observed variables were standardized, the mean of each observed variable is 0, and can expressed as linear combinations of latent variables, as follows:

$$x_v = \alpha_{v1}F_1 + \alpha_{v2}F_2 + \cdots \alpha_{vp}F_p + \varepsilon_v \tag{3}$$

where x_v is the I-dimensional vector of the sample v, α_{vk} is the partial regression weight of the latent variable F_k, and ε_v is an independently distributed error term with zero mean and finite variance. In matrix form, equation (3) is expressed as

$$X = \Lambda F + E \ . \tag{2}$$

In this case, the item number is I, and then X and E are the (q×I) matrix composed of the observed data and error terms, respectively. The partial regression coefficients of each latent variable are indicated as elements of Λ, the (q×p) latent interaction matrix. In the matrix Λ, each column corresponds to a factor and each row corresponds to an observed sample, and thus each element of Λ

indicates the strength of the regulation from each latent variable to each sample. The matrix F is the latent variable vector.

The variance-covariance matrix between the observed variables is structurized by parameters, as follows:

$$Var[X] = E[(X-U)(X-U)^t] = \Sigma = \Lambda\Phi\Lambda^t + \Psi^2. \quad (3)$$

Here, Ψ^2 is the covariance matrix of error terms, Λ is the factor loading matrix of latent variables, and Φ is the covariance matrix among factors. From this structurized matrix, the values of the partial regression weight matrix Λ and the variances of the "errors" are estimated.

2.3. Estimation of subclass number

To estimate the optimal subclass number for dividing the data, the LPA and CFA results were compared. In LPA, the number of divided classes should be set as the class-parameter, and the prevalence of each class is calculated depending on the set parameter. Thus, we applied LPA with all possible numbers as the class-parameter.

In CFA, the number of latent variables should be assumed in the models, and all possible models were tested in this study. Basically, the number of latent variables is smaller than the number of samples, and thus the minimum number of latent variables is two and the maximum number is (I-1). By CFA, the partial regression weights of the latent variables are calculated for each sample. Although the weights of all assumed latent variables are estimated, an observed sample can be divided into some groups depending on the largest absolute values of the partial regression weight.

The ratio of samples that are classified into groups in CFA is considered to have the same meaning as the class prevalence in LPA. Thus, a comparison of the class prevalence in LPA and the classification ratio in CFA is useful to estimate a suitable number of subclasses.

2.4. Calculation of estimated subclass data

In LPA, the individual data x_{vi} is assumed to depend on both the latent class prevalence π_g and the effect score f_{ig}, which is the product of the latent class-specific effect $f_{.g}$ and the item-specific effect $f_{i.}$. By CFA, the latent class g-specific effect of the sample v is estimated as the partial regression weight α_{vg}. According to the class prevalence π_g and the partial regression weight α_{vg}, equation (1) is expressed as

$$x_{vi} = \kappa_{vi} \sum_{g=1}^{G} \pi_g \alpha_{vg} \quad (4)$$

where κ_{vi} is the individual weight for each x_{vi}. The data of each subclass can be calculated as follows:

$$X_{vi}^{g} = \kappa_{vi} \pi_g \alpha_{vg}. \quad (5)$$

Here, X^g_{vi} is the element of subclass data matrix X^g with the class prevalence π_g. By LPA and CFA, the number of subclasses is estimated as G, and thus the data will be divided into G subclasses. In this model, the sum of X^g_{vi} is equal to the former data, which were equal to x_{vi}.

2.5. Iteration algorithm

Since the classified subclasses still have heterogeneity, an iteration algorithm to classify the subclass data into smaller subclasses was developed, as follows:

Step 1: Application of LPA with all possible numbers of classes

In LPA, the range of the class-parameter number is from two to the number of samples. Thus, we applied LPA with all possible numbers as class-parameters. The Structural Equation Modelling (SEM) software Mplus is utilized to apply LPA. The quality of classification by LPA is estimated as entropy static in Mplus, and thus we evaluated the set class-parameter numbers by the entropy static values. Values of entropy static close to 1 indicate high classification accuracy, whereas values close to 0 indicate low classification certainty. In this study, we only used the entropy static value equal to 1.

Step 2: Application of CFA with all possible numbers of latent variables

In CFA, the number of latent variables should be smaller than the number of samples. We employed the CFA models with the combinations of all possible numbers of latent variables, all possible factoring methods, and all possible rotation methods.

Step 3: Comparison of the results of LPA and the classified variables in CFA

By comparing the absolute values of the factor loadings in CFA, each sample can be classified into one specific latent group. The ratio of classified samples is compared with the class prevalence in LPA for the same number of subclasses. When the classified ratio in CFA and the class prevalence in LPA are similar, the subclass number is considered to be suitable. If the results of CFA and LPA are not similar in any cases, then we decide that the data could not be separated.

Step 4: Calculation of subclass data

According to section 2.4, each data subclass is calculated as a $v \times I$ matrix. In this step, the individual weight κ_{vi} of the observed data is calculated from the individual data x_{vi}, the class prevalence π_g, and the partial regression α_{vg}. After the individual weight κ_{vi} is calculated, the individual data x_v is divided into G different numerical data by multiplication for partial regression weight and class prevalence.

Step 5: Return to step 1 for classification of subclass data

From Step 1 to Step 4, the original data can be divided into some smaller subclasses, but not sufficiently. To classify the data depending on their homogeneity, these classification steps are iterated for each subclass data matrix, which is calculated in Step 4. The stopping criterion in this iteration algorithm is that the LPA results could not fit the CFA results.

3. Results & Discussion

We applied our developed method to simulated data and some omics data compiled from the GEO database (http://www.ncbi.nlm.nih.gov/geo/). To evaluate the classification ability of our developed method, we created 4 sample datasets composed of 5 different Gaussian distributions. In this simulation, the individual effect κ_{vi} was not set for each data point, but the group effect κ_v was set for each sample. The 4 created samples were divided into 6 different subclasses, which is a larger number than the set of different types, in advance.

The omics data multiply measured under the same conditions were compiled for our evaluation. In the real data, the samples were classified into more subclasses than the measured points. Interestingly, the LPA results tended to be similar to the CFA results with no rotation. For both data types, our method facilitated division into groups depending on the latent effects, without a sample number constraint. Our method is useful to clarify all types of omics data, including transcriptome, proteome and metabolic information.

4. References

[1] Chu Y and Corey D R 2012 *Nucleic Acid Ther.* **22(4)** 271
[2] Mirza S P and Olivier M 2008 *Physiol. Genomics* **33(1)** 3
[3] Griffin J L and Vidal-Puig A 2008 *Physiol. Genomics* **34(1)** 1–5
[4] Altschuler S J and Wu L F 2010 *Cell* **141(4)** 559
[5] Newman J R, Ghaemmaghami S, Ihmels J, Breslow D K, Noble M, DeRisi J L and Weissman J S 2006 *Nature* **441(7095)** 840
[6] Torres-Padilla M E and Chambers I 2014 *Development* **141(11)** 2173
[7] Aburatani S 2012 Bioinformation 8(14) 652
[8] Bartholomew D J, Steel F, Moustaki I and Galbraith J I 2002 *The Analysis and Interpretation of Multivariate Data for Social Scientists* 2nd ed. (London: Chapman and Hall/CRC Press)
[9] Geiser C 2010 *DATA ANALYSIS WITH MPLUS* (New York: The Guilford Press)
[10] Spirtes P, Glymour C and Scheines R 2001 *Causation, Prediction, and Search* 2nd ed. Cambridge: The MIT Press