Published in final edited form as:
Comput Stat Data Anal. 2014 September 1; 77: 146–156. doi:10.1016/j.csda.2014.02.017.

Variable Assessment in Latent Class Models

Q. Zhang[*] and **E. H. Ip**
Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston Salem, NC, USA

Abstract

The latent class model provides an important platform for jointly modeling mixed-mode data — i.e., discrete and continuous data with various parametric distributions. Multiple mixed-mode variables are used to cluster subjects into latent classes. While the mixed-mode latent class analysis is a powerful tool for statisticians, few studies are focused on assessing the contribution of mixed-mode variables in discriminating latent classes. Novel measures are derived for assessing both absolute and relative impacts of mixed-mode variables in latent class analysis. Specifically, the expected posterior gradient and the Kolmogorov variation of the posterior distribution, as well as related properties are studied. Numerical results are presented to illustrate the measures.

Keywords

Latent class analysis; variable selection; mixed data type; total variation; posterior gradient; cross entropy; Kolmogorov distance

1. Introduction

Heterogeneous data types have become commonplace in many sciences. In the medical sciences, clinical studies often collect data that are continuous (e.g., blood pressure), binary (whether or not the subject has diabetes), ordinal (severity level of a disease), categorical (medication used), and other types such as count and time-to-event. The identification of clinically meaningful phenotypes in the population using a heterogeneous data type is thus an important area of research. Everitt (1988,1993) [9, 10] referred to heterogeneous data types as mixed-mode data in the context of latent class and mixture analysis, in which multiple data types are used as indicators for putting similar objects into groups [see also 14, 24]. The terms latent class model and mixture are used interchangeably. The idea here is to cluster a vector of mixed-mode responses $\mathbf{Y} = (Y_i)$ for indicators $i = 1, \ldots, m$ into S distinct latent classes $Z = 1, \ldots, S$. There are at least two general approaches for mixed-mode latent

---

[*] Department of Biostatistical Sciences, Wake Forest School of Medicine, Medical Center Blvd, Winston Salem, NC 27157, USA. Phone: (336) 716-5141; Fax: (336) 716-6427. qizhang@wakehealth.edu (Q. Zhang), eip@wakehealth.edu (E. H. Ip).

[☆] Some mathematical derivation results are included as annexes in the electronic version of this article.

class analysis (MM-LCA). The first approach is to relate the manifest categorical response to an underlying multivariate Gaussian distribution such that continuous normal variables and categorical variables can be jointly modeled [13, 23]. As pointed out by [7], the underlying Gaussian approach has limitations, one of which is that it cannot easily accommodate general data types such as counts. An alternative is to use the generalized linear mixed-model approach proposed by [22] and later extended by [17], [7], [5], [25], and [3]. This approach can accommodate any mixture of outcomes from an exponential family. Under the assumption of conditional independence given latent class Z, the likelihood for an individual subject in an MM-LCA can be expressed as:

$$f(\boldsymbol{Y}|\theta)=\sum_{z=1}^{S}\alpha_z\prod_{i=1}^{m}p_{iz}(y_i|\theta_z), \quad (1)$$

where θ contains the vector of parameters θ_z for each individual class z, which has a prior probability $\alpha_z = p(Z = z)$. Within an exponential family framework, different link functions can be specified for the conditional distribution p_{iz} for different data types.

One question that arises from the generalized mixed-model approach for latent class analysis and latent variable in general is how the different types of data “impact” the likelihood. It is possible that one data type “overwhelms” another data type in the likelihood and becomes dominant in defining the structure of the latent class model. Because data values are not measured on the same scale, it is not easy to promptly assess the impact of a variable on the overall likelihood. This question is directly related to a second question: if only a limited number of mixed-mode indicators can be included in a latent class analysis, which variables should be selected for maximally “discriminating” between the classes? Interestingly, the latter question can also be reformulated as a variable-selection problem and solved by a search algorithm using criteria such as the BIC [20, 6].

Two measures are proposed for assessing a variable’s contribution to the classification of latent classes. In LCA class labels are not known a priori; the term classification here refers to the extent to which a variable contributes to discriminating the classes, or in the case the class label is known (e.g., in a simulation setting) the accuracy in retrieving class membership. The first measure, the expected posterior gradient (EPG), measures the absolute contribution of a variable to MM-LCA. The second measure, based on the Komolgorov variation of the posterior distribution (KVP), can be interpreted in terms of the relative contribution of a variable by comparing classification accuracies with and without the variable in the MM-LCA. Interestingly, both measures can be related to the statistical distance between the prior distribution $p(z)$ and the posterior distribution $p(z|y)$. There are several advantages in using the EPG and KVP. First, they both have strong theoretical foundations, which will be described in the following two sections under the heading “Justification of measures.” Second, the measures can be universally applied to all kinds of mixed-mode data - continuous, discrete, and count data. Furthermore, computationally the two measures are straightforward to compute and closed form solutions are available for EPG. For the remainder of the paper, Section 2 describes the EPG measure, and the procedure of how the measure can be derived and used in practice. Section 3 describes KVP and specifically its relation to the total variation measure, which is commonly used in the image processing literature. In Section 4, two numerical examples of MM-LCA are provided to illustrate the proposed methods. A brief discussion is given in Section 5.

2. Expected posterior gradient for variable assessment

Consider an S-class latent class model that includes both continuous and discrete random variables, $Y = (Y_1, \ldots, Y_m)$, with class-conditional distributions of normal, exponential, Gamma, Poisson, ordinal, or binomial distributions, given the latent random variable $Z \in S$, $S = \{1, \ldots, S\}$. Class-conditional independence is assumed among all the variables — i.e.,

$$p(Y_1=y_1,\ldots,Y_m=y_m|Z=z)=\prod_{i=1}^{m}p(Y_i=y_i|Z=z). \quad (2)$$

Denote the posterior probability $p(Z = z|Y_1 = y_1, \ldots, Y_m = y_m)$ by τ_z, the class-conditional probability $p(Y_i = y_i|Z = z)$ by $\pi_{y|z}=\prod_{i=1}^{m}\pi_{y_i|z}$. These quantities are related by the Bayes formula:

$$\tau_z=\frac{\alpha_z\pi_{y|z}}{p(y)}, \quad (3)$$

where $p(y)$ is the marginal probability of observing the outcome vector, $y = (y_1, \ldots, y_m)$, and

$$p(y)=\sum_{z=1}^{S}\alpha_z\pi_{y|z}. \quad (4)$$

The EPG measure for assessing the impact of variable y_i on the MM-LCA is denoted by B_i and its definition is given by:

$$B_i=\sum_{z\in\mathscr{S}}\alpha_z\left|E_y\left(\frac{\partial\log(\tau_z)}{\partial y_i}\right)\right|. \quad (5)$$

The idea behind EPG is that when a change in a specific variable has a strong impact on the posterior distribution, it would imply that the variable contains substantial information about the classification of latent class - i.e., the membership of subject given the response pattern. In other words, if class membership of a subject is sensitive to a change in the value of the variable, then the variable has a strong impact. Because the gradient of $\log(\tau_s)$ by y_i is a function of y, the expectation with respect to y in (5) is taken.

A few examples of the term $E(\partial \log(\tau_z))/(\partial_i)$, are presented in closed form for parametric distributions. Taking the logarithm of Eq. (3) and forming the expected partial derivative leads to:

$$E\left(\frac{\partial\log(\tau_z)}{\partial y_i}\right)=E\left(\frac{\partial\log(\pi_{y_i|z})}{\partial y_i}\right)-E\left(\frac{\partial\log(p(y))}{\partial y_i}\right). \quad (6)$$

And from (2) and (4),

$$\frac{\partial p(y)}{\partial y_i}=\sum_{z\in\mathscr{S}}\alpha_z\pi_{y|z}\frac{\partial \log(\pi_{y_i|z})}{\partial y_i}. \quad (7)$$

It is easy to derive the following result for the expected log posterior gradients using the identities (6) and (7). Details of the derivation are given in the supplementary file.

1. For a continuous variable Y_i, without loss of generality, centered at 0, and class-conditional normal distributions, $N(\mu_z, \sigma_z)$,

$$E_y\left(\frac{\partial \log(\tau_z)}{\partial y_i}\right)=\frac{\mu_z}{\sigma_z^2}. \quad (8)$$

2. For a continuous variable Y_i with unconditional mean $\bar{\lambda}$ and class-conditional exponential distributions, $Exp(\lambda_z)$,

$$E_y\left(\frac{\partial \log(\tau_z)}{\partial y_i}\right)=\overline{\lambda}-\lambda_z. \quad (9)$$

3. For a continuous variable Y_i and class-conditional gamma distributions, $G(k_z, \theta_z)$,

$$Ey\left(\frac{\partial \log(\tau_z)}{\partial y_i}\right)=(k_z-1)\sum_{j=1}^{S}\frac{\alpha_j}{(k_j-1)\theta_j}-\frac{1}{\theta_z}. \quad (10)$$

4. For a count variable Y_i with unconditional mean $\bar{\lambda}$ and class-conditional Poisson distributions, $Pois(\lambda_z)$,

$$E_y\left(\frac{\Delta\tau_z}{\tau_z(y_i+1)}\right)=1-\frac{\overline{\lambda}}{\lambda_z}, \quad (11)$$

where $\Delta\tau_z = \tau_z(y_i + 1) - \tau_z(y_i)$, and $\tau_z(y_i) = p(Z = z|Y_i = y_i, i = 1, \ldots, p)$.

5. For a discrete variable Y_i with class-conditional binomial distributions, $B(n, \pi_z)$,

$$Ey\left(\frac{\Delta\tau_z}{\tau_z(y_i+1)}\right)=1-\frac{\overline{O}}{O_z}, \quad (12)$$

where $\overline{O}=\sum_{z=1}^{S}\alpha_z O_z$ and $O_z = \pi_z/(1 - \pi_z)$.

In the univariate mixture model of S normal distributions with equivariance, if $\mu_1 = \ldots = \mu_S$, then $\mu_1 = \ldots = \mu_S = 0$, the degenerating case mentioned in [2]. In such a case, the continuous variable has zero expected contribution to the posterior measure. A similar degenerating case is seen in the Poisson distribution when $\lambda_1 = \ldots = \lambda_S = \bar{\lambda}$, and in the Gamma distribution when $k_1 = \ldots = k_S$ and $\theta_1 = \ldots = \theta_S$.

From (12), it can be seen that if a binary variable has uniform conditional probabilities across all classes — i.e., $\forall_z$, $\pi_{1|z} = c$, then the EPG is zero. On the other hand, if at least one of $\pi_{1|z}$ is close to 0 or 1, the EPG, or the contribution to the reduction in entropy, would be large.

2.1. Justification of the measure

Intriguingly, the EPG is related to the gradient of change of entropy from knowing the prior distribution of the latent class to knowing the distribution of the latent classes after observing the data. For example, if the prior distribution is uniform and the posterior distribution is highly skewed with probability mass concentrated on one particular class, then the change of entropy is large. More formally, denote the cross entropy between two distributions q_1 and q_2 by

$$CH(q_1, q_2) = -\int q_1(x) \log q_2(x) dx. \quad (13)$$

Cross entropy is a statistical distance measure that is directly related to the Kullback-Leibler distance D_{KL} through the following equation:

$$D_{KL}(q_1 \parallel q_2) = H(q_1) + CH(q_1, q_2), \quad (14)$$

where $H(x)$ is the Shannon entropy. Specifically for the prior distribution of $Z, H(Z) = -\sum_{z=1}^{S} \alpha_z \log \alpha_z$.

The cross entropy for the distribution of Z and the distribution of Z given Y is given by

$$CH(Z, Z|Y) = -\sum_{z=1}^{S} \alpha_z \log \tau_z. \quad (15)$$

The quantity $\partial CH(Z,Z|Y)/\partial y_i$ signifies the relative change of cross entropy between the prior distribution $p(Z)$ and the posterior $p(z|y)$ after observing Y with respect to a change in the variable y_i. Thus, the cross entropy gradient quantifies the impact of the variable y_i on the MM-LCA in terms of change in the classification distribution of the latent class variable Z.

Using the Jensen inequality, it is easy to show that the EPG and cross entropy is related by the following inequality:

$$E_y\left(\left|\frac{\partial CH(Z, Z|Y)}{\partial y_i}\right|\right) \leq B_i = \sum_{z \in \mathscr{S}} \alpha_z \left| E_y\left(\frac{\partial \log(\tau_z)}{\partial y_i}\right)\right|, \quad (16)$$

Thus, the EPG is the upper bound of cross entropy gradient. This bound is important for variable assessment because if a variable y_i has small EPG value B_i, then it is not possible for large cross entropy values to exist, and therefore the variable is deemed to have insignificant impact in terms of its discriminatory power for distinguishing between latent classes. In general, the further the class-conditional distribution of y_i is away from a uniform distribution, the more contribution it provides to B_i. Taking absolute value of gradients in (16) is necessary because given a univariate standard-Gaussian variable and $\sigma_z = 1$ for $z \in S$, without the absolute sign, B_i can be zero,

$$B_i = \sum_{z \in \mathscr{S}} \alpha_z \frac{\mu_z}{\sigma_z} = 0, \quad (17)$$

even when class-conditional means, $\{\mu_z, \ldots, \mu_S\}$, are very different.

To illustrate the EPG, a small data set of $n = 1,000$ is simulated, using a 2-class mixture ($S = 2$, $Z = 1, 2$, and $p(Z = 1) = p(Z = 2) = 0.5$) in which Y_1 is a mixture of $N(-1, 1)$ and $N(1, 1)$ respectively for $Z = 1$ and $Z = 2$. Two additional variables Y_2 and Y_3 were independently randomly sampled from $N(0, 1)$ and added to the mix. Then an LCA software, Latent Gold, used to estimate the 2-class model. For Y_1, the class-conditional statistics were $\hat{\mu}_{z=1} = -0.99$, $\hat{\sigma}^2_{z=1} = 1.33, \text{and} \hat{\mu}_{z=2} = 0.99, \hat{\sigma}^2_{z=2} = 0.74$. Therefore, the estimated $B_1 = 0.5 \times (|-0.99/1.33| + |0.99/0.74|) = 1.04$, as opposed to the population value of 1.0. For Y_2 and Y_3, the sample estimates of EPG were respectively $B_2 = 0.015$ and $B_3 = 0.002$, showing that these two variables had little or no impact. In general, the EPG can be estimated by first fitting an MM-LCA to data, and then using the easy-to-compute analytical formulas to calculate B_i for variable ranking and selection. In the above example, Y_1 would be ranked as the variable with the highest impact on the posterior distribution and retained in the model.

3. Komolgorov variation of posterior distribution

The second measure for variable assessment, the Komolgorov variation of posterior distribution, or KVP, is based on the statistical distance between the class-conditional distributions. Formally, the Komolgorov distance [1] between two distributions, f and g, is defined as

$$d_{KD}(f, g) = \int |f(x) - g(x)| dx, \quad (18)$$

and the measure of KVP is defined as:

$$C_i = \max_{s \in \mathscr{P}(\mathscr{S})} \sum_{z \in s} \alpha_z d_{KD}(\pi_{y_i|z+1}, \pi_{y_i|z}), \quad (19)$$

where $\mathscr{P}(\mathscr{S})$ represents the set of all permutations of class index set $\mathscr{S}$, and $\pi_{y_i|\mathscr{S}+1} = \pi_{y_i|1}$.

3.1. Justification of the measure

The KVP can be linked to a concept of "classification accuracy" in MM-LCA. The idea behind KVP is that for a candidate variable to be included in an MM-LCA it needs to have a substantial impact on the "classification accuracy" of the latent classes, and one way to assess impact is to compare the "classification accuracies" of the MM-LCA with and without the specific variable in the model. If there is a significant reduction in classification accuracy without the variable in the model, then the variable is deemed important. Otherwise, the variable is considered not important and could be a candidate for removal in an iterative model assessment process. This section shows how the concept of classification accuracy is operationalized in terms of total variation and how KVP can be linked to total variation.

Consider a generative latent class model in which each person is given a class membership and mixed-model variables are generated given class membership. Under such a scenario, a method that allows the accurate recovery of class membership is highly desirable. One measure of “classification accuracy” is the total variation (TV) of the posterior distribution, which is commonly used in the signal processing and compressive sensing literature [4, 19]. Total variation in the context of MM-LCA is defined as

$$TV(\tau)=\max_{\mathscr{S}\in\mathscr{P}(\mathscr{S})}\sum_{z\in\mathscr{S}}|\tau_{z+1}-\tau_z|, \quad (20)$$

where $\mathscr{P}(\mathscr{S})$ is defined in (19). With this definition, the measure $TV(\tau)$ is order-invariant, and thus the measure does not change even when the class labels are switched. For operational convenience, it is assumed $\tau_{\mathscr{S}+1}=\tau_1$. Figure 1 shows the TVs of three example posterior distributions, $\tau=\{\tau_z|z=1,\dots,4\}$, respectively with TV values of 0,1, and 2 (from left to right). From the classification perspective, the third example (TV=2) would be the most desirable. Briefly, higher TV value implies better discrimination between classes. It is interesting to note that the TV of any posterior distribution is bounded between 0 and 2, as demonstrated respectively by the first and the third examples in Fig. 1. The proof of the bounds are straightforward and not included here. It is important to note that from the definition of total variation, the measure TV is not dependent on knowing the actual or true class membership. In a sense, the TV can be thought of as a measure of (non)uniformity of the classification probabilities computed over all pairs of classes for an observed data point.

Total variation as defined in (20) is a function of the observed outcome y. The expectation of total variation is a more useful summary for the purpose of assessing the overall impact of a variable. The expectation of TV is given by

$$E_y(TV(\tau))=\int TV(\tau)p(y)dy. \quad (21)$$

The TV measure inherits two desirable features directly from its very definition: that it can be applied to mixed-mode data, and that it is order invariant, which means that one does need to be concerned about label switching when compiling comparison statistics for different models. Besides these two features, there are two less trivial properties of the TV measure that make it desirable for assessing mixed-mode variables. The first property concerns the lower bound of its value and states that using any variable for classification in an MM-LCA could not deteriorate classification accuracy compared to the classification based on the prior distribution. Specifically this non-negativity property is stated as:

$$E_y(TV(\tau))-TV(\alpha)\geq 0. \quad (22)$$

The non-negativity property guarantees that any variable added to the MM-LCA cannot decrease the expected TV as compared to the TV of the prior distribution. In other words, one would avoid a potential awkward situation in which using a variable in the model could make the classification accuracy worse than simply using the prior distribution of the classes without knowledge of any manifest variable.

The second property of the expected TV is related to a relative measure of impact - a comparison of the TVs of the posterior distributions with and without the given variable in a given model. Define, for a given variable y_i, $\tau^- = p(z|y^-)$, where y^- denotes the remainder of the set of variables after deleting y_i. Then under the condition that a monotonic relationship exists between the joint probability, $p(z, y^-)$, and the class-conditional probability, $\pi_{y_i|z}$ — that is, $\pi_{y_i|z_1} \geq \pi_{y_i|z_2}$ if $p(z_1, y^-) \geq p(z_2, y^-)$— the following inequality stands, i.e.,

$$TV(\tau) \geq TV(\tau^-), \quad (23)$$

— that is, the total variation of the posterior distribution after adding an extra variable y_i would be no less than the total variation of the posterior distribution without y_i. The proofs for the two properties are included in the supplementary file.

The proposed measure KVP takes advantage of the properties of TV, and the connection between the two can be summarized by the following theorem, which states that the absolute value of the expected total variation difference between models with and without a given variable in an MM-LCA is bounded by a weighted sum of KVPs, where the weights are the priors of the latent classes.

Theorem 1. *The expected difference of TV (τ) and TV (τ^-) is upper-bounded by the KVP:*

$$|E_y(TV(\tau)) - E_y - (TV(\tau^-))| \leq C_i, \quad (24)$$

where C_i is given in (19).

Although not obvious, C_i has an upper bound as 2 and a lower bound 0 as well. The upper bound is achieved when $\forall_z$, $d_{KD}(\pi_{y_i|z+1}, \pi_{y_i|z}) = 2$. Unlike the EPG B_i, closed form solutions of C_i may not be immediately available. However, simple numerical procedures such as finite difference, can be used for computing C_i. The proof of Theorem 1 is given in an appendix.

Theorem 1 states that the change of the expected total variation of the posterior distribution between models with and without a variable, is upper-bounded by the weighted sum of the Kolmogorov distances between class-conditional distributions. This suggests that variables with large C_i's are more desirable. If the increase of classification accuracy by a variable as measured by expected TV is upper-bounded by a small value, the variable would not be very useful. The reasoning is similar as that for using the upper bound of cross entropy in EPG; the difference here is that the upper bound for TV is used, and the two quantities - TV and KVP - share the same lower and upper bounds of variation. Their common bounds can be viewed as yet another desirable property because this implies that C_i could not be too far off from the total variation measure. To put it differently, it allows a tighter interpretation of C_i in terms of total variation.

4. Numerical results

The following numerical examples used both simulated and real data for illustrating the proposed two measures of EPG and KVP. One challenge of evaluating the performance of a new procedure for LCA - or for unsupervised learning procedures in general - is that if real data are used, there is no known label for the latent classes and hence it is hard to evaluate classification accuracy. In the first example using simulated data, LCA parameters derived from a real data set of archaeometry data are used to simulate a set of continuous and discrete variables. Therefore, the class labels for latent classes are known and can be used in evaluating the performance of the proposed measures. In the second example, real genetic data from several ethnic groups is used. The latent classes are based on the ethnic group, therefore the class labels are known as well.

4.1. Simulated Archaeometry data

The first example used archaeometry data from [18], in which 3 latent classes were estimated from 21 continuous normal variables, and 12 binary variables selected from the original 19 binary variables. These variables were measurements taken from archeological artifacts - such as floor vases - dated back to ancient Italy. The continuous measures provided chemical information whereas the binary variables were obtained from petrological analysis. The discrete and continuous data are simulated using the reported estimated parameters in [18] for a sample size of 1,000 and used the data for analysis. Without refitting the original latent class model, the simulated data and the known model parameters are used to compute the posterior probabilities and the classification accuracy. The two defined measures for assessing variables were included in the analysis: the log EPG ($\log(B_i)$), and the KVP (C_i). Scatterplots of the two measures are shown in Fig. 2, in which near-monotonic relationships are clearly seen between the two measures. However, the two measures appear to assess continuous and binary variables differently. This is further evidenced in the following analysis when both measures are compared to classification accuracy.

For the n^{th} subject, the classification is $z_n = \max_z p(z|y_{n1}, \ldots, y_{nm})$, and the classification accuracy $r(y)$ is the percent of correctly classified subjects, given the observed variables. In Fig. 3, the two measures log EPG and KVP are evaluated against the actual classification accuracy using each individual variable, i.e., $z_n = \max_z p(z|y_{ni})$, and r_i is the percent of correctly classified subjects, given y_i. Because class membership is known in this simulated data set, the measure r_i can be considered an objective reference for performance. From Fig. 3, C_i has a clear and strong linear monotonic relationship with r_i, while the monotonic relationships is less clear in $\log(B_i)$, even though the continuous and binary variables each exhibits a high degree of monotonicity. Based on the result in Fig. 3, the measure C_i appeared to be more consistent with the objective measure r_i and subsequently was used to rank all the 33 continuous and binary variables. Table 1 shows the values of the conditional probabilities across the three latent classes, B_i, C_i, and the rank order of the binary variables based on C_i, whereas Table 2 shows the conditional means, B_i, C_i, and the rank order of the continuous variables. The ordering of the variables in both tables are based on B_i. The tables show that the top 5 ranked variables using C_i are continuous, and appear to dominate the binary variables for classification. The measure EPG, on the other hand, tends to report high values for binary variable and ranks four of the binary variables higher than all the continuous variables. One reason for this to arise is that the conditional probabilities for these binary variables are close to the boundary which lead to large odds ratios. As a result, the ratio for computing EPG in equation (12) gives exceedingly high EPG values.

Finally, in this example, the top 10 variables that have the largest values of C_i are selected. In this case, the selected continuous variables were 12, 6, 9, 2, 9, 11, 4, and 8, and the selected binary variables were 13, and 14. Using only these ten variables, which are less than one third of the original number of variables, the resulting MM-LCA could achieve a classification accuracy of 99.71%.

4.2. Hapmap2 data

The data set for this example was extracted from Hapmap2 [11], which is a second-generation version of the haplotype map of the human genome. A Hapmap2 dataset of 210 subjects including three populations – central European, African, and Chinese/Japanese, with respective sample sizes of 60, 60 and 90 was acquired. For each subject, 1% of the original two million SNPs was randomly sampled to result in 20, 293 SNPs, for each of which the major and minor alleles frequencies were counted. As the ethnicity of an individual was known, the three populations are treated as latent classes. Thus, the variable assessment question here is to select a subset of the SNPs to achieve a high level of classification accuracy.

The EPG, B_i, and the KVP, C_i, are computed for each SNP and compared them with two commonly used statistics in population differentiation – the likelihood ratio [12] and the F_{st} statistic, which are defined as follows.

The likelihood ratio for the i^{th} SNP is given by,

$$\mathrm{LRT}_i=2\sum_{j=1}^{S}\sum_{k=1}^{2}N_{ijk}\frac{N_{ijk}}{E_{ijk}}, \quad (25)$$

where S is the number of populations, N_{ijk} is the k^{th} allele count of the i^{th} SNP in the j^{th} population, and p_{ijk} is the corresponding expected allele frequency.

The F-statistic F_{st} is defined as,

$$F_{st}^{(i)}=\sum_{k=1}^{2}\frac{\sum_{j=1}^{S}(p_{ijk}-p_{ik})^2}{p_{ik}(1-p_{ik})}, \quad (26)$$

where $p_{ijk} = N_{ijk}/n_j$, n_j is the number of subject in the j^{th} population, and $p_{ik}=\sum_{j=1}^{S}N_{ijk}/\sum_{j=1}^{S}n_j$ is the probability of observing the k^{th} allele in the i^{th} SNP in all populations.

Of note, the F-statistic is similar to Efrons pseudo R-squared measure [8], and the LRT is similar to McFaddens pseudo R-squared measure [15].

The relationships between the three statistics, C_i, F, and LRT, are shown in Fig. 4, wherein scatterplots between pairs of the three statistics are plotted, in which both x and y-axis are in log scales. Overall, all three statistics are monotonically related to each other, and F_{st} is more linearly related to LRT, while LRT and F_{st} appear to have greater variations than C_i. The scatterplot between B_i and C_i (not shown) is also monotonic and approximately linear.

The performances of four measures, i.e., F_{st}, LRT, B_i and C_i, are compared in terms of classification accuracy. All 20,293 SNPs are first separately ranked according to each of the four measures. Starting with the top ranked SNP, the classification accuracy is calculated only using this SNP, and then the next highest ranked SNP is incrementally added. Classification accuracies were calculated for each measure as a function of the number of SNPs in the LCA. In a way similar to the first example, the classification accuracy was calculated for each set of chosen SNP-i.e., the posterior distribution of population classes for each subject is computed, and the total number of correctly classified subjects is counted. Figure 5 shows that the accuracy curves for all four measures limited to the first 50 SNPs. The curve for C_i rises fastest among the four, suggesting that if only a small number of SNPs were used for classification, then C_i would be the most effective measure. Paradoxically, as more SNPs are included, the classification accuracies of all measures, to various extents, tend to fluctuate. When the number of SNPs are more than 20, F_{st} performs the best, although its performance also seems to decrease after including 45 SNPs. The performance of C_i and LRT tend to converge, and for B_i its performance is comparable to the other measures. One remark is that LRT is applicable to mixed-mode data while F_{st} only applies to discrete data, so the above results are not necessarily generalizable to data with different modes.

To compare how similar, or different, the several measures selected the respective most important SNPs, we calculated the number of overlapping top ten SNPs selected by each measure. We found that the top-10 SNPs selected by the measures substantially overlapped. For example, 9 out of 10 overlapped between F_{st} and LRT, 7 out of 10 between C_i and LRT, 6 out of 10 between C_i and F_{st}, 3 out of 10 between B_i and F_{st}, and 3 out of 10 between B_i and LRT. Overall, compared to B_i, C_i had a stronger overlap with F_{st} and LRT.

In summary, this example illustrates the practical relevance of the proposed measures for efficiently identifying among a large number of SNPs, a small number of SNPs that could be used for clustering and classification purpose.

4.3. Guidelines for practitioners

A summary of how EGP and KVP can be computed and used in practice is listed below.

1. Collect all the available variables, which could be of mixed mode, and fit a latent class model.
2. Use the class-conditional parameters estimated in the previous step to compute the proposed EPG and KVP measures.
3. Select the most discriminating variables and eliminate the weak variables based on the measures.

4. Re-run LCA using the selected variables. Using the classification accuracy of the full model as a reference, compute the classification rate of the reduced model, which assumes the same latent structure in the number of classes.
5. Assess the adequacy of the reduced model. Repeat steps 1 to 4 if necessary until a satisfactory reduced set of variables is selected.

5. Discussion

Two measures - the EPG and KVP, are presented for assessing variable impact in the context of mixed-mode latent class modeling. The mathematical foundations for these measures are elaborated respectively in terms of cross entropy and total variation. Two empirical studies - one using simulated data based on parameters derived from an archaeometry study, and a real data set in genomic that contains > 20, 000 manifest binary variables, are used to illustrate the two measures. The several findings from this investigation are summerized as follows. (1) EPG and KVP both can be used for assessing mixed mode variables in LCA, and both measures are straightforward to compute given an MM-LCA. (2) Both measures seem to perform reasonably well compared to existing measures such as LRT and F_{st}. (3) The result from EPG may have to be handled with caution for binary variables in cases when the conditional probabilities are close to the boundary. (4) The KVP measure appears to have an advantage in its interpretation - its value is bounded between 0 and 2, so both absolute and relative interpretations of the “scale” is possible (e.g., value of 1.8 implies high impact on an absolute scale, and variable A that has a value of 1.1 has stronger impact than another variable B that has a value of 0.9.) While it is too early to offer any definitive conclusion about the proposed measures of EPG and KVP at this point given the preliminary nature of the empirical studies, both the theoretical and the empirical investigations demonstrate promising properties, as well as potential pitfalls, in the two measures. More extensive work such as Monte Carlo experiments will be required to further assess the performance of EPG and KVP.

Besides the lack of a large-scale empirical study validating the performance of the proposed measures, there are also other limitations. Theoretical results are limited by the assumption that the prior probabilities, $\{\alpha_z | z = 1, \ldots, S\}$, remain constant as variables are added or removed in the latent class model. In real data-driven MM-LCA estimation, the prior will vary as a function of the set of variables included. The reported theoretical and empirical results should be construed as focusing on comparison of the impact of individual variables within the same latent class model.

The proposed measures are different from item fit statistics [21], where two items that both have good fit do not necessarily have similar impact on the likelihood function. Item fit statistic can be used to first eliminate poorly fitted items - continuous or discrete — separately, and the proposed measures can be subsequently used to evaluate the remaining variables.

The proposed measures can also be applied to mixtures of regression models [16] which specify conditional distributions, for example, mixtures of regression models for normal, binary and Poisson variables. For instance, when the conditional distributions are normal given the mixture component, an EPG-based regression procedure replaces the conditional mean μ_j in Eq. (7) with the component-specific linear predictors, keeps the same σ_j , and then compares $\{\mu_j | \sigma_j^2, j=1, S\}$. Such work is now in progress.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is sponsored by NIH grants R01AG031827A, U01HL101066-01 (PI: E.H. Ip).

Appendix A

Proof of Theorem 1

First the permutation is ignored. By the definition of $E_y(TV(\tau))$,

$$\begin{aligned} E_y(TV(\tau)) &= \sum_{z=1}^{S} \int |\tau_{z+1} - \tau_z| p(y) dy \\ &= \sum_{z=1}^{S} \int |\alpha_{z+1}\pi_{y^-|z+1}\pi_{y_i|z+1} - \alpha_z \pi_{y^-|z}\pi_{y_i|z}| dy \\ &= \sum_{z=1}^{S} \int |\alpha_{z+1}\pi_{y^-|z+1}\pi_{y_i|z+1} - \alpha_z \pi_{y^-|z}\pi_{y_i|z+1} + \alpha_z \pi_{y^-|z}\pi_{y_i|z+1} - \alpha_z \pi_{y^-|z}\pi_{y_i|z}| dy. \end{aligned} \quad \text{(A.1)}$$

From (A.1), two inequalities can be derived, the first of which is

$$Ey(TV(\tau)) \le \sum_{z=1}^{S} \int |\alpha_{z+1}\pi_{y^-|z+1} - \alpha_z \pi_{y^-|z}\pi_{y_i|z+1}| dy + \sum_{z=1}^{S} \alpha_z \int |\pi_{y_i|z+1} - \pi_{y_i|z}| \pi_{y^-|z} dy, \quad \text{(A.2)}$$

where the first term on the RHS is $E_{y^-}(TV(\tau^-))$ and the second term is the C_i. It is straightforward to derive the second inequality:

$$E_y(TV(\tau)) \ge E_y(TV(\tau^-)) - \sum_{z=1}^{S} d_{KD}(\pi_{y_i|z+1}, \pi_{y_i|z}). \quad \text{(A.3)}$$

These two inequalities together lead to (24). Because the equality above is true for any permutation, the inequality holds for the permutation that gives the maximum TV of τ.

References

1. Ali SM, Silvey SD. A general class of coefficients of divergence of one distribution from another. Journal of the Royal Statistical Society. Series B. 1966; 28(1):131–142.
2. Bauer DJ, Curran PJ. Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. Psychological Methods. 2003; 8(3):338–363. [PubMed: 14596495]
3. Cai JH, Song XY, Lam KH, Ip EHS. A mixture of generalized latent variable models for mixed mode and heterogeneous data. Computational Statistics & Data Analysis. 2011; 55(11):2889–2907.

4. Chan, Tony; Shen, Jianhong. Image processing and analysis: variational, PDE, wavelet, and stochastic methods. 2005 Siam,
5. Daniels MJ, Normand ST. Longitudinal profiling of health care units based on continuous and discrete patient outcomes. Biostatistics. 2006; 7(1):1. [PubMed: 15917373]
6. Dean N, Raftery AE. Latent class analysis variable selection. Annals of the Institute of Statistical Mathematics. 2010; 62(1):11–35. [PubMed: 20827439]
7. Dunson DB. Dynamic latent trait models for multidimensional longitudinal data. Journal of the American Statistical Association. 2003; 98(463):555–563.
8. Efron, Bradley. Regression and anova with zero-one data: Measures of residual variation. Journal of the American Statistical Association. 1978; 73(361):113–121.
9. Everitt BS. A finite mixture model for the clustering of mixed-mode data. Statistics & probability letters. 1988; 6(5):305–309.
10. Everitt, BS. Cluster Analysis. 1993. London, UK: Edward Arnold and Halsted Press; 1993.
11. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al. A second generation human haplotype map of over 3.1 million snps. Nature. 2007; 449(7164):851–861. [PubMed: 17943122]
12. Goudet J, Raymond M, De-Meeus T, Rousset F. Testing differentiation in diploid populations. Genetics. 1996; 144(4):1933. [PubMed: 8978076]
13. Joreskog, KG. A general method for estimating a linear structural equation system. In: Goldberger, AS.; Duncan, OD., editors. Structural Equation Models in the Social Sciences. New York: Seminar Press; 1973. p. 85-112.
14. Lawrence CJ, Krzanowski WJ. Mixture separation for mixed-mode data. Statistics and Computing. 1996; 6(1):85–92.
15. McFadden, Daniel. Conditional logit analysis of qualitative choice behavior. 1973
16. McLachlan, G.; Peel, D. Finite mixture models. Vol. ume 299. Wiley-Interscience; 2000.
17. Moustaki I, Knott M. Generalized latent trait models. Psychometrika. 2000; 65(3):391–411.
18. Moustaki I, Papageorgiou I. Latent class models for mixed variables with applications in Archaeometry. Computational statistics & data analysis. 2005; 48(3):659–675.
19. Osher, Stanley; Burger, Martin; Goldfarb, Donald; Xu, Jinjun; Yin, Wotao. An iterative regularization method for total variation-based image restoration. Multiscale Modeling & Simulation. 2005; 4(2):460–489.
20. Raftery AE, Dean N. Variable selection for model-based clustering. Journal of the American Statistical Association. 2006; 101(473):168–178.
21. Reise SP. A comparison of item-and person-fit methods of assessing model-data fit in irt. Applied Psychological Measurement. 1990; 14(2):127–137.
22. Sammel MD, Ryan LM, Legler JM. Latent variable models for mixed discrete and continuous outcomes. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1997; 59(3):667–678.
23. Shi JQ, Lee SY. Latent variable models with mixed continuous and polytomous data. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2000; 62(1):77–87.
24. Vermunt, JK.; Magidson, J. Hagenaar, JA.; McCutcheon, AL. Applied latent class analysis. New York: Cambridge University Press; 2002. Latent class cluster analysis; p. 89-106.
25. Yang M, Dunson DB. Bayesian semiparametric structural equation models with latent variables. Psychometrika. 2010:1–19.

HIGHLIGHTS

- Two measures are proposed for assessing continuous and discrete variables in LCA.
- Both measures are either in closed form or straightforward to compute.
- Both measures perform reasonably well compared to existing measures such LRT and F_{st}.
- Both absolute and relative interpretations of one measure are possible.

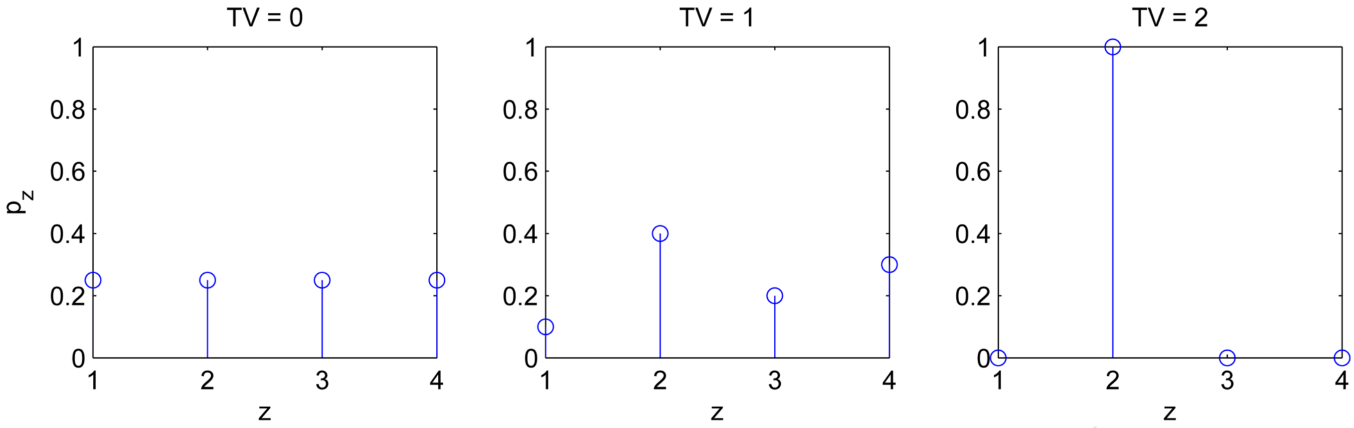

Figure 1.

Three example posterior distributions with increasing TVs.

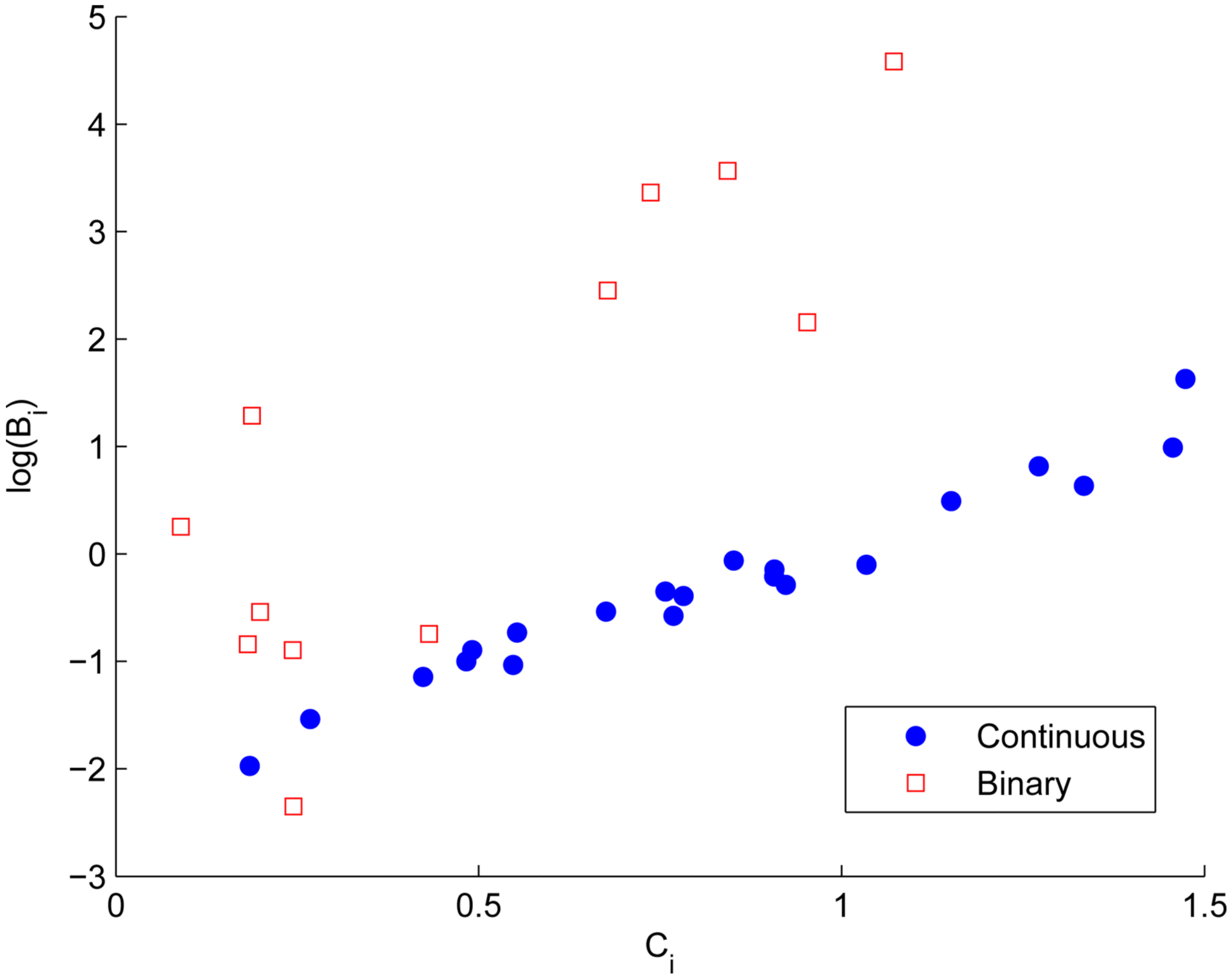

Figure 2.
Scatter plots of the two measures of both binary and continuous variables.

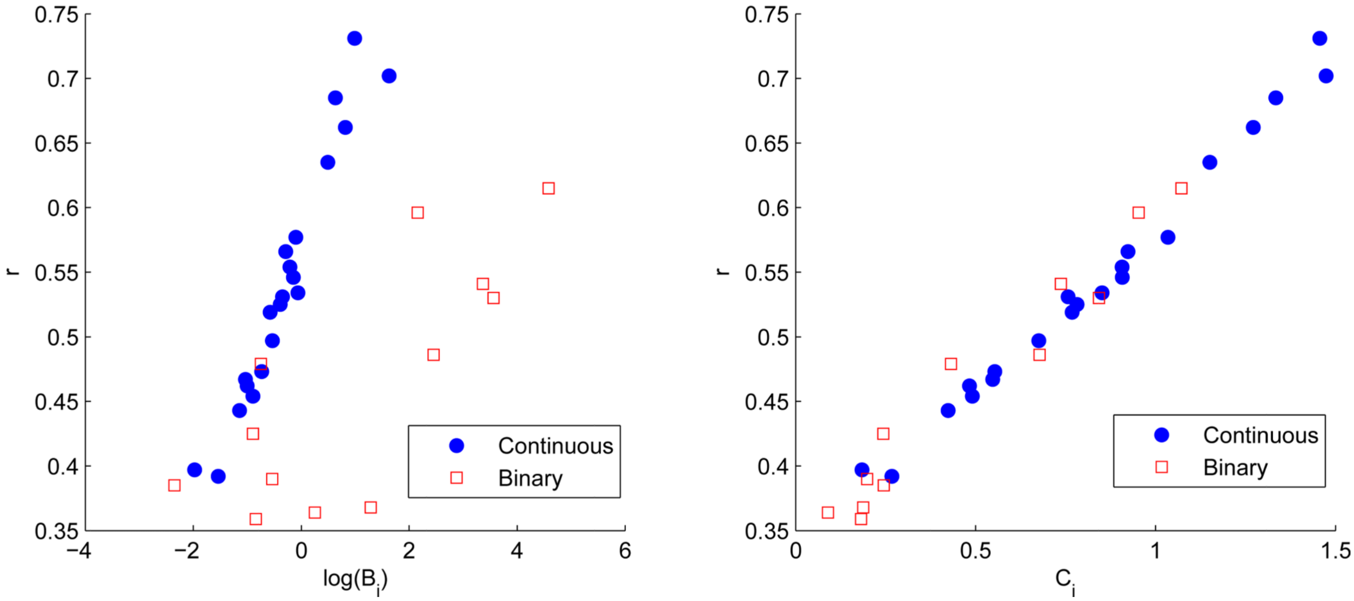

Figure 3.

Scatter plots between the classification accuracy and each of the two measures of both binary and continuous variables.

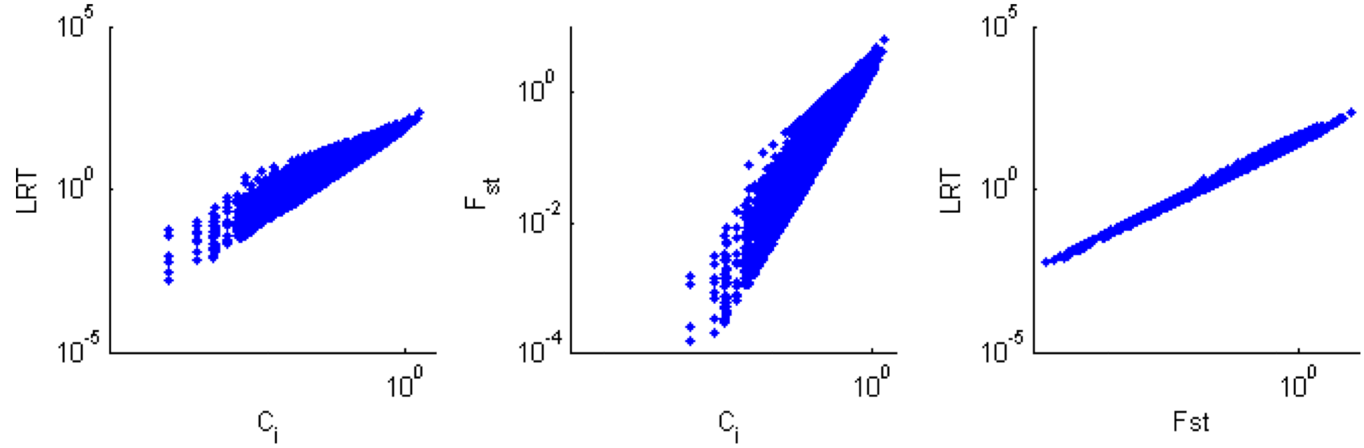

Figure 4.

Scatter plots for comparing C_i, LRT and F_{st}.

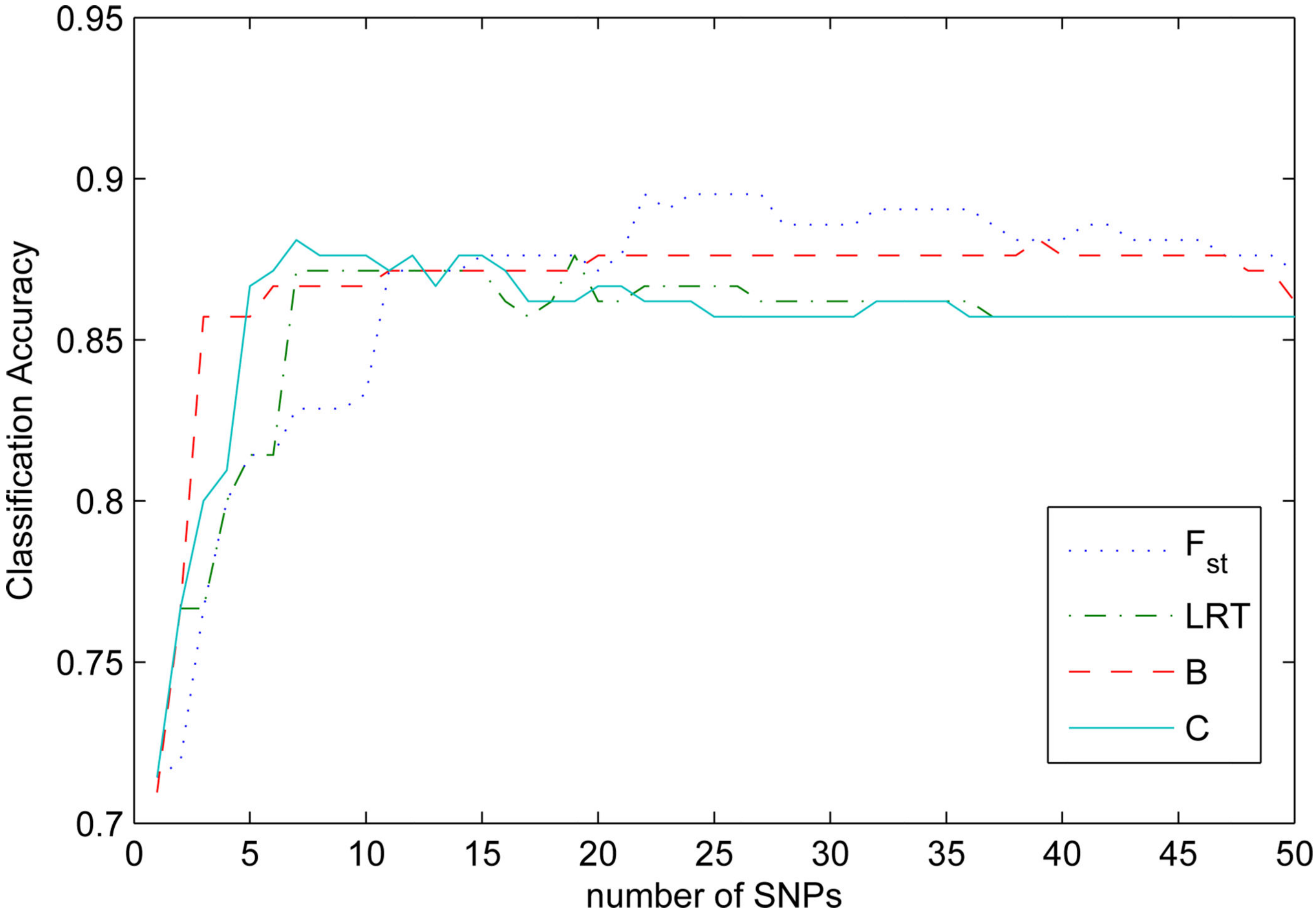

Figure 5.
Classification accuracy as a function of the number of variables incrementally added to model according to their rank order.

Table 1

An example of binary variables in a latent class model.

Variable	π_1	π_2	π_3	B_i	C_i	Rank
13	1	0.19	1	98.38	1.07	6
5	0.67	0.59	0	35.70	0.84	13
6	1	0.44	1	29.37	0.74	17
9	0.5	0.07	0	12.03	0.68	18
14	1	0.85	1	4.06	0.95	8
15	0.92	0.93	1	1.66	0.19	30
8	0.17	0.07	0.82	0.95	0.09	33
1	0.04	0.11	0.19	0.85	0.20	29
19	0.5	0.81	0.77	0.79	0.43	24
10	0.08	0.11	0.27	0.71	0.24	28
17	0.12	0.19	0.05	0.71	0.18	32
18	0.46	0.59	0.64	0.29	0.25	27

Table 2

An example of continuous normal variables in a latent class model.

Variable	μ_1	μ_2	μ_3	B_i	C_i	Rank
12	−0.78	1.17	−0.59	5.10	1.47	1
6	−0.36	1.06	−0.9	2.69	1.46	2
2	−0.72	1.04	−0.49	2.26	1.27	4
9	0.24	0.73	−1.16	1.88	1.33	3
11	−0.55	−0.44	1.14	1.63	1.15	5
4	−0.74	−0.05	0.87	0.90	1.03	7
13	0.87	−0.4	−0.46	0.94	0.85	12
8	0.54	0.26	−0.9	0.86	0.91	10
18	0.14	0.58	−0.86	0.81	0.91	11
1	0.67	0.05	−0.8	0.75	0.92	9
3	−0.29	−0.41	0.82	0.70	0.76	16
20	−0.6	0.65	−0.07	0.68	0.78	14
5	−0.59	−0.03	0.68	0.56	0.77	15
17	−0.52	0.6	−0.17	0.59	0.68	19
5	0.59	−0.25	−0.34	0.48	0.55	20
21	−0.37	0.47	−0.17	0.41	0.49	22
16	0.46	−0.01	−0.49	0.36	0.55	21
14	0.14	0.3	−0.53	0.37	0.48	23
15	0.36	−0.38	0.08	0.32	0.42	25
7	0.31	−0.14	−0.16	0.22	0.27	26
19	−0.07	0.18	−0.15	0.14	0.18	31