Genotyping of single nucleotide polymorphism using model-based clustering

H. Fujisawa1,∗, S. Eguchi1, M. Ushijima2, S. Miyata2, Y. Miki2, T. Muto2 and M. Matsuura2

1Institute of Statistical Mathematics, Tokyo 106-8569, Japan and 2Genome Center, Japanese Foundation for Cancer Research, Tokyo 170-8455, Japan

ABSTRACT

Motivation: Single nucleotide polymorphisms have been investigated as biological markers and the representative high-throughput genotyping method is a combination of the Invader assay and a statistical clustering method. A typical statistical clustering method is the k-means method, but it often fails because of the lack of flexibility. An alternative fast and reliable method is therefore desirable.

Results: This paper proposes a model-based clustering method using a normal mixture model and a well-conceived penalized likelihood. The proposed method can judge unclear genotypings to be re-examined and also work well even when the number of clusters is unknown. Some results are illustrated and then satisfactory genotypings are shown. Even when the conventional maximum likelihood method and the typical k-means clustering method failed, the proposed method succeeded.

Contact: fujisawa@ism.ac.jp

INTRODUCTION

Single nucleotide polymorphisms (SNPs) have been investigated as biological markers to identify the genes relevant to diseases and adverse effects. A great deal of effort has been put into research to find new SNPs. Many SNPs have been discovered and used for association studies in humans. Such studies require a fast and reliable genotyping method for each SNP. One high-throughput genotyping method is a combination of the Invader assay (Third Wave Technologies, Madison, WI, USA) and the k-means clustering method (Olivier et al., 2002). However, the k-means method is not always effective. This paper proposes a model-based clustering method using a normal mixture model and a penalized likelihood to overcome some problems.

In experiment using the Invader assay, we can examine many individuals simultaneously for each SNP and observe two fluorescence intensities for each individual, as shown in Figure 1, in which a point corresponds to an individual and its coordinates indicate two observed fluorescence intensities, respectively. The most typical example is Figure 1(a). The observations can clearly be clustered into four groups near the origin and along the x-axis, the y-axis, and the line with the angle π/4. The last three groups correspond to three genotypes and such clustering is essential for genotyping. We can also see different typical examples in Figure 1, but the observations can only roughly be clustered but not clearly.

The clustering of individuals for genotyping was considered in Ranade et al. (2001) and Olivier et al. (2002). The k-means method is a typical clustering method and sometimes provides successful results when the groups are clearly separated and the number of groups is known. However, the k-means method has some drawbacks: it is not effective for handling different within-variations (variations within each cluster) and for finding outliers. In addition, it is necessary to fix the number of groups a priori.

The proposed clustering method is based on the normal mixture model (McLachlan and Peel, 2000), in which the observations are arranged into appropriate angle data. The importance of the angle data was illustrated in Mein et al. (2000). The heteroscedasticity of the model can handle different within-variations. The outlier can be detected using the confidence region. Using the Bayes rule, the observation is assigned to a certain group to be genotyped or is regarded as an unclear genotype. The detection of the outlier and the unclear genotype is very important and helpful for experimenters, so that such observations can be re-examined. The remaining problem is that a simple application of the model generally forces the number of groups to be fixed a priori, as in the k-means method. We can propose a method for identifying the number of groups (e.g. cross-validation), but this would require a great deal of effort and would not be reliable. The method described in this paper overcomes this problem using typical characteristics of the Invader assay. The examples imply typical angles, e.g. one of which is π/4. Such typical characteristics allow a penalized likelihood approach using the Kullback–Leibler divergence as a penalty function. This concept eliminates the requirement for prior identification of the number of groups beyond a conventional meaning of the penalty. The proposed method works well even when
some groups have no observation or only one observation. Furthermore, it is known that the usual likelihood method is not robust to outliers. A modification proposed here can achieve robustness to outliers.

The model-based clustering has recently been applied to microarray expression data in Yeung et al. (2001), McLachlan et al. (2002), and Ghosh and Chinnaiyan (2002). The penalized likelihood method has been employed in Chen and Kalbfleisch (1996), Ormoneit and Tresp (1998), Chen et al. (2001) and Eguchi and Yoshioka (2001). The characteristic aspect of the proposed method is the assignment of a penalty only to the normal components but not to the proportional parameters, differently from the past methods. See Hampe et al. (2001) and Riva and Kohane (2002) for recent research into SNPs.

This paper is organized as follows. First, the Invader assay and motivated examples are explained. Next, the method is constructed and then analyses of the examples are demonstrated. Finally, extensive variations of the proposed method are discussed.

**MOTIVATION**

**Invader assay**

We assume that there are two types of alleles for each SNP, which are denoted symbolically by 1 and 2. An individual has two alleles for each SNP and therefore the genotype of the individual is regarded as 1/1, 1/2 or 2/2, because the genotype is a combination of two alleles. Based on the Invader assay (Ohnishi et al., 2001; Olivier et al., 2002), we can know what alleles the individual has and hence decide the genotype of the individual. In the following, we give a simple explanation on the Invader assay and specify its characteristics.

We prepare two allele-specific probes, two probes labeled with two fluorescence dyes, and so on. We react them on DNAs of the individuals for each SNP. If the individual has an allele, the corresponding allele-specific probe responds to the allele and then we observe an expression of the corresponding fluorescence. Consequently, we can know what alleles the individual has and hence decide the genotype of the individual.

There are many factors affecting a fluorescence intensity. The fluorescence intensity depends on the velocity of the invader response of the relevant probe reaction. Different response velocities are due to different amounts of DNA and/or impurity in each sample and different invader reagents of DNA. Experimenters may adjust the response time to obtain a sufficient response and a relatively high intensity after obtaining a poor response. The response time affects the fluorescence intensity and furthermore results based on a few different response times might be combined into a plane data. Extra untrue expressions are usually observed for some reasons, which is referred to as a background response,

**Fig. 1.** Two fluorescence intensities of individuals for six SNPs.
although the velocity of the background response is slower than that of the normal invader response. To measure a background response, we often react the probes not only on DNAs but also on blanks, which are referred to as no-target blanks. In spite of many disturbed factors, we can say that a strong fluorescence intensity indicates a strong possibility of having a specific allele.

**Example**

The Invader assay generates two fluorescence intensities for individuals. The examples of six SNPs are shown in Figure 1, where the \(x\)-axis and the \(y\)-axis correspond to two fluorescence intensities, respectively. When the \(x\) (or \(y\)) value is sufficiently large, then the individual will have allele 1 (or 2). The individual in the lower-right part along the \(x\)-axis is genotyped as 1/1 because the \(x\)-value is large and the \(y\)-value is small. Similarly, the individuals in the upper-right part and the upper-left part along the line with the angle \(\pi/4\) and the \(y\)-axis are genotyped as 1/2 and 2/2, respectively. They are referred to as groups 1, 2 and 3, respectively. The angles of the observations belonging to their groups are ideally 0, \(\pi/4\) and \(\pi/2\), respectively. The reason for this is that when the individual has only one of two alleles, ideally there is an expression of the corresponding allele but no expression of the other allele. The ideal situation might hold if there should be no background response, although the observations in fact show various variations due to many factors as described already. The lower-left part near the origin has a different meaning. An observation near the origin results from a no-target blank or from a failure of invader response. This is referred to as group 0. The observations as in Figure 1 have been sampled for a wide range over SNPs in the Japanese Foundation for Cancer Research (JFCR), based on informed consent.

The most typical example of the data generated by the Invader assay is given in Figure 1(a). The clustering of the observations will be easy since four groups are clearly separated. Figure 1(a) illustrates some characteristics of the observations. Let the variations in the \(x\)-direction and the \(y\)-direction be named by the \(x\)-variation and the \(y\)-variation, respectively. Group 1 exists along the \(x\)-axis and its \(x\)-variation is large but its \(y\)-variation is small, because the \(x\)-variation results from a normal invader response but the \(y\)-variation results from a weaker background response. Similar results hold for group 3. Group 2 has both variations and hence is widely spread. There may be a tendency for the ranges of heterozygous group 2 to resemble the ranges for the \(x\)-direction of homozygous group 1 and the \(y\)-direction of homozygous group 3. This is not always true, because results based on a few different response times might be combined.

The remaining examples in Figure 1 are also typical but imply difficult aspects of clustering. In Figure 1(b), group 3 along the \(y\)-axis moves to group 2 as the \(y\)-value increases, because of the background response. Figure 1(c) shows a tilted case, in which the angle of group 2 deviates slightly from the standard angle \(\pi/4\) because the response velocity of allele 2 along the \(y\)-axis is slower than the other. In Figure 1(d), the background response along the \(x\)-axis is larger than the usual and the observation near \((1.5, 1.5) \times 10^4\) should not be assigned to group 2 or 3. The observations can only be clustered roughly but not clearly. In Figure 1(e), no homozygous group 3 along the \(y\)-axis is observed because an allele frequency may be relatively small compared with the sample size. In Figure 1(f), the number of observations is not enough. These examples present some commonly encountered characteristics. The within-variations are very different. The observations near the origin belonging to group 0 are not always far from the other groups. The angles of the observations allow a rough clustering.

**METHOD**

The purpose of this paper is to propose a fast and reliable method for clustering the observations. Various situations and problems have been illustrated already. One applicable clustering method is a model-based clustering using the normal mixture model, but a simple application of this method cannot overcome all of the problems. This section proposes a modification of the method, using a penalized likelihood.

**Adjustment of data**

Before applying the normal mixture model, we adjust the observations. As stated already, the angles of the observations are important information. The observations will be clustered on the basis of angle data only. There are other reasons for using only the angle data. The length of the observation depends on many factors, including an artificial and uncontrollable adjustment of response time and therefore the use of the length will be uncertain. The scatter of the angle data is more suitable for normality of the model than the scatter of the original data. Other adjustments are discussed later.

The origin should not be defined as zero when we use the angle data, as shown in Figure 1(d). In this paper, the origin is adaptively selected by using the following concept. There will be two candidates for the origin. If the no-target blanks are known a priori, then an appropriate origin will be located at the center of them. If not, there will be no definitive way to select the origin. For convenience, this paper employs a pair of the minimums of the \(x\)-values and the \(y\)-values as the origin of the angle because no-target blanks usually exist for each assay. This method for setting the origin was founded to work well, as described later.

We can empirically regard an observation of length of less than \(10^4\) as a no-target blank or as a failure of the invader response. In the following, such observations are not taken into consideration. This concept has little influence on the genotyping because such observations belong to group 0.
Normal mixture model

Let $x$ be a random observation and let $z = (z_j)$ be a $k$-dimensional label where the $j$-th element is one and the other elements are zero when $x$ is sampled from the group $j$ according to $N(\mu_j, \sigma_j^2)$ for $j = 1, \ldots, k$. In the context of the genotyping, $x$ is the angle of observation, $j$ is the number of genotype group, and $k = 3$. Let $\omega_j$ be the proportional probability, where $x$ is sampled from the group $j$, and therefore $\sum_{j=1}^{k} \omega_j = 1$. Let $\phi(x; \xi)$ be the normal density function, where $\xi = (\mu, \sigma^2)$ with mean $\mu$ and variance $\sigma^2$. The probability density function of $(x, z)$ is given by

$$f(x, z; \theta) = \frac{1}{(2\pi)^{d/2}} \prod_{j=1}^{k} \omega_j \phi(x; \xi_j)^z_j,$$

where $\xi_j = (\mu_j, \sigma_j^2)$ and $\theta$ is the parameter vector of $\omega, \mu_j, \sigma_j$. Note that the variable $z$ is unobservable in the context of the genotyping. Summing out unknown labels, we can get the following probability density function:

$$f(x; \theta) = \sum_z f(x, z; \theta) = \sum_{j=1}^{k} \omega_j \phi(x; \xi_j),$$

which is called a normal mixture model.

Suppose that the parameters are known. The observations can be clustered by the Bayes rule. Let $p_j$ be the posterior probability given that the observation $x$ belongs to the group $j$ for $j = 1, \ldots, k$:

$$p_j = P(z_j = 1 | x) = \frac{\omega_j \phi(x; \xi_j)}{f(x; \theta)}.$$

The observation $x$ is assigned to the group $j$ if $p_j$ is the maximum of $p_j$s or if $p_j$ is more than a certain threshold probability. In the latter case, some observations are not clearly assigned to any group. The outlier can be detected on the basis of confidence interval, for which this paper adopts the simple view that if the observation is outside all confidence intervals of the normal components, then the observation is judged to be an outlier.

The parameter will be replaced by an estimate because the parameter is generally unknown. The estimation method is proposed in the next subsection.

Estimation method

The key to the application of the normal mixture model is an estimation method used for the parameter. A standard method is the maximum likelihood estimate. Remember that the number of groups is not always known. If the maximum likelihood method is directly applied to the data with an unnecessarily large number of groups, then the method will force a group to be divided into some groups or give no solution. This is a well-known drawback. The number of groups can be selected a priori, e.g. by way of model selection although it is often unstable and requires a great deal of effort. Furthermore, the maximum likelihood method is not robust to outliers. The following method improves these drawbacks by using prior information.

Suppose that the number of groups is known to be at most $k$ and that these groups have characteristic parameters on the normal components. This paper takes them into consideration and proposes a simple modification of the maximum likelihood method with the penalty. Let the log-likelihood function and the Kullback–Leibler divergence be denoted by

$$l(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(x_i; \theta),$$

$$KL(\xi_0, \xi) = \int \phi(z; \xi_0) \log \frac{\phi(z; \xi_0)}{\phi(z; \xi)} dz.$$

Let $\xi_{j0} = (\mu_{j0}, \sigma_{j0}^2)$ be the characteristic parameter of the $j$-th group. The penalized likelihood function is given by

$$PL(\theta) = l(\theta) - \frac{\lambda}{n} \sum_{j=1}^{k} KL(\xi_{j0}, \xi_j).$$

This formula is often seen in the framework of conjugate prior. Note that the label is assumed to be known on the penalty term, differently from the likelihood term (Eguchi and Yoshioka, 2001). The characteristic of the penalty is to assign no penalty to the proportional probabilities $\omega_j$s, differently from the past methods. The estimate of $\theta$ is defined as the maximizer of the penalized likelihood function:

$$\hat{\theta} = \arg \max_{\theta} PL(\theta).$$

The estimate will yield a large value for the likelihood function $l(\theta)$ in the neighborhood of the characteristic parameter. The tuning parameter $\lambda$ is set a priori in this paper, which is discussed later.

The above penalty shows us another advantage beyond a conventional meaning of the penalty. The penalized likelihood method eliminates the requirement for prior identification of the number of groups and allows us to make an appropriate estimate, even if the true number of groups is less than $k$. For example, consider the case where $k = 3$, group 3 has no member, the groups are separated, and the parameters $\xi_j$s lie in the neighborhood of the characteristic parameters. The estimate of $\omega_3$ should be zero. It can be shown that the estimate is generally $\omega_3 = 0$ and $\xi_3 = \xi_{30}$. The reason for this is as follows. Let $x$ be the observation belonging to group 1 or 2, since group 3 has no member. If the likelihood $\phi(x; \xi_3)$ is much smaller than $\phi(x; \xi_1)$ (or $\phi(x; \xi_2)$), then the penalized likelihood $PL(\theta)$ is maximized at $\omega_3 = 0$ because

$$\sum_{j=1}^{3} \omega_j \phi(x; \xi_j) \leq (\omega_1 + \omega_3) \phi(x; \xi_1) + \omega_2 \phi(x; \xi_2).$$
Other cases can be demonstrated in a similar way. It holds that the PL(θ) depends on ξ3 only in KL(ξ30, ξ3) when ω3 = 0, which implies that the PL(θ) is maximized at ξ3 = ξ30. This phenomenon also holds for population.

The estimate ϑ is not robust to outliers because it is based on the conventional log-likelihood. The robustness can be accomplished by virtue of the modified likelihood (Basu et al., 1998; Jones et al., 2001):

\[ l_{β}(θ) = \frac{1}{nβ} \sum_{i=1}^{n} f(x_i; θ)^β - b_{β}(θ), \]

where β > 0 and

\[ b_{β}(θ) = \frac{1}{1+β} \int f(x; θ)^{1+β} dx. \]

It may be noted that the cases β = 0 and β = 1 correspond to the conventional log-likelihood and the empirical L2-distance of the probability density functions, respectively. The latter is known to give a strong robust estimation with less efficiency. The penalized likelihood can be robustified by replacing the conventional log-likelihood l(θ) with the modified likelihood l_{β}(θ). This paper adopts β = 0.2 because Fujisawa and Eguchi (2002) reported that the default β = 0.2 performs with satisfactory robustness and efficiency in the normal mixture model.

EM algorithm

The penalized likelihood method has been described above, When there is no penalty, the convenient EM algorithm can be constructed (McLachlan and Krishnan, 1997). Even if we add the penalty described above, a similar update algorithm can be proposed in a similar way to find the convenient EM algorithm:

\[ \omega_j^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} z_{ij}^{(r)}, \]

\[ μ_j^{(r+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(r)} x_i + λμ_j}{\sum_{i=1}^{n} z_{ij}^{(r)} + λ}, \]

\[ (σ_j^{(r+1)})^2 = \frac{\sum_{i=1}^{n} z_{ij}^{(r)} (x_i - λμ_j)^2 + λ(σ_j^2 + μ_j^2)}{\sum_{i=1}^{n} z_{ij}^{(r)} + λ} - (μ_j^{(r+1)})^2, \]

where

\[ z_{ij}^{(r)} = \frac{ω_j^{(r)} ϕ(x_i; ξ_j^{(r)})}{f(x_i; ξ_j^{(r)}).} \]

This update rule implies the favorable monotone property

\[ PL(θ^{(r+1)}) ≥ PL(θ^{(r)}). \]

The proof is given in the appendix. The update step for normal parameters is different from that for the conventional EM algorithm. A similar discussion is presented in Ormoneit and Tresp (1998).

When the conventional log-likelihood l(θ) is replaced by the modified likelihood l_{β}(θ), it seems to be difficult to construct an update algorithm with a monotone property. When there is no penalty, Fujisawa and Eguchi (2002) proposed an update algorithm and investigated its favorable properties. The same concept yields a similar update algorithm to the above, as given in the appendix.

DATA ANALYSIS

In this section, the examples are analyzed using the proposed method. Remember the flow of the analysis with Figure 2. The original observations are displayed in Figure 2(a). The observations are transformed into angle data and the normal mixture model is applied to the angle data, as shown in Figure 2(b). The data are clustered using the Bayes rule, as stated in Figure 2(c). The clustered data are restored into two-dimensional data in Figure 2(d), which is the result of clustering. The numbers 1, 2 and 3 denote the corresponding group numbers to which the observation belongs. Points are neglected data for clustering because the fluorescence intensity is excessively small. The numbers 7 and 9 indicate the outlier and the unspecified observation, respectively, which are identified using the confidence interval and the Bayes rule, respectively. The outlier detection was performed before the application of the Bayes rule.

The proposed method requires that the tuning parameter λ, the typical parameter ξ0, the threshold value of the Bayes rule, and the confidence level be chosen. The first three values are set to 1, an empirical value, and 0.9, respectively, and the confidence levels are set to 0.999, 0.95 and 0.999 for three groups, respectively. Details of the reason for these assignments are discussed later.

All of the results, as shown in Figure 3, indicate that the proposed method works well. Figures 3(a) and (b) show that the proposed method can handle different within-variations. The k-means method will fail for this typical data, as described later. Figure 3(c) indicates that the proposed method can pursue a tilted group. Although it is generally difficult to type the data near the origin, more precisely, near (1.2, 0.2) \times 10^4, the transformation to the angle data overcomes this problem. Figure 3(d) illustrates that the proposed method responds to a change in the origin and points out the outlier indicated as the symbol 7 near (1.5, 1.5) \times 10^4. This outlier should be re-examined in the next assay. It is very important and helpful for experimenters to detect such doubtful points as outliers. Figure 3(e) demonstrates that the proposed method can work even if the number of groups is less than three. When the maximum likelihood method with k = 3 was applied, the method forced group 2 to be divided into two groups. The lack of the group as shown in Figure 3(e) happens when an allele frequency is relatively small compared with the sample size. The proposed method can adequately deal with such a situation. Figure 3(f) demonstrates that the proposed method can work even if the number of observations is very small. The
Fig. 2. The flow of analysis: (a) original observations; (b) transformed angle data and applied normal mixture model; (c) Clustering result; (d) restored result.

Fig. 3. Clustering results of Figure 1.
outlier near \((3, 1) \times 10^4\) was correctly detected by the proposed method. When the robust method was not adopted, the variance of group 2 was overestimated because of the outlier, so that the outlier was incorrectly assigned to group 2. In Figures 3(a) and (d), there are questionable outliers and unclustered data, which is discussed later.

The \(k\)-means method was also applied to the examples. As stated already, the number of groups must be identified a priori before applying the \(k\)-means method. Even if the number of groups is correctly set, the clustering often fails; e.g., when the \(k\)-means method was applied to Figure 1(d), the observations near \((2.5, 0.8) \times 10^4\) were incorrectly assigned to group 1 because of different within-variations (Fig. 4).

DISCUSSION

The proposed method was applied to the present examples and many other data which were provided for genomic research in the JFCR, and satisfactory results were obtained. Some extensive variations of the proposed method are discussed in this section.

Some values constructing the proposed method are set a priori without a detailed explanation. Let us discuss other possibilities. The tuning parameter \(\lambda\) was set to be one. It was investigated for \(\lambda = 0.1, 1\). If the penalty is merely thought of in the conventional meaning, a change in the parameter \(\lambda\) will largely affect the clustering results. However, the results were only slightly affected by a change of the parameter \(\lambda\). This follows from the true meaning of the penalty, because the largest contribution made by the penalty is in eliminating the requirement for prior identification of the number of groups. The characteristic parameter \(\xi_0\) was empirically set; \(\xi_{10} = (0.025, 0.05^2)\), \(\xi_{20} = (0.726, 0.05^2)\), \(\xi_{20} = (1.38, 0.05^2)\). Changes in the empirical values only slightly affected the clustering results (not essentially). This is true for the same reason as is the case for the tuning parameter \(\lambda\). The threshold value of the Bayes rule was set to 0.9. A high threshold value of the Bayes rule is favorable because it expresses a degree of confidence. However, an excessively high threshold value was not appropriate for some examples. For example, assume that group 2 has an excessively large within-variation and then the tail of the normal distribution of group 2 becomes long. This implies that the posterior probability of the member of groups 1 and 3 cannot be high. In fact, when the threshold value was 0.999, the members of group 1 in Figure 1(d) were not assigned to any group. After the value 0.9 was empirically selected, the questionable clustering was not observed. The confidence levels were set to 0.999, 0.95 and 0.999 for groups 1, 2 and 3, respectively. Groups 1 and 3 tended to have a small within-variation. When the confidence level was set to 0.95 in common, unfavorable outliers appeared as a result of excessively small confidence regions.

The various values described above can be changed according to the needs of the researcher. If the researcher demands more confidential results, then one can set a larger threshold value of the Bayes rule and a smaller confidence level, although this increases the observations which should be re-examined, and the converse is also true. The above setting is sometimes severe when the sample size is small.

Reconsider Figure 3(e) and apply the normal mixture model with three components to it, using the conventional maximum likelihood method, which gives Figure 5. (Note that the threshold value of the Bayes rule was set to 0.7 for a helpful understanding. When it was 0.9, all of the numbers 3 were changed to 9.) This example shows an interesting fact. We do not think this figure is good for genotyping because the typical angles are not supported. However, if we have no prior knowledge like typical angles, then it seems to be difficult to deny the normal mixture model with three components, which illustrates an importance of using the prior knowledge.

If the number of groups were to be known and set correctly (to less than three), the conventional maximum likelihood method might appear better than the proposed method.
However, the proposed method has another merit in that the proposed EM algorithm offers a more rapid convergence than the conventional EM algorithm. This will result from the inclusion of prior knowledge.

This paper has dealt with cases where a sensible clustering is available. In some experiments, unclear shapes were displayed, e.g. only one mass, which results from a doubtless failure of the experiment. When the algorithm fails to converge or the overlap of the components is large, we can conclude that the corresponding SNP should be re-examined.

We often use a log-transformation to adjust a bias of positive data. We tried various log-transformations including $\log(x)\log(y), \log\tan^{-1}(y/x), (\log(x+a), \log(y+b))$, where $a$ and $b$ are appropriate constants. However, we could not obtain any effective results and therefore we have used the raw angle data.

ACKNOWLEDGEMENTS

The authors would like to thank M.D. Minoru Isomura for his valuable suggestions and comments. This work was supported by Grant-in-Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science and Technology and by grants from New Energy and Industrial Technology Development Organization.

REFERENCES


APPENDIX

EM algorithm for the penalized likelihood

Let the penalized likelihood based on the labeled data be denoted by

$$PL_{\lambda}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(x_i, z_i; \theta) - \frac{\lambda}{n} \sum_{j=1}^{k} KL(\xi_j \parallel \xi_j),$$

and let

$$Q(\theta; \theta') = E \left[ PL_{\lambda}(\theta) \left| \prod_{i=1}^{n} f(z_i | x_i; \theta') \right. \right],$$

where

$$H(\theta; \theta') = \frac{1}{n} \sum_{i=1}^{n} E[\log f(z_i | x_i; \theta) | f(z_i | x_i; \theta')].$$

Genotyping of SNPs using model-based clustering
Assume that $Q(\theta''; \theta') \geq Q(\theta'; \theta')$. It then follows that

$$PL(\theta'') - PL(\theta') = Q(\theta''; \theta') - Q(\theta'; \theta') + H(\theta'; \theta') - H(\theta''; \theta')$$

and $H(\theta'; \theta') - H(\theta''; \theta')$ becomes the Kullback–Leibler divergence. It is easy to show that

$$\theta^{(r+1)} = \arg \max_{\theta} Q(\theta; \theta^{(r)})$$

It holds that $PL(\theta^{(r+1)}) \geq PL(\theta^{(r)})$. The proof of the monotone property is complete.

Update algorithm based on the penalized modified likelihood

Let

$$c_{jm}(\theta) = \int \omega_j s_{jm} \phi(x; \mu_j, \sigma^2_j) f(x; \theta) \phi \, dx,$$

where $s_{j0} = 1, s_{j1} = x - \mu_j, s_{j2} = x^2 - (\sigma^2_j + \mu^2_j)$. Let

$$z_{ij}^{(0r)} = z_{ij}^{(r)} - c_{j0}(\theta^{(r)}) + \sum_{j=1}^{k} \omega_j^{(r)} c_{j0}(\theta^{(r)})$$

and

$$z_{ij}^{(mr)} = x_i^m z_{ij}^{(r)} - c_{jm}(\theta^{(r)}), \quad m = 1, 2.$$

The algorithm is given by

$$\omega_j^{(r+1)} = \frac{\omega_j^{(r)} \phi(x_i; \beta_j^{(r)})}{\sum_{j=1}^{k} \sum_{i=1}^{n} z_{ij}^{(r)}},$$

$$\mu_j^{(r+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(1r)} + \lambda \mu_j^{(0)}}{\sum_{i=1}^{n} z_{ij}^{(1r)} + \lambda},$$

$$\sigma^2_j^{(r+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(2r)} + \lambda (\sigma^2_j^{(0)} + \mu^2_j^{(0)})}{\sum_{i=1}^{n} z_{ij}^{(2r)} + \lambda} - (\mu_j^{(r+1)})^2.$$