A paper presented at the 1997 Annual Meeting of the American Educational Research Association (Chicago, IL, USA, March 1997)

Bayesian Finite Mixtures for Nonlinear Modeling of Educational data

Henry Tirri and Tomi Silander
Complex Systems Computation Group (CoSCo)*
P.O.Box 26, Department of Computer Science
FIN-00014 University of Helsinki, Finland

Kirsi Tirri
P.O. Box 38,Department of Teacher Education
FIN-00014 University of Helsinki, Finland

Abstract

In this paper we discuss a Bayesian approach for finding latent classes in the data. In our approach we use finite mixture models to describe the underlying structure in the data, and demonstrate that the possibility to use full joint probability models raises interesting new prospects for exploratory data analysis. The concepts and methods discussed are illustrated with a case study using a data set from a recent educational study. The Bayesian classification approach described has been implemented, and presents an appealing addition to the standard toolbox for exploratory data analysis of educational data.

1 Introduction

Quantitative research methods in education have traditionally been based on a standard "toolbox" of methods for analyzing the data collected: e.g., linear regression, discriminant analysis, exploratory and confirmatory factor analysis (Klecka, 1981; Basilevsky, 1994). In spite of the popularity of multivariate

*URL: http://www.cs.Helsinki.FI/research/cosco/

factor analysis among the practitioners, utilization of the power of general latent variable models for data analysis has been low, and based almost exclusively on linear model families. This is partly due to the controversial nature of the latent variable approaches as practiced in the applied end of the spectrum, exploratory factor analysis being a prime example of the continuing debates on the validity and arbitrariness of the method (see e.g., the discussion in (Chatfield, 1980)).

On the other hand recent years have seen an impressive growth of interest in building complex latent variable models of natural phenomena and man-made systems. Although in computer science, and related fields, nonlinear modeling has been studied for more than three decades, it is only recently that the availability of increased computing power has made the approaches more appealing, and made their application more feasible. In particular, the developments in building latent variable models expressed with graphical structures such as Bayesian networks (Heckerman, 1996; Lauritzen, 1996) and in Bayesian analysis using Markov Chain Monte Carlo methods (Gilks et al., 1996) have completely changed the level of complexity that can be addressed in modeling of data.

There is no reason to doubt that Bayesian latent variable approaches with nonlinear models will have a profound impact on modeling of social phenomena also. Unfortunately the techniques that have already proved their applicability for modeling in the context of industrial, economical or biological processes, are almost unknown to the practitioners in the educational sciences. At the same time the accelerated embedding of computer technology into all sectors of society by computerized services has made increasing volumes of data available to the analyst, thus motivating the search for better methods in model building and testing.

In this paper our purpose is to gradually introduce into the reader's mind a, perhaps less familiar, Bayesian approach to modeling. In particular, we will be here interested in the problem of unsupervised *classification*, i.e., of finding latent classes in the data. One should observe that the word "classification" is ambiguous. In discriminant analysis it means the procedure of assigning a new case to one of an existing set of possible classes. As used in this paper, however, it means *finding the class structure* from a given set of "unclassified" cases. This view of classification is also sometimes known as "conceptual clustering". Obviously, once such a set of classes has been found, they can be the basis for classifying new cases in the first sense.

Classification aims at discovering natural classes in the data. Consequently, as we will argue, these classes reflect basic regularities in the processes that generate the data, which make some cases look more like each other than the rest of the cases. Therefore classification is a powerful tool for exploratory analysis. For example, in our teacher education case study we can find "prototypical" teacher profiles reflecting different general views on teacher education. This type of discovery of previously unknown structure occurs most frequently when there are many relevant variables describing each case, because humans are poor at seeing structure in large dimensional spaces. Such situation is naturally quite prevalent in the educational data analysis, where typical questionnaires can easily have more than 100 associated assertions. A practitioner can view this Bayesian classification as a new interesting "tool" for the data analysis toolbox, but we would like to point out that underlying notions of modeling discussed in Section 3 are quite fundamental, and widely applicable outside the particular problem at hand.

To make the underlying ideas as accessible as possible, we keep the technical level of the discussion very moderate, and try to frequently refer to sources, where the technically oriented reader can find more formal treatment of the issues discussed in this paper. In some sections, such as Section 4, technical details are unavoidable, but they are not necessary for understanding the main points of the paper.

2 Example data

In order to illustrate the Bayesian classification approach, we use a typical data sample from a recent educational research project. This educational data was gathered for the research project "Effectiveness of Teacher Education in Finland" in the spring 1996. The objective of the project was to evaluate the effectiveness of Finnish teacher education at various levels from individual to international teacher education policy. A more detailed description of the framework and research conducted in the project is discussed in (Niemi and Tirri, 1996). The data adopted to this study was gathered to investigate how well the Finnish teacher education had been able to achieve the goals set to it. The goals were selected from school-law, programs of teacher education and other documents describing teachers' work at school. The teachers and their educators from four different teacher education departments in Finland were asked to perform self-evaluation on the success of teacher education for helping teachers to achieve these goals. The evaluation instrument consisted of 41 behavior statements (and information about the teacher education department), and used a Likert scale from 1 to 5 for the assertions. The results of this evaluation study are reported in the forthcoming study (Niemi and Tirri, 1997).

The data sample used for our comparison is derived from the teachers' data in the study described above. This data consist of ratings of 204 Finnish teachers. The subjects were teaching at two levels, one half being elementary school class teachers (N=110) and the other half secondary school subject teachers (N=94). These teachers came from four different teacher education departments in three different counties of Finland. The gender distribution was representative to that of Finnish teacher population—about 25% were males.

3 Bayesian modeling

Inductive modeling One of the most fundamental questions in statistical inference is finding good models. In the Bayesian terminology we could rephrase this problem as the question: "Given some data and weak prior domain knowledge, what is the most probable model of the domain?"

In this work we will focus on the problem of *inductive* model construction, in which the basic issue is distinguishing the underlying structure from noise. It is well-known that one can always find a sufficiently complex model to "explain" any data set. However, the fundamental problem here is to find a model that reflects only the general structure of the domain, not the individual idiosyncrasies of the cases (the "noise"). This *overfitting problem* is inherent to any model construction process, and the so called "Ockhams razor" principle (William of Ockham, c. 1285-1349) tells us not to overfit the data.

The solution to this overfitting problem is to find a tradeoff between the fit to data, and the complexity of the model. A model as complex as the data itself can fit the data exactly, but such a model has very little predictive value for new, unseen data. Conversely, models with little structure do not predict the given data or new data well. The real question is to find an appropriate balance between these two aspects.

Bayesian theory (Bernardo and Smith, 1994) (together with its information theoretic interpretation (Rissanen, 1989; Wallace and Freeman, 1987)) explicitly trades model complexity, as determined by prior probabilities, against the fit to the data. This trade-off is in fact a direct consequence of Bayes' theorem discussed below.

Notation Let us first introduce some general notation, used subsequently throughout the paper. The data D denotes a (random) sample of N independent and identically distributed (i.i.d.) data vectors $\vec{d}_1, \ldots, \vec{d}_N$. For our case study we have a data vector for each teacher that has answered the query, and the data vector contains background information and answers to the questionnaire questions. For simplicity, in all our discussion we assume that the data is coded by using only discrete, i.e., finite-valued, variables $X_1, \ldots, X_m$. More precisely, we regard each variable X_i as a random variable with possible values from the set $\{x_{i1}, \ldots, x_{in_i}\}$. Consequently, each data vector $\vec{d}$ is represented as a value assignment of the form $(X_1 = x_1, \ldots, X_m = x_m)$, where $x_i \in \{x_{i1}, \ldots, x_{in_i}\}$.

It will be also useful to talk about a set of models, which we will call a *model family* $\mathcal{M}$. Examples of model families include the set of linear functions (Basilevsky, 1994), or the set of graphical structures describing independence assumptions (Heckerman et al., 1995). For the classification problem, a model Θ simply means a description of the classes in terms of the joint probability distribution of $X_1, \ldots, X_m$. It is also often useful to partition the models within a model family $\mathcal{M}$ to some finite number of subsets, *model classes* M_i, where all the models within a model class share the same parametric form, i.e., the same number of parameters. Consequently, the model classes usually correspond to some specific model structure. Examples of such structure is the degree of the polynomial in polynomial regression models, or in the present case the number of classes, i.e., M_K means models with K classes. Hence, finally a *model* Θ can be defined as a parameter instantiation within some parametric model class M_i, fully determining a probability distribution in the data vector space.

Bayesian inference—an information theoretic view In Bayesian inference one searches for the most probable model Θ in a given model family $\mathcal{M}$. This search for probable models can be described alternatively in an intuitively appealing form using information-theoretic concept. Since this complementary view of Bayesian inference is not widely known, we will use it here as a tool for intuitive explanation of the method. Obviously we could have formulated this discussion also directly in terms of distributions.

It can be argued that the most probable model Θ is the one that has *the shortest encoding of the model and the data combined.* If a new data vector is described using the existing abstraction (model), a shorter total encoding will result. An example from our case study illustrates this issue. Let us assume that a set of teachers have answered the questionnaire in a very similar manner, and call this set of teachers as "Class A" teachers. Similarly another set of teachers have answers that are quite alike, let us call them "Class B" teachers. Now if we need to transmit information for specific teacher responses, the trivial way is to send the questionnaire information for each teacher. However, typically it is more efficient to first send the description of the responses for Class A teachers and Class B teachers, and then for each teacher the information about his/her "type" (Class A or B) and the differences from the standard answers in the class in question. If the answers of a particular teacher differs very much from Class A and Class B answers the approach does not essentially save anything, i.e., the encoding is not shorter than sending the answers directly.

This is how Bayesian model building method finds structure in the data—if a new data vector (teacher's answers) cannot be compactly described in terms of abstract structure of the sample data, it means that the sample data has very little predictive value for that particular data vector. Now why do we call this encoding approach a Bayesian approach?

In standard Bayesian inference text book approach one assumes that the researcher has selected a set of discrete mutually exclusive and exhaustive models $\{\Theta_1, \Theta_2, \ldots, \Theta_n\}$, and has assigned some *prior* probabilities $p(\Theta|I)$, where I is the general context of the modeling problem. Using such models we can calculate the *likelihood* $p(D|\Theta_i)$, i.e., the probability of the sample given a model Θ_i. Searching for the most probable model means finding the model Θ that maximizes the probability $p(\Theta_i|D)$, which is called the *posterior* probability. The prior, likelihood and posterior are connected via the Bayes' theorem (see e.g., (Bernardo and Smith, 1994):

$$p(\Theta \mid D) = \frac{p(D \mid \Theta)p(\Theta)}{p(D)}. \tag{1}$$

Taking the negative logarithm of this expression turns the products into sums, and gives us

$$-\log p(\Theta|D) = -\log p(D|\Theta) - \log p(\Theta) + \text{constant}. \tag{2}$$

Since we are only interested in the relative probability of the different models Θ, the last term in equation (1) can be ignored. Now the connection between

Bayesian probability theory and the coding approach becomes clear: from information theory we know that $-\log p(\vec{d_i})$ is the theoretically optimal minimum message length to encode a particular data vector $\vec{d_i}$ (Cover and Thomas, 1991).

The minimum message length in (2) is the sum of two terms. The first term is the information to describe the model Θ, which is greater for more complex, and thus less probable, models. The second term is the information required to encode the data, given the model Θ, and decreases for suitably selected more complex models. The trade-off between these two terms is another way of expressing the inherent "Ockham's razor" in Bayesianism.

We can summarize the discussion above as follows. In the Bayesian approach for finding structure in the sample *we look for regularities that allow us to predict the data in the sample well.* If we predict well, we can also use short encodings for the data. The tradeoff between too complex models and short encodings of the data (equation 2) with the model prevents us from finding models that are too closely reflecting the properties of the sample rather than the full population.

Bayesian classification We can now explain the intuition underlying the Bayesian classification with the above information theoretic argumentation. The data in the sample can be modeled by first describing a set of classes, then describing the data vectors using the prototypical class descriptions. Each description gives the probabilities of the observables, assuming that the data vector belongs to the class. The class descriptions need to be chosen in such a way that the information required to describe data vectors in the class is reduced, because they resemble the class prototype. The information reduction results from the fact that only the differences between the observed and expected values need to be described. More classes makes it possible to describe individual data vectors with less difference information, and thus the data set encoding is shorter. However, it takes a certain amount of information to describe a set of classes as probabilities of the variable values (given that the data vector belongs to the class). Thus the Bayesian classification approach involves finding the set of classes that minimizes the total information

```
total description =
          class descriptions + sample description
          given the class descriptions
```

If the sample is "random", i.e., exhibit no regularities, it is very unlikely that one can find class descriptions for which the total information is less than what needs to be used to describe each data vector in the sample individually. One should notice that this discussion implies that one is able to have a rigorous means to determine the proper number of latent classes indicated by the sample data—a problem which is very difficult to solve rigorously by other approaches. For more detailed discussion see e.g., (Cheeseman and Stutz, 1996; Kontkanen et al., 1996a; Kontkanen et al., 1996b; Kontkanen et al., 1997).

An interested reader can find more formal treatment of the general ideas discussed above in the seminal works by Rissanen (Rissanen, 1987; Rissanen, 1989) and Wallace et al. (Wallace and Boulton, 1968; Wallace and Freeman, 1987); the Bayesian classification is addressed in (Cheeseman and Stutz, 1996).

4 Model family: finite mixtures

Like any other Bayesian inference, Bayesian classification is always relative to a model family $\mathcal{M}$. For the classification problema very natural model family is the set of *discrete finite mixtures* ((Everitt and Hand, 1981), (Titterington et al., 1985)), where the joint domain probability distribution is approximated as a weighted sum of mixture distributions.

Let $X_1, \ldots, X_m$ be a set of m ($m \geq 1$) discrete (random) variables, and $\vec{d} \in D$ is a sample from the joint distribution of the variables $X_1, \ldots, X_m$. Then the *finite mixture* distribution for $\vec{d}$ can be written as ($K \geq 1$)

$$\begin{aligned} p(\vec{d}) &= p(X_1 = x_1, \ldots, X_m = x_m) \\ &= \sum_{k=1}^{K} \left(p(Y = y_k) p(X_1 = x_1, \ldots, X_m = x_m | Y = y_k) \right), \end{aligned} \tag{3}$$

where Y denotes a latent *clustering random variable*, the values of which are not given in the data D, and K is the number of possible values of Y.

Thus in finite mixture models the problem domain probability distribution is approximated by a weighted sum of mixture distributions, where each mixture component $p(X_1 = x_1, \ldots, X_m = x_m | Y = y_k)$ models one data producing mechanism. If the variables $X_1, \ldots, X_m$ are independent, given the value of the clustering variable Y, equation (3) becomes

$$p(\vec{d}) = \sum_{k=1}^{K} \left(p(Y = y_k) \prod_{i=1}^{m} p(X_i = x_i | Y = y_k) \right). \tag{4}$$

For the Mixture Density Networks considered here this independence assumption holds and consequently computation uses equation (4).

A finite mixture model partitions the data to K clusters. This partitioning can be modeled by introducing for each data vector $\vec{d}_j$ an unobserved latent variable Z_j, the value of which gives the the cluster index for the cluster vector $\vec{d}_j$ belongs to. We can now think a vector $Z = (z_1, \ldots, z_N)$, consisting of the values of the latent variables $Z_1, \ldots, Z_N$, as a random sample from the distribution of Y like D is a random sample from the joint distribution of $X_1, \ldots, X_m$. However, for technical reasons it is more convenient to consider each value z_j as a vector of *cluster indicator variable* values, $z_j = (z_{j1}, \ldots, z_{jK})$, where

$$z_{jk} = \begin{cases} 1, & \text{if } \vec{d}_j \text{ is sampled from } P(\cdot|Y = y_k), \\ 0, & \text{otherwise.} \end{cases}$$

Finite mixtures as defined in equation (4) is a generic model family, as we still have to fix the cluster distribution $p(Y)$ and the intra-class conditional distributions $p(X_i|Y = y_k)$[1]. Most commonly used component functions in the literature are the univariate normal distributions (see e.g., (Titterington et al., 1985)). In educational domains the variables are usually discrete, thus we can drop the assumption of the form of the distribution. For the univariate case a binomial model could be used, but for the general case with $m > 1$ a natural choice is the multivariate generalization of the binomial distribution called the *multinomial distribution*

$$p(\vec{c}|\Theta) = \begin{pmatrix} N' \\ c_1 \ldots c_{n_i} \end{pmatrix} \prod_{j=1}^{n_i} \theta_j^{c_j}$$

where $\vec{c} = (c_1, \ldots, c_{n_i})$ is the vector of counts of the number of observations of each value of X_i. In addition the sum of probabilities $\sum_{j=1}^{n_i} \theta_j = 1$ and $\sum_{j=1}^{n_i} c_j = N'$ (N' is the total number of observations). Since we are interested in the data distribution, i.e., $p(X_i|Y = y_k)$ the multinomial distribution form

[1]Here we consider only mixtures in which all the component distributions come from the same parametric class.

simply reduces to a product of probabilities θ_j. Analogously we assume that the cluster distribution $p(Y)$ is multinomial. Thus in order to get a model, we need to fix the number of the mixing distributions (K), and determine the values of the model parameters. For technical reasons it will be convenient to make a notational distinction between the mixture weight parameters and the parameters of the intra-class conditional distributions, i.e., $\Theta = (\alpha, \Phi), \Theta \in \Omega$, where $\alpha = (\alpha_1, \ldots, \alpha_K)$ and $\Phi = (\Phi_{11}, \ldots, \Phi_{1m}, \ldots, \Phi_{K1}, \ldots, \Phi_{Km})$, with the denotations $\alpha_k = P(Y = y_k)$, $\Phi_{ki} = (\phi_{ki1}, \ldots, \phi_{kin_i})$, where $\phi_{kil} = P(X_i = x_{il}|Y = y_k)$.

Since our estimation of the network parameters will be Bayesian (Bernardo and Smith, 1994) we need to fix the prior distributions for the parameters. The family of Dirichlet (multivariate Beta) densities is conjugate to the family of multinomials, therefore we assume that prior distributions of the parameters are $(\alpha_1, \ldots, \alpha_K) \sim \mathrm{Di}\,(\mu_1, \ldots, \mu_K)$ and $(\phi_{ki1}, \ldots, \phi_{kin_i}) \sim \mathrm{Di}\,(\sigma_{ki1}, \ldots, \sigma_{kin_i})$, $(1 \leq k \leq K, 1 \leq i \leq m)$, where

$$\{\mu_k, \sigma_{kil} \mid 1 \leq k \leq K; 1 \leq i \leq m; 1 \leq l \leq n_i\}$$

are called the *hyper parameters* of the corresponding distributions. Assuming that the parameter vectors α and Φ_{ki} are independent, the joint prior distribution of all the parameters can be expressed as

$$\mathrm{Di}\,(\mu_1, \ldots, \mu_K) \prod_{k=1}^{K} \prod_{i=1}^{m} \mathrm{Di}\,(\sigma_{ki1}, \ldots, \sigma_{kin_i}).$$

The finite mixture model family is universal in the sense that it can approximate any distribution arbitrarily close as long as a sufficient number of components is used. Unfortunately such generality typically implies also that parameter estimation can become computationally inefficient.

5 On Bayesian explorative analysis

In traditional educational research the data, such as in our case study, would typically be analyzed by factor analysis. Factor analysis is usually motivated by the fact that observed variables can be correlated in such a way that one is able to reconstruct their correlation by a smaller set of parameters, which could represent the underlying structure in a concise and interpretable form.

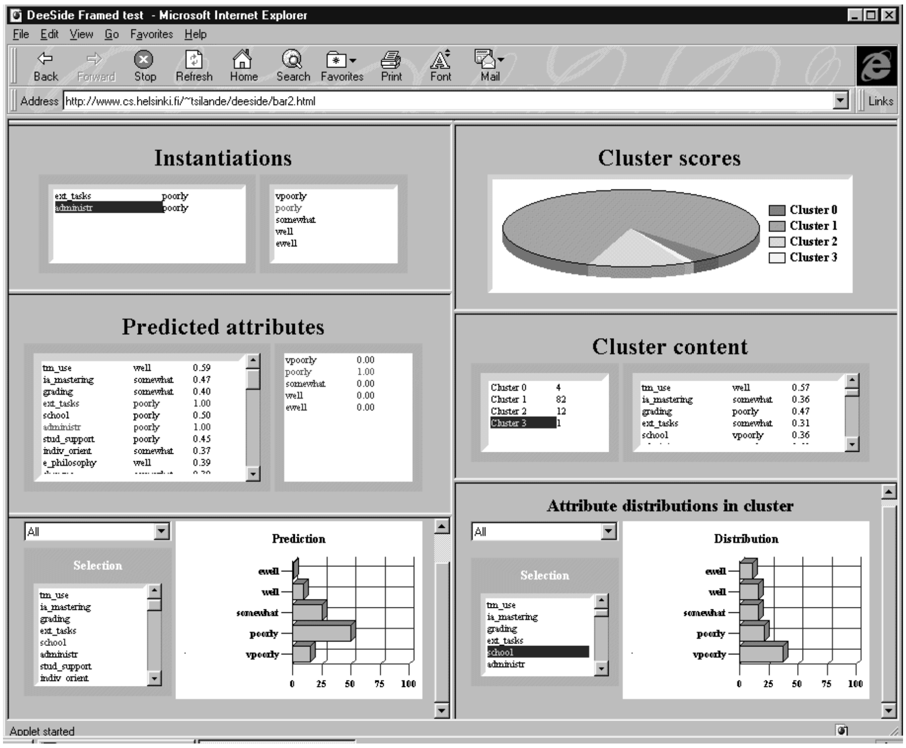

Figure 1: A snapshot of the interface of the NONE software tool.

In the Bayesian finite mixture based classification we have an interesting different approach for finding interpretable structures from the data. As discussed earlier, a class can be viewed as a "prototype", i.e., an abstract description which reflects dependencies between the values of the observables. Such prototypes can be understood as "conceptual sufficient statistics"—they summarize the general tendencies existing in the data.

In factor analysis one is often interested in factor loadings, i.e., in the measure how much a variable X_i is representative of, or agrees to, the factor in question. In our Bayesian finite mixture approach the corresponding notion would be the Kullback-Leibler distance of the unconditional and conditional marginal likelihood of X_i, i.e.,

$$\mathcal{D}_{\mathrm{KL}}(p(X_i|Y=k,\Theta), p(X_i|\Theta)),$$

where $\mathcal{D}_{\mathrm{KL}}(p, q)$ is the relative entropy between p and q (Cover and Thomas, 1991). Similarly we can also study how different the multivariate class distribution is from the unconditioned joint distribution (a "Bayesian Wilk's lambda"), defined as the relative entropy between the unconditional and conditional joint distributions, i.e.,

$$\mathcal{D}_{\mathrm{KL}}(p(\vec{X}|Y=k,\Theta), p(\vec{X}|\Theta)).$$

However, as finite mixtures model the joint probability distribution of all the variables $X_1, \ldots, X_m$, we can in fact *explore the predictive (marginal) distribution of any variable* X_i given the values of other variables. Modeling the full joint distribution gives us an extremely powerful exploratory tool. Explorations can be done in the setting, where we study the variable predictive distributions (Bernardo and Smith, 1994) of a new (actual or imaginary) data vector. Here we only want to briefly address some interesting question types that can be answered by such a tool:

- **Variable distributions for a given explaining variable assignment.** In the extreme case we can fix in the new data vector only the value of a background variable, e.g., the sex, after which we can calculate all the marginal predictive distributions. This means that one can study the distribution of any variable conditioned by the fact that the data vector $\vec{d}$ satisfies the assignment. For example in our case study we can fix one teacher education department, and then explore what is the predicted attitude towards readiness for multimedia teaching for teachers that graduated from that particular department.

- **Variable distribution of an explaining variable given some combination of other variable values.** We can reverse the situation in the previous item, and explore the effect of some value combination of variables $X_i, X_j, \ldots$ for predicting a background variable. Again, to give an example, we could explore which of the teacher education departments seems to have given the least readiness to teachers for using computers and multimedia in their teaching.

- **Variable predictive distribution of any variable given some combination of other variable values.** Similarly, based on the model Θ, one could also explore the predictive distribution of any variable X_i given

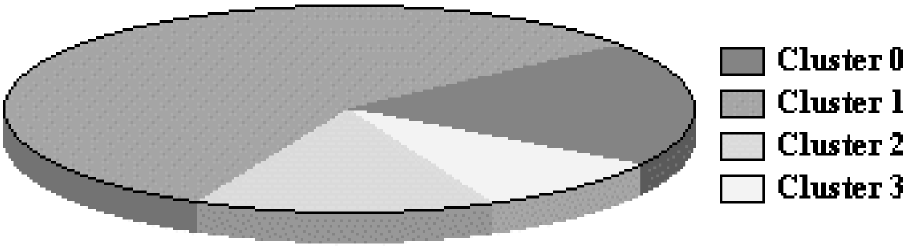

Figure 2: The class "influence" distribution as shown by the NONE tool.

the values of some other variables X_j, X_k, etc. This allows us to see non-linear dependencies between the variable values analogously to the linear correlations in factor analysis.

6 Case study

The finite mixture based approach has been implemented, and runs on a Pentium PC under Linux operating system. Figure 1 illustrates the experimental software tool called NONE, which provides a flexible graphical interface for studying Bayesian finite mixture models, and exploring the predictive distributions. NONE is programmed in Java, and thus can be used with any Java-compatible Internet browser. A running Java^{TM} demo of the software can be accessed through our WWW homepage at URL "http: //www.cs.Helsinki.FI/ research/ cosco/". We will now proceed and illustrate the Bayesian approach described with a case study using the Effectiveness data set. The standard factor analysis results for this same data set are reported in (Niemi and Tirri, 1997).

General explorative analysis As described in Section 5 the methods estimating the domain joint probability distribution can be used in exploring much more complex dependency patterns than simple covariances. This is due to the fact that we are using a more general model family than multivariate normal. However, it is also beneficial to just explore problem domain structure by mixing components, since they are amenable to (sometimes even deceptively) easy interpretation. In order to compare the mixture model approach to the

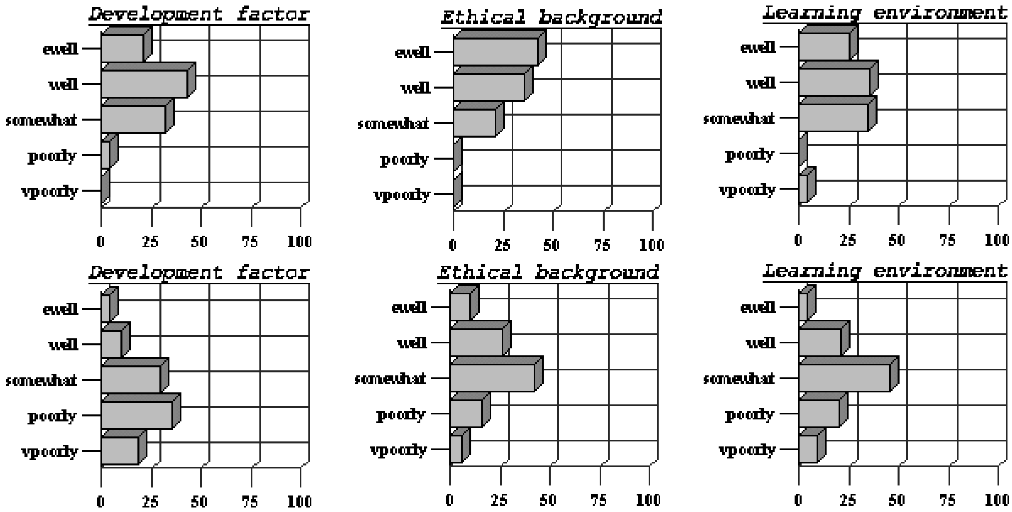

Figure 3: Comparing the marginal distributions of some attributes in the Cluster 2 (above) to the corresponding marginal distibutions of the full distribution (below).

previously run factor analysis we studied a "four class solution" for the data. In Figure 2 we can see the four classes, biggest of which ("Cluster 1" in the Figure) seems to model an "average teacher". This average teacher answers neutrally to most of the questions, and never deviates much from the mean of the population.

On the other hand, 14% of the full domain distribution is influenced by a class ("Cluster 2"), which seems to grasp the tendency that could best be characterized by "Increased social awareness". The most distinctive single feature of this class is the increased awareness of the teacher's role as a development factor in the society. This tendency is accompanied by positive evaluations on the development of teachers own educational philosophies, their awareness of the ethical background of the teacher's profession, and the renovation of the learning environment.

On the other hand one of the classes ("Cluster 0"), seems to model teachers that in general evaluate the teaching received below the average, most notably in issues dealing with internationality and multiculturality, quite unlike the tendency present in Cluster 2.

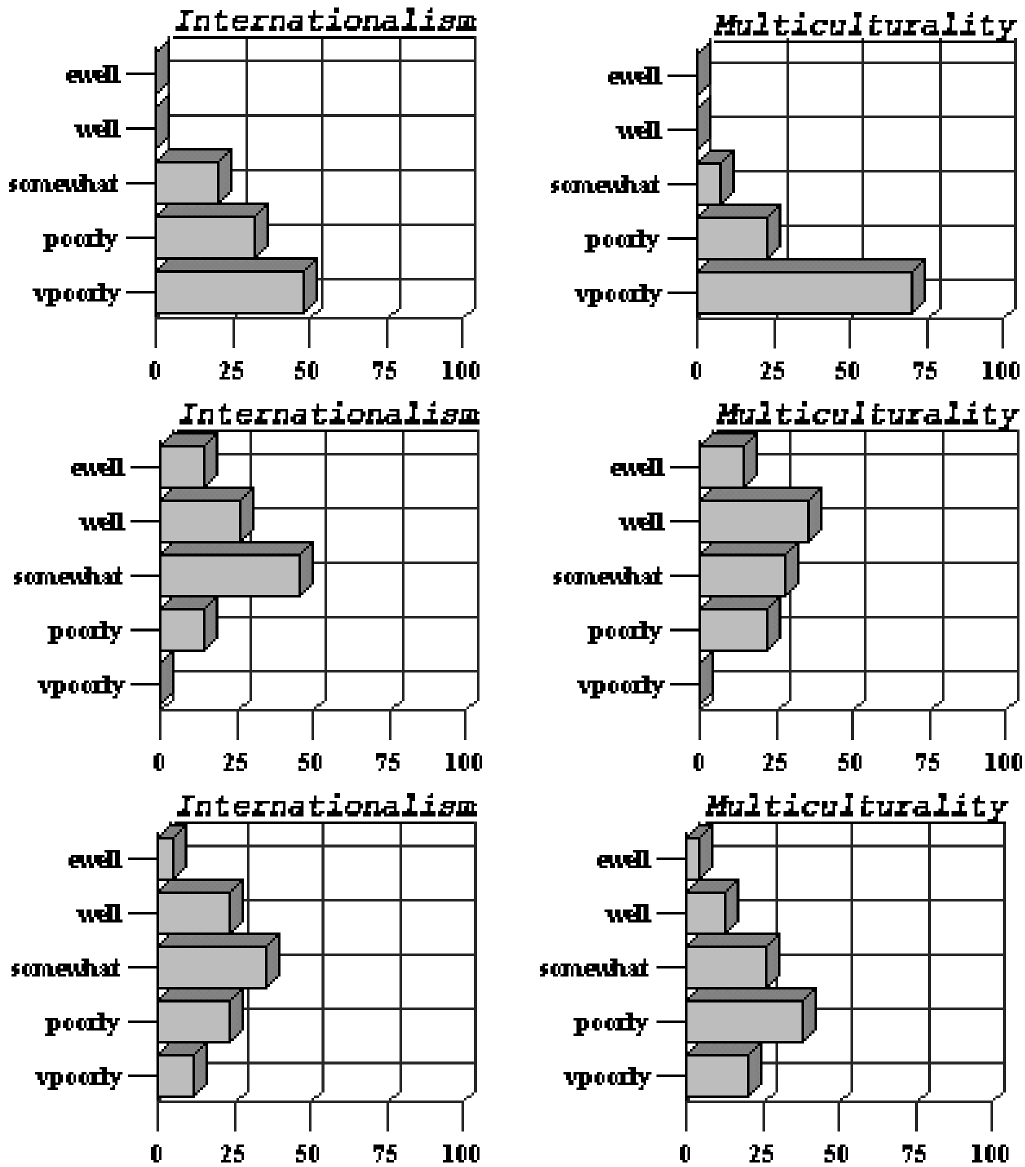

Figure 4: Comparing the marginal distributions of some attributes in the Cluster 0 (above) to the corresponding marginal distibutions of Cluster 2 (middle) and the full distribution (below).

Exploration of more complex dependencies As an example of explorative questions of more complex nature, we may characterize teachers as a function of how they evaluate teaching concerning ethical awareness and readiness to guide students to use modern information technology. Here we can notice that the teachers performing well in both areas can be seen as a balanced mixture of two classes (Clusters 1 and 2). Due to the positive influence of Cluster 2, these teachers also feel that they have received better readiness to promote equality between sexes. Changing awareness to its maximum value changes the situation so that Cluster 2 clearly dominates (85%), and the third class (Cluster 3) also appears as an explaining factor. Here we can also see that readiness to promote equality is even stronger.

To give another example, the teachers feeling that they received good preparedness for their own educational philosophies also seem to be very satisfied with their skills in managing student well-being. Practically no such dependency exists among those who felt that the teacher education did not prepare them to critically reflect their own profession.

It should be observed that due to the increased expressiveness of the model there are exponentially many complex situations one can explore, some of which are more natural and better motivated than others. However, interactive use of NONE tool with joint probability distribution offers the researcher a principled way to study these complex interactions among variables of his/her choice, including situations that are not explicitly present in the sample.

7 Conclusion

In this paper we have discussed some of the methodological issues of using a Bayesian approach with finite mixture models for finding latent classes in the data. We demonstrated that the use of full joint probability models raises interesting methodological questions, some of which were addressed in our discussion. The concepts and methods discussed were illustrated with a case study using an educational data set. The Bayesian classification approach as described here has been implemented, and will be extended in near future. This paper has discussed ongoing research, and more extensive theoretical and experimental treatment as well as comparison to standard approaches is a topic for future work.

Acknowledgements

This research has been supported by the Technology Development Center (TEKES), and by the Academy of Finland.

References

Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods. Theory and Applications*. John Wiley & Sons, New York.

Bernardo, J. and Smith, A. (1994). *Bayesian theory*. John Wiley.

Chatfield, C. a. A. C. (1980). *Introduction to Multivariate Analysis*. Chapman and Hall, New York.

Cheeseman, P. and Stutz, J. (1996). Bayesian classification (AutoClass): Theory and results. In Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors, *Advances in Knowledge Discovery and Data Mining*, chapter 6. AAAI Press, Menlo Park.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. John Wiley & Sons, New York, NY.

Everitt, B. and Hand, D. (1981). *Finite Mixture Distributions*. Chapman and Hall, London.

Gilks, W. R., Richardson, S., and J., S. D. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall, London, GB.

Heckerman, D. (1996). A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division, One Microsoft Way, Redmond, WA 98052.

Heckerman, D., Geiger, D., and Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243.

Klecka, W. (1981). *Discriminant analysis*. Sage Publications, Beverly Hills, CA.

Kontkanen, P., Myllymäki, P., and Tirri, H. (1996a). Comparing Bayesian model class selection criteria by discrete finite mixtures. In Dowe, D., Korb, K., and Oliver, J., editors, *Information, Statistics and Induction in Science*, pages 364–374, Proceedings of the ISIS'96 Conference, Melbourne, Australia. World Scientific, Singapore.

Kontkanen, P., Myllymäki, P., and Tirri, H. (1996b). Predictive data mining with finite mixtures. In Simoudis, E., Han, J., and Fayyad, U., editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 176–182, Portland, Oregon.

Kontkanen, P., Myllymäki, P., and Tirri, H. (1997). Experimenting with the Cheeseman-Stutz evidence approximation for predictive modeling and data mining. In *Proceedings of Tenth International FLAIRS Conference (to appear)*, Daytona Beach, Florida.

Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.

Niemi, H. and Tirri, K. (1996). Effectiveness of teacher education. New challenges and approaches to evaluation. Technical Report A 6/1996, Department of Teacher Education in Tampere University.

Niemi, H. and Tirri, K. (1997). Readiness for teaching profession evaluated by teachers and teacher educators. In Press.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society*, 49(3):223–239 and 252–265.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Company, New Jersey.

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, New York.

Wallace, C. and Boulton, D. (1968). An information measure for classifiation. *Computer Journal*, 11:185–194.

Wallace, C. and Freeman, P. (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society*, 49(3):240–265.