Variable Selection for Model-Based Clustering

Adrian E. RAFTERY and Nema DEAN

We consider the problem of variable or feature selection for model-based clustering. The problem of comparing two nested subsets of variables is recast as a model comparison problem and addressed using approximate Bayes factors. A greedy search algorithm is proposed for finding a local optimum in model space. The resulting method selects variables (or features), the number of clusters, and the clustering model simultaneously. We applied the method to several simulated and real examples and found that removing irrelevant variables often improved performance. Compared with methods based on all of the variables, our variable selection method consistently yielded more accurate estimates of the number of groups and lower classification error rates, as well as more parsimonious clustering models and easier visualization of results.

KEY WORDS: Bayes factor; BIC; Feature selection; Model-based clustering; Unsupervised learning; Variable selection.

1. INTRODUCTION

In classification, or supervised learning problems, the structure of interest often may be contained in only a subset of the available variables, and inclusion of unnecessary variables in the learning procedure may degrade the results. In these cases some form of variable selection made before or incorporated into the fitting procedure may be advisable. Similarly, in clustering, or unsupervised learning problems, the structure of greatest interest to the investigator may be best represented using only a few of the feature variables. This may give the best clustering model to describe future data, or fewer variables may give a better partition of the data into clusters closer to the true underlying group structure. But in clustering the classification is not observed, and there is usually little or no a priori knowledge of the structure being looked for in the analysis, so there is no simple preanalysis screening technique available to use. In this case it makes sense to consider including the variable selection procedure as part of the clustering algorithm.

In this article we introduce a method for variable or feature selection for model-based clustering. The basic idea is to recast the variable selection problem as one of comparing competing models for all of the variables initially considered. Comparing two nested subsets is equivalent to comparing two models, in one of which all of the variables in the bigger subset carry information about cluster membership, whereas in the other the variables considered for exclusion are conditionally independent of cluster membership given the variables included in both models. This comparison is made using approximate Bayes factors. This model comparison criterion is combined with a greedy search algorithm to give an overall method for variable selection. The resulting method selects the clustering variables, the number of clusters, and the clustering model simultaneously.

The variable selection procedure suggested in this article is tailored specifically for model-based clustering and as such incorporates the advantages of this paradigm relative to some of the more heuristic clustering algorithms. These advantages include an automatic method for choosing the number of clusters, the need for only one user-defined input (the maximum number of clusters to be considered) that is easily interpretable, and a basis in statistical inference.

Adrian E. Raftery is Professor of Statistics and Sociology (E-mail: *raftery@stat.washington.edu*) and Nema Dean is Graduate Research Assistant (E-mail: *nemad@stat.washington.edu*), Department of Statistics, University of Washington, Seattle, WA 98195. This research was supported by National Institutes of Health grant 8 R01 EB002137-02. The authors thank Chris Fraley, Peter W. F. Smith, and the associate editor and referees for helpful comments.

A brief review of model-based clustering is given in Section 2.1. The statistical model behind the variable selection method is explained in Section 2.2, and the greedy search algorithm is introduced in Section 2.3. The specific example of clustering with Gaussian components and allowing different covariance formulations is discussed in Section 2.4. Results comparing the performance of model-based clustering with and without variable selection are given in Section 3 for simulated data and in Section 4 for some real data examples. The advantages and limitations of the method are discussed in Section 5, where some other work on the problem is also mentioned.

2. METHODOLOGY

2.1 Model-Based Clustering

Model-based clustering is based on the idea that the observed data come from a population with several subpopulations. The general idea is to model each of the subpopulations separately and the overall population as a mixture of these subpopulations, using finite mixture models. Model-based clustering goes back at least to work of Wolfe (1963), and reviews of the area have been given by McLachlan and Peel (2000) and Fraley and Raftery (2002).

The general form of a finite mixture model with G subpopulations or groups is

$$f(\mathbf{x}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}),$$

where π_g is the proportion of the population in the gth group and $f_g(\cdot)$ is the probability density function for the gth group. The subpopulations are often modeled by members of the same parametric density family, in which case the finite mixture model can be written as

$$f(\mathbf{x}) = \sum_{g=1}^{G} \pi_g f(\mathbf{x}|\boldsymbol{\phi}_g),$$

where $\boldsymbol{\phi}_g$ is the parameter vector for the gth group.

The mixture model can be used to partition the data into clusters using the optimal or Bayes rule for classification. This classifies observation $\mathbf{x}$ to cluster g if the posterior probability that it belongs to group g is greater than the posterior proba-

Journal of the American Statistical Association
March 2006, Vol. 101, No. 473, Theory and Methods
DOI 10.1198/016214506000000113

bilities that it belongs to any other group; that is, if $\tau_g(\mathbf{x}) > \tau_h(\mathbf{x})$, $h = 1, \ldots, G$, where $\tau_h(\mathbf{x}) = \pi_h f_h(\mathbf{x}) / \sum_{g=1}^{G} \pi_g f_g(\mathbf{x})$ is the posterior probability that it belongs to the hth group. Because the denominator is the same in all posterior probabilities, it is possible to simply define the Bayes rule as follows: Classify $\mathbf{x}$ into cluster g if $g = \arg\max_h \pi_h f_h(\mathbf{x})$. We can approximate the Bayes rule by replacing the unknown parameters by their estimated values. This is called the plug-in rule. In our examples, we compare the partition given by the finite mixture model defined on the subset of selected variables with a known underlying classification, to assess how much improvement the variable selection procedure gives in clustering.

A difficulty of some of the more heuristic clustering algorithms is the lack of a statistically principled method for determining the number of clusters. Because it is an inferentially based procedure, model-based clustering can use model selection methods to make this decision. Bayes factors (Kass and Raftery 1995) are used to compare the models. This permits comparison of the nonnested models that arise in this context.

The Bayes factor for a model M_1 against a competing model M_2 is equal to the posterior odds for M_1 against M_2 when their prior model probabilities are equal. It is computed as the ratio of the integrated likelihoods for the two models. This ratio can be hard to compute, and we use the easily calculated Bayesian information criterion (BIC) as the basis for an approximation. This is defined by

$$BIC = 2 \times \log(\text{maximized likelihood}) - (\text{no. of parameters}) \times \log(n), \quad (1)$$

where n is the number of observations. Twice the logarithm of the Bayes factor is approximately equal to the difference between BIC values for the two models being compared. We choose the number of groups and the parametric model by recognizing that each different combination of number of groups and parametric constraints defines a model, which can then be compared with others. Keribin (1998) showed BIC to be consistent for the choice of the number of clusters. Differences of less than 2 between BIC values are typically viewed as barely worth mentioning, whereas differences greater than 10 are often regarded as constituting strong evidence (Kass and Raftery 1995).

2.2 Model-Based Variable Selection

To address the variable selection problem, we recast it as a model selection problem. We have a dataset Y, and at any stage in our variable selection algorithm, it is partitioned into three sets of variables, $Y^{(1)}$, $Y^{(2)}$, and $Y^{(3)}$, as follows:

- $Y^{(1)}$, the set of already selected clustering variables
- $Y^{(2)}$, the variable(s) being considered for inclusion into or exclusion from the set of clustering variables
- $Y^{(3)}$, the remaining variables.

The decision for inclusion or exclusion of $Y^{(2)}$ from the set of clustering variables is then recast as one of comparing the following two models for the full dataset:

$$\begin{aligned} M_1: \quad & p(Y|\mathbf{z}) \\ & = p\big(Y^{(1)}, Y^{(2)}, Y^{(3)}|\mathbf{z}\big) \\ & = p\big(Y^{(3)}|Y^{(2)}, Y^{(1)}\big) p\big(Y^{(2)}\big|Y^{(1)}\big) p\big(Y^{(1)}|\mathbf{z}\big), \\ M_2: \quad & p(Y|\mathbf{z}) \\ & = p\big(Y^{(1)}, Y^{(2)}, Y^{(3)}|\mathbf{z}\big) \\ & = p\big(Y^{(3)}\big|Y^{(2)}, Y^{(1)}\big) p\big(Y^{(2)}, Y^{(1)}|\mathbf{z}\big), \end{aligned} \quad (2)$$

where $\mathbf{z}$ is the (unobserved) set of cluster memberships. Model M_1 specifies that given $Y^{(1)}$, $Y^{(2)}$ is conditionally independent of the cluster memberships (defined by the unobserved variables $\mathbf{z}$); that is, $Y^{(2)}$ gives no additional information about the clustering. Model M_2 implies that $Y^{(2)}$ does provide additional information about clustering membership, after $Y^{(1)}$ has been observed.

An important aspect of the model formulation is that it does not require that irrelevant variables be independent of the clustering variables. If instead the independence assumption $p(Y^{(2)}|Y^{(1)}) = p(Y^{(2)})$ were used in model M_1, then we would be quite likely to include redundant variables related to the clustering variables but not to the clustering itself. We assume that the remaining variables $Y^{(3)}$ are conditionally independent of the clustering given $Y^{(1)}$ and $Y^{(2)}$ and belong to the same parametric family in both models. The difference between the assumptions underlying the two models is illustrated in Figure 1, where arrows indicate dependency.

Models M_1 and M_2 are compared via an approximation to the Bayes factor that allows the high-dimensional $p(Y^{(3)}|Y^{(2)}, Y^{(1)})$ to cancel from the ratio. The Bayes factor, B_{12}, for M_1 against M_2 based on the data Y is given by

$$B_{12} = \frac{p(Y|M_1)}{p(Y|M_2)},$$

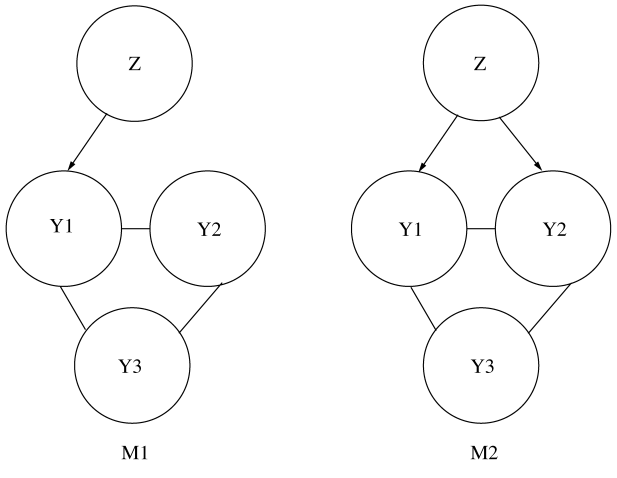

Figure 1. Graphical Representation of Models M_1 and M_2 for Clustering Variable Selection. In model M_1, the candidate set of additional clustering variables, $Y^{(2)}$, is conditionally independent of the cluster memberships, $\mathbf{z}$, given the variables $Y^{(1)}$ already in the model. In model M_2, this is not the case. In both models, the set of other variables considered, $Y^{(3)}$, is conditionally independent of cluster membership given $Y^{(1)}$ and $Y^{(2)}$, but may be associated with $Y^{(1)}$ and $Y^{(2)}$.

where $p(Y|M_k)$ is the integrated likelihood of model M_k, $k = 1, 2$, namely

$$p(Y|M_k) = \int p(Y|\boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k|M_k)\,d\boldsymbol{\theta}_k. \quad (3)$$

In (3), $\boldsymbol{\theta}_k$ is the vector-valued parameter of model M_k and $p(\boldsymbol{\theta}_k|M_k)$ is its prior distribution (Kass and Raftery 1995).

Let us now consider the integrated likelihood of model M_1, $p(Y|M_1) = p(Y^{(1)}, Y^{(2)}, Y^{(3)}|M_1)$. From (2), the model M_1 is specified by three probability distributions: the finite mixture model that specifies $p(Y^{(1)}|\boldsymbol{\theta}_1, M_1)$ and the conditional distributions $p(Y^{(2)}|Y^{(1)}, \boldsymbol{\theta}_1, M_1)$ and $p(Y^{(3)}|Y^{(2)}, Y^{(1)}, \boldsymbol{\theta}_1, M_1)$. We denote the parameter vectors that specify these three probability distributions by θ_{11}, θ_{12}, and θ_{13}, and we assume that their prior distributions are independent. It follows that the integrated likelihood itself factors as follows:

$$p(Y|M_1) = p\big(Y^{(3)}|Y^{(2)}, Y^{(1)}, M_1\big) \times p\big(Y^{(2)}|Y^{(1)}, M_1\big)p\big(Y^{(1)}|M_1\big), \quad (4)$$

where $p(Y^{(3)}|Y^{(2)}, Y^{(1)}, M_1) = \int p(Y^{(3)}|Y^{(2)}, Y^{(1)}, \theta_{13}, M_1) \times p(\theta_{13}|M_1)\,d\theta_{13}$, and similarly for $p(Y^{(2)}|Y^{(1)}, M_1)$ and $p(Y^{(1)}|M_1)$. Similarly, we obtain

$$p(Y|M_2) = p\big(Y^{(3)}|Y^{(2)}, Y^{(1)}, M_2\big)p\big(Y^{(2)}, Y^{(1)}|M_2\big), \quad (5)$$

where $p(Y^{(2)}, Y^{(1)}|M_2)$ is the integrated likelihood for the model-based clustering model for $(Y^{(2)}, Y^{(1)})$ jointly.

The prior distribution of the parameter, θ_{13}, is assumed to be the same under M_1 as under M_2. It follows that $p(Y^{(3)}|Y^{(2)}, Y^{(1)}, M_2) = p(Y^{(3)}|Y^{(2)}, Y^{(1)}, M_1)$. We thus have

$$B_{12} = \frac{p(Y^{(2)}|Y^{(1)}, M_1)p(Y^{(1)}|M_1)}{p(Y^{(2)}, Y^{(1)}|M_2)}, \quad (6)$$

which has been greatly simplified by the cancellation of the factors involving the potentially high-dimensional $Y^{(3)}$. The integrated likelihoods in (6) are hard to evaluate analytically, and so we approximate them using the BIC approximation of (1).

2.3 Combined Variable Selection and Clustering Procedure

Here we propose a greedy search algorithm. At each stage it searches for the variable to add that most improves the clustering as measured by BIC, and then assesses whether one of the current clustering variables can be dropped. At each stage, the best combination of number of groups and clustering model is chosen. The algorithm stops when no local improvement is possible.

Here is a summary of the algorithm:

1. Select the first clustering variable to be the one that has the most evidence of univariate clustering.
2. Select the second clustering variable to be the one that shows the most evidence of bivariate clustering including the first variable selected.
3. Propose the next clustering variable to be the one that shows the most evidence of multivariate clustering including the previous variables selected. Accept this variable as a clustering variable if the evidence favors this over its not being a clustering variable.
4. Propose the variable for removal from the current set of selected clustering variables to be the one for which the evidence of multivariate clustering including all variables selected versus multivariate clustering only on the other variables selected and not on the proposed variable is weakest. Remove this variable from the set of clustering variables if the evidence for clustering is weaker than that for not clustering.
5. Iterate steps 3 and 4 until two consecutive steps have been rejected, then stop.

2.4 Variable Selection for Gaussian Model-Based Clustering

We now consider in more detail the case where the mixture components have multivariate normal distributions, that is, $f(\cdot|\boldsymbol{\phi}_g) = \mathrm{MVN}(\cdot|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$. For variable selection in this setting, we consider only the case where $Y^{(2)}$ contains just one variable, in which case $p(Y^{(2)}|Y^{(1)}, M_1)$ represents a normal linear regression model with an intercept and main effects only. This follows from the standard result for conditional multivariate normal means. The BIC approximation to this term in (6) is

$$\begin{aligned} &2\log p\big(Y^{(2)}|Y^{(1)}, M_1\big) \\ &\quad \approx BIC_{\mathrm{reg}} \\ &\quad = -n\log(2\pi) \\ &\qquad - n\log(\mathrm{RSS}/n) - n - \big(\dim\big(Y^{(1)}\big) + 2\big)\log(n), \end{aligned} \quad (7)$$

where RSS is the residual sum of squares in the regression of $Y^{(2)}$ on the variables in $Y^{(1)}$.

One practical issue with multivariate normal modeling of the components is that if the model is unconstrained, the number of parameters grows rapidly with the dimension and with the number of clusters, leading to possible overfitting and degradation of performance. For instance, our first example in Section 4 is fairly small, with 4 groups and 5 variables, but the full multivariate normal mixture model still has 83 parameters.

One way of alleviating this is to impose restrictions on the covariance matrices. The covariance matrix can be decomposed, as was done by Banfield and Raftery (1993) and Celeux and Govaert (1995), as

$$\boldsymbol{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g,$$

where λ_g is the largest eigenvalue of $\boldsymbol{\Sigma}_g$ and controls the volume of the gth cluster; $\mathbf{D}_g$ is the matrix of eigenvectors of $\boldsymbol{\Sigma}_g$, which control the orientation of that cluster; and $\mathbf{A}_g$ is a diagonal matrix with the scaled eigenvalues as entries, which control the shape of that cluster. By imposing constraints on the various elements of this decomposition, a large range of models is available, ranging from the simple spherical models that have fixed shape to the least parsimonious model where all elements of the decomposition are allowed to vary across all clusters. A list of the models available in the *mclust* software (Fraley and Raftery 2003), which allows this type of eigenvalue decomposition Gaussian clustering, is given in Table 1. We can choose the parametric model by recognizing that each different combinations of number of groups and parametric constraints defines a model, which can then be compared with others using BIC.

Table 1. Parameterizations of the Covariance Matrix $\boldsymbol{\Sigma}_g$ Currently Available in the mclust *Software*

Name	*Model*	*Distribution*	*Volume*	*Shape*	*Orientation*
EII	$\lambda\mathbf{I}$	Spherical	Equal	Equal	NA
VII	$\lambda_g\mathbf{I}$	Spherical	Variable	Equal	NA
EEI	$\lambda\mathbf{A}$	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_g\mathbf{A}$	Diagonal	Variable	Equal	Coordinate axes
EVI	$\lambda\mathbf{A}_g$	Diagonal	Equal	Variable	Coordinate axes
VVI	$\lambda_g\mathbf{A}_g$	Diagonal	Variable	Variable	Coordinate axes
EEE	$\lambda\mathbf{D}\mathbf{A}\mathbf{D}^T$	Ellipsoidal	Equal	Equal	Equal
VVV	$\lambda_g\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g^T$	Ellipsoidal	Variable	Variable	Variable
EEV	$\lambda\mathbf{D}_g\mathbf{A}\mathbf{D}_g^T$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_g\mathbf{D}_g\mathbf{A}\mathbf{D}_g^T$	Ellipsoidal	Variable	Equal	Variable

NOTE: When the data are of dimension 1, only two models are available: equal variances (E) and unequal variances (V).

Different choices of subsets of clustering variables also require different covariance structures for the subpopulations. In our examples, we used the *mclust* software, but the method could also be implemented using other mixture modeling software. We used hierarchical agglomerative model-based clustering to give the starting values needed in the EM algorithm used to estimate model parameters.

3. SIMULATION EXAMPLES

We now present results for two simulation examples. Here we use the term "groups" to refer to the true unknown partition and the term "clusters" to refer to the partition estimated by the clustering algorithm.

3.1 First Simulation Example: Two Groups

In this simulation there are 150 data points on 7 variables. The data are simulated from a mixture of two multivariate normal distributions with unconstrained (VVV) covariance matrices, so that there are two groups in the data. Only the first two variables contain clustering information. The remaining five variables are irrelevant variables independent of the clustering variables, so that the distribution of these variables is multivariate normal independent of group membership. The pairs plot of all the variables is given in Figure 2, where X1 and X2 are the clustering variables and X3–X7 are the independent irrelevant variables.

When forced to cluster on all seven variables, a five-cluster diagonal EEI model yields the highest BIC value. The model yielding the next highest BIC value is a four-cluster EEI model. The two-cluster model with the highest BIC value is the two-cluster EEE model, but there is a substantial difference (20 points) between this and the model with the highest BIC. This would lead to the (incorrect) choice of a five-group structure for this data.

The step-by-step progress of the greedy search selection procedure is shown in Table 2. Two variables are chosen, X1 and X2; these are the correct clustering variables. The two-cluster VVV model has the highest BIC for clustering on these variables by a decisive margin; this gives both the correct number of groups and the correct clustering model (two VVV clusters).

Because the data are simulated, we know the underlying group memberships of the observations, and we can check the quality of the clustering in this way. The partition arising from clustering on the selected two variables gives 100% correct classification. The confusion matrix for the clustering on all variables is as follows:

	Group 1	*Group* 2
Cluster 1	53	0
Cluster 2	4	30
Cluster 3	34	0
Cluster 4	1	13
Cluster 5	0	15

The error rate is 44.7%. This is calculated by taking the best matches of clusters with the groups (i.e., group 1 $\leftrightarrow$ cluster 1 and group 2 $\leftrightarrow$ cluster 2), which gives us the minimum error rate over all matches between clusters and groups. If we were to correctly amalgamate clusters 1 and 3 and identify them as one cluster, and to amalgamate clusters 2, 4, and 5 and identify them as another cluster, we would get an error rate of 3.3%. However, this assumes knowledge that the investigator would not typically have in practice.

Finally, we pretend (as do many heuristic clustering algorithms) that we know the number of groups (two) correctly in advance, and cluster on all of the variables allowing only two-cluster models. The two-cluster model with the highest BIC is the EEE model, which has an error rate of 3.3%.

In this example, variable selection led to a clustering method that gave the correct number of groups and a 0% error rate. Using all variables led to considerable overestimation of the number of groups and a large error rate. Even when the five clusters found in this way were optimally combined into two clusters (with their own mixtures), or when the correct number of groups was assumed known, clustering using all of the variables led to a nonzero error rate with five errors. Table 3 summarizes the classification results.

3.2 Second Simulation Example: Irrelevant Variables Correlated With Clustering Variables

Again, we have 150 points from 2 clustering variables, with 2 (VVV) groups. To make the problem more difficult, we allow different types of irrelevant variables. There are three independent normal irrelevant variables, seven irrelevant variables that are allowed to be dependent on other irrelevant variables (multivariate normal), and three irrelevant variables that have a linear relationship with either or both of the clustering variables. This gives a total of 15 irrelevant variables.

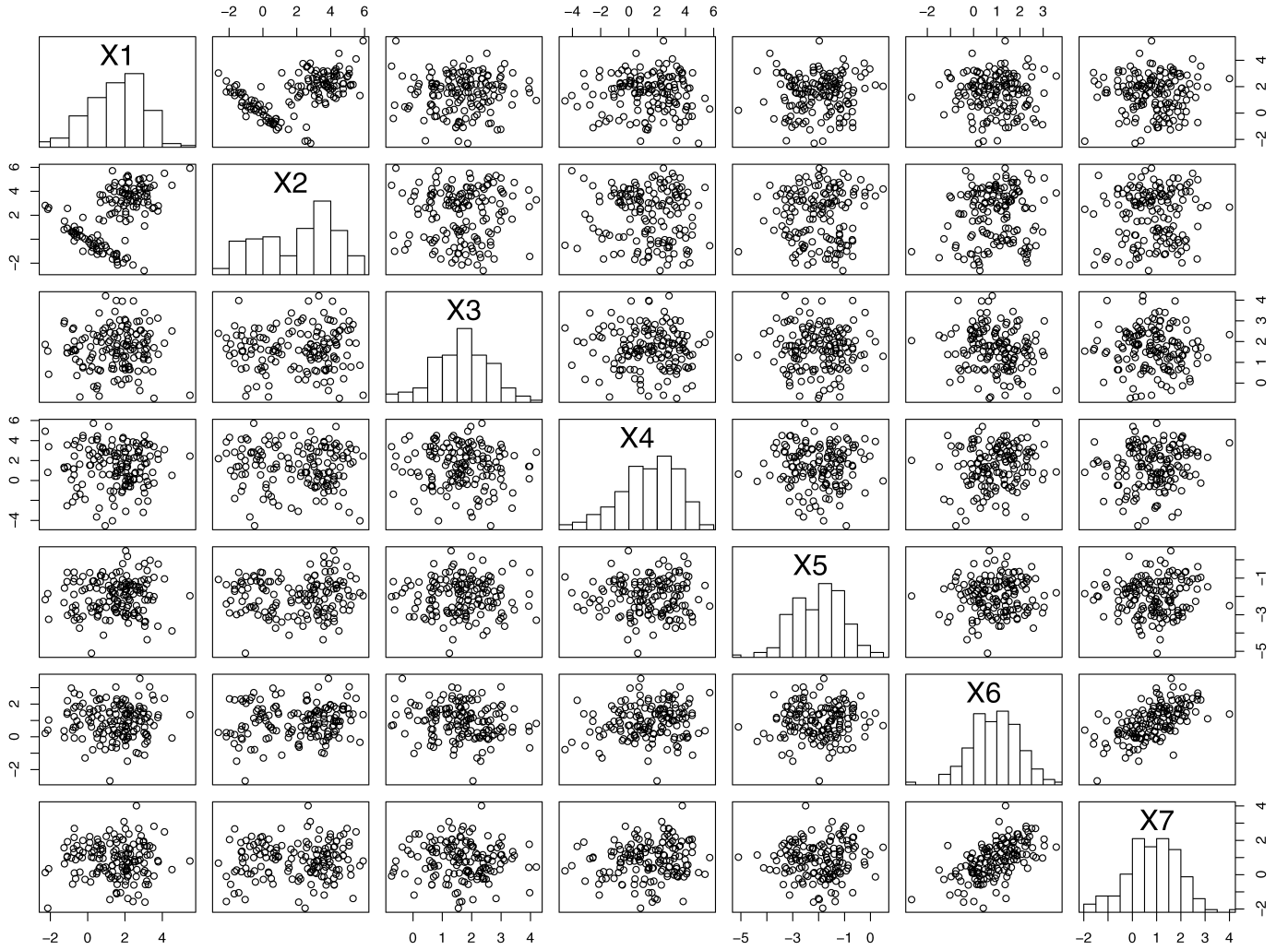

Figure 2. First Simulation Example: Pairs Plot of the Data.

The pairs plot of a selection of the variables is given in Figure 3. Variables X1 and X2 are the clustering variables, variable X3 is an independent irrelevant variable, variables X6 and X7 are irrelevant variables that are correlated with one another, variable X13 is linearly dependent on the clustering variable X1, variable X14 is linearly dependent on the clustering variable X2, and variable X15 is linearly dependent on both clustering variables X1 and X2.

When forced to cluster on all 15 variables, a 2-cluster diagonal EEI model yields the highest BIC. The model yielding the next-highest BIC value is a 3-cluster diagonal EEI model, with a difference of 10 points between the two. In this case the investigator would probably decide on the correct number of groups, based on this evidence. The error rate for classification based on this model is 1.3%.

The results of the steps when the greedy search selection procedure is run are given in Table 4. Two variables are selected, and these are precisely the correct clustering variables. The model with the highest BIC for clustering on these variables is a two-cluster VVV model, and the next highest model in terms of BIC is the three-cluster VVV model. There is a difference of 27 between the 2 BIC values, which typically would be considered strong evidence.

We compare the clustering memberships with the underlying group memberships and find that clustering on the selected variables gives a 100% correct classification, that is, no errors. In

Table 2. Individual Step Results From Greedy Search Algorithm for the First Simulation

Step no.	*Best variable proposed*	*Proposed for*	*BIC difference*	*Model chosen*	*Number of clusters chosen*	*Result*
1	X2	Inclusion	15	V	2	Included
2	X1	Inclusion	136	VVV	2	Included
3	X6	Inclusion	−13	VVV	2	Not included
4	X1	Exclusion	136	VVV	2	Not excluded

NOTE: The BIC difference is the difference between the BIC for clustering and the BIC for not clustering for the best variable proposed, as given in (A.1) in the Appendix.

Table 3. Classification Results for the First Simulation Example

Variable selection procedure	*Number of variables*	*Number of clusters*	*Error rate (%)*
None (all variables)	7	5	44.7
None (all variables)	7	2(c)	3.3
Greedy search	2	2	0

NOTE: The correct number of groups was 2. (c) indicates that the solution was constrained to this number of clusters.

contrast, using all 15 variables gives a nonzero error rate, with 2 errors. Variable selection has the added advantage in this example that it makes the results easy to visualize, because only two variables are involved after variable selection. Table 5 summarizes the classification results.

4. EXAMPLES

We now give the results of applying our variable selection method to three real datasets in which the correct number of groups is known.

4.1 *Leptograpsus* Crabs Data

This dataset consists of 200 subjects: 100 of species orange (50 male and 50 female) and 100 of species blue (50 male and 50 female). This gives a possible four-group classification, so we are hoping to find a four-cluster structure. There are five measurements on each subject: width of frontal lip (FL), rear width (RW), length along the midline of the carapace (CL) maximum width of the carapace (CW) and body depth (BD) in mm. The dataset was published by Campbell and Mahon (1974) and was further analyzed by Ripley (1996) and McLachlan and Peel (1998, 2000).

The variables selected by the variable selection procedure were (in order of selection) CW, RW, FL, and BD. The error rates for the different clusterings are given in Table 6. The error rates for the seven-cluster models were the minimum error rates over all matchings between clusters and groups, where each group was matched with a unique cluster.

When no variable selection was done, the number of groups was substantially overestimated, and the error rate was 42.5%, as can be seen in the following confusion matrix for clustering on all variables:

	Group 1	*Group* 2	*Group* 3	*Group* 4
Cluster 1	32	0	0	0
Cluster 2	0	31	0	0
Cluster 3	0	0	28	0
Cluster 4	0	0	0	24
Cluster 5	0	0	0	21
Cluster 6	18	19	0	0
Cluster 7	0	0	22	5

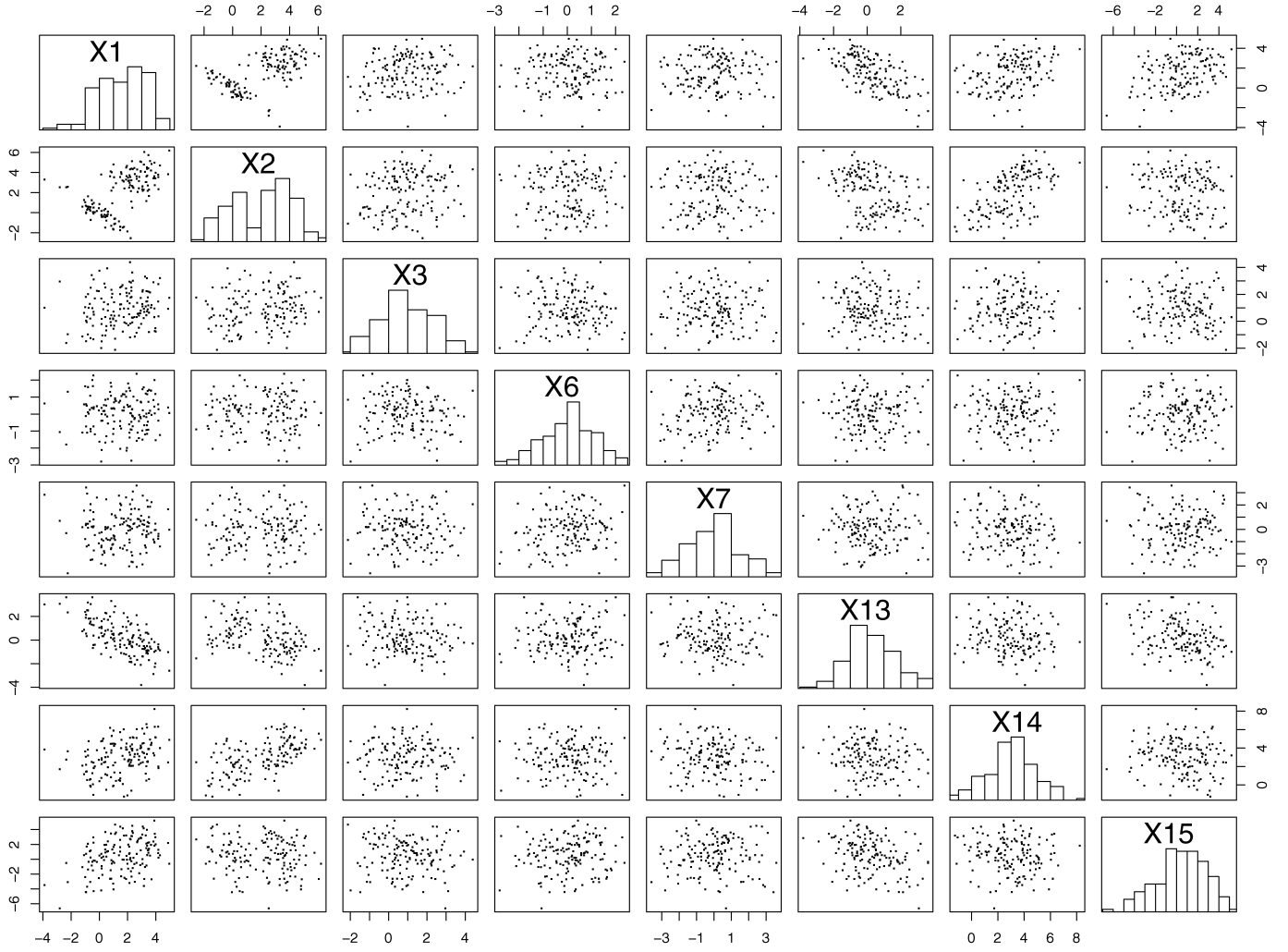

Figure 3. Second Simulation Example: Pairs Plot of 8 of the 15 Variables.

Table 4. Individual Step Results From the Greedy Search Algorithm for the Second Simulation Example

Step no.	Best variable proposed	Proposed for	BIC difference	Model chosen	Number of clusters chosen	Result
1	X11	Inclusion	17	V	2	Included
2	X2	Inclusion	5	EEE	2	Included
3	X1	Inclusion	109	VVV	2	Included
4	X11	Exclusion	−19	VVV	2	Excluded
5	X4	Inclusion	−9	VVV	2	Not included
6	X2	Exclusion	153	VVV	2	Not excluded

NOTE: The BIC difference is the difference between the BIC for clustering and the BIC for not clustering for the best variable proposed, as given in (A.1) in the Appendix.

When our variable selection method was used, the correct number of groups was selected, and the error rate was much lower (7.5%), as can be seen in the following confusion matrix for the clustering on the selected variables:

	Group 1	*Group* 2	*Group* 3	*Group* 4
Cluster 1	40	0	0	0
Cluster 2	10	50	0	0
Cluster 3	0	0	50	5
Cluster 4	0	0	0	45

Variable selection reduced the number of classification errors to a striking extent, especially given that the method selected four of the five variables, so not much variable selection was actually done in this case. This example suggests that the presence of only a single noise variable in even a low-dimensional setting can cause the clustering results to deteriorate.

In clustering it is common practice to work with principal components of the data and to select the first several as a way of reducing the data dimension. Our method could be used as a way of choosing the principal components to be used, and it has the advantage that one does not have to use the principal components that explain the most variation, but can automatically select the principal components that are most useful for clustering. To illustrate this, we computed the five principal components of the data and used these instead of the variables. The variable selection procedure chose (in order) principal components 3, 2, and 1.

Once again, when all of the principal components were used, the number of groups was overestimated, and the error rate was high, at 42.5%. When variable selection was carried out, our method selected the correct number of groups without invoking any previous knowledge of the number of groups, and the error rate was much reduced, at 6.5%. Even when the number of groups was assumed to be correctly known in advance but no variable selection was done, the error rate was higher than with variable selection, at 9.0%.

Chang (1983) showed that the practice of reducing the data to the principal components that account for the most variability before clustering is not justified in general. He showed that the principal components with the largest eigenvalues do not necessarily contain the most information about the cluster structure, and that taking a subset of principal components can lead to a major loss of information about the groups in the data. Chang demonstrated this theoretically, by simulations, and in applications to real data. Similar results have been found by other researchers, including Green and Krieger (1995) for market segmentation and Yeung and Ruzzo (2001) for clustering gene expression data. Our method rescues the principal component reduction dimension reduction approach to some extend, because it allows one to use all or many of the principal components and then for clustering select only those that are most useful for clustering, not those that account for the most variance. This avoids Chang's criticism.

In this example the EM algorithm used for estimating the parameters when clustering on all variables was sensitive to starting values, and the best starting values came from randomly generating posterior probabilities rather than from hierarchical agglomerative model-based clustering. The variable selection EM clustering was not as sensitive to the starting values, and hierarchical clustering was used to initialize the EM algorithm in that case.

Table 5. Classification Results for the Second Simulation Example

Variable selection procedure	Number of variables	Number of clusters	Error rate (%)
None (all variables)	15	2	1.3
Greedy search	2	2	0

4.2 Iris Data

The well-known Iris data consist of four measurements on 150 samples of either *Iris setosa*, *Iris versicolor*, or *Iris virginica* (Anderson 1935; Fisher 1936). The measurements are sepal length, sepal width, petal length, and petal width (cm). When one clusters using all of the variables, the model with the highest BIC is the two-cluster VEV model, with the three-group VEV model within one BIC point of it. Thus an analyst might conclude that these data do not contain enough information to determine whether there are two or three groups. The two-group clustering puts *I. versicolor* and *I. virginica* together, and these are known to be very closely related; their identification as separate species is based in part on information not included in this dataset (Anderson 1936). If one does select the two-group clustering model favored slightly by BIC, the confusion matrix is as follows:

	I. setosa	*I. versicolor*	*I. virginica*
Cluster 1	50	0	0
Cluster 2	0	50	50

The *I. setosa* group is well picked out, but *I. versicolor* and *I. virginica* have been amalgamated. This leads to a minimum error rate of 33.3%.

Table 6. Classification Results for the Crabs Data

Variable selection procedure	*Number of variables/components*	*Number of clusters*	*Model selected*	*Error rate (%)*
Original variables				
None (all variables)	5	7	EEE	42.5
None (all variables)	5	4(c)	EEE	7.5
Greedy search	4	4	EEV	7.5
Principal components				
None (all components)	5	7	EEE	42.5
None (all components)	5	4(c)	EEV	9.0
Greedy search	3	4	EEV	6.5

NOTE: The correct number of groups is four. (c) indicates that the number of clusters was constrained to this value in advance. The error rates for the seven-cluster models were calculated by optimally matching clusters to groups.

The confusion matrix from the three-group clustering is as follows:

	I. setosa	*I. versicolor*	*I. virginica*
Cluster 1	50	0	0
Cluster 2	0	45	0
Cluster 3	0	5	50

This gives a 3.3% error rate and reasonable separation. However, given the BIC values, an investigator with no reason to do otherwise might well have erroneously chosen the two-cluster model.

The variable selection procedure selects three variables (all but sepal length). The highest BIC model is the three-cluster VEV model, with the next highest model being the four-cluster VEV model; the BIC difference is 14. The confusion matrix from the three-cluster clustering on these variables is as follows:

	I. setosa	*I. versicolor*	*I. virginica*
Cluster 1	50	0	0
Cluster 2	0	44	0
Cluster 3	0	6	50

This reflects a 4% error rate. A summary of the results from the different methods is given in Table 7. For these data, clustering on all variables gives an ambiguous result, whereas the correct number of groups is decisively chosen when variable selection is done.

4.3 Texture Dataset

The texture dataset was produced by the Laboratory of Image Processing and Pattern Recognition (INPG–LTIRF) in the development of the Esprit project ELENA No. 6891 and the Esprit working group ATHOS No. 6620. The original source was Brodatz (1966). This dataset consists of 5,500 observations with 40 variables, created by characterizing each pattern using estimation of fourth-order modified moments in four orientations: 0, 45, 90, and 135 degrees (see Guérin-Dugué and Avilez-Cruz 1993 for details). There are 11 classes of types of texture: grass lawn, pressed calf leather, handmade paper, raffia looped to a high pile, cotton canvas, pigskin, beach sand, another type of beach sand, oriental straw cloth, another type of oriental straw cloth, and oriental grass fiber cloth (labeled groups 1–11). We have 500 observations in each group.

When we clustered on all available variables, we found that the model with the highest BIC value was the one-cluster model (with an error rate of 90.9%). When we used the greedy search procedure with a maximum number of 15 clusters, allowing only the unconstrained VVV model because the search space was already so large, we selected 35 variables (all but variables 1, 11, 15, 31, and 40), which, when clustered allowing all models, chose (via BIC) the 11-cluster VVV model. The classification results are given in Table 8.

The classification matrix for the model based on the selected variables is given in Table 9. The classification from this model is much closer to the true partition than that from the model based on all of the variables, in terms of both the number of groups being correct and the group memberships. We can see that most groups except groups 8, 1, and 6 are picked out well.

5. DISCUSSION

We have proposed a method for variable or feature selection in model-based clustering. The method recasts the variable selection problem as one of model choice for the entire dataset and addresses it using approximate Bayes factors and a greedy search algorithm. For several simulated and real data examples, the method gives better estimates of the number of clusters, lower classification error rates, more parsimonious clustering models, and hence easier interpretation and visualization than clustering using all of the available variables.

Our method for searching for the best subset of variables is a greedy search algorithm, and of course this will find only a local optimum in the space of models. The method works well in our experiments, but it may be possible to improve its performance by using a different optimization algorithm, such as

Table 7. Classification Results for the Iris Data

Variable selection procedure	*Number of variables*	*Number of clusters*	*Error rate (%)*
None (all variables)	4	2	33.3
None (all variables)	4	3(c)	3.3
Greedy search	3	3	4

NOTE: The correct number of groups is three. (c) indicates that the number of clusters was constrained to this value in advance.

Table 8. Classification Results for the Texture Data

Variable selection procedure	*Number of variables*	*Number of clusters*	*Error rate (%)*
None (all variables)	40	1	90.9
None (all variables)	40	11(c)	40.7
Greedy search	35	11	13.6

NOTE: The correct number of groups is 11. (c) indicates that the number of clusters was constrained to this value in advance.

Table 9. Texture Data: Confusion Matrix for the Clustering Based on the Selected Variables

	Group 4	Group 5	Group 9	Group 10	Group 3	Group 11	Group 7	Group 2	Group 8	Group 1	Group 6
Cluster 8	**500**	0	0	0	0	0	0	0	0	0	0
Cluster 9	0	**500**	0	0	0	0	0	0	0	0	0
Cluster 10	0	0	**500**	0	0	0	0	0	0	0	0
Cluster 11	0	0	0	**500**	0	1	0	0	0	0	0
Cluster 6	0	0	0	0	**499**	0	0	0	0	0	0
Cluster 4	0	0	0	0	0	**497**	0	0	0	0	0
Cluster 2	0	0	0	0	0	0	**474**	0	0	200	0
Cluster 3	0	0	0	0	0	0	0	**439**	0	0	0
Cluster 5	0	0	0	0	0	0	0	0	**383**	0	341
Cluster 1	0	0	0	0	1	0	0	61	0	**300**	0
Cluster 7	0	0	0	0	0	2	26	0	117	0	**159**

NOTE: The largest count in each row is boxed.

Markov chain Monte Carlo or simulated annealing. Our method is analogous to stepwise regression, which has been found to be often unstable, as noted by Miller (1990), for example. This was not a problem for the analyses conducted in this article, but it remains an issue to be aware of. Also, when the number of variables is vast (e.g., in microarray data analysis when thousands of genes may be the variables being used), the method is too slow to be practical as it stands. Combining our basic approach with prescreening (where subsets of variables are selected prior to using the variable selection clustering procedure) and alternative model search methods, such as Badsberg's (1992) headlong procedure could yield a method that would be feasible for such cases.

The method is feasible for quite large datasets. For example, when the method was run on a simulated dataset with two clusters, 10,000 observations, and 10 variables (of which 8 were clustering variables), using hierarchical clustering on a subset of 1,000 observations, a maximum allowed number of 9 clusters and the VVV model only, the CPU time on a laptop with 512 MB of memory and a 1.5-GHz processor was just under 11 hours.

Less work has been done on variable selection for clustering than for classification (also called discrimination or supervised learning), perhaps reflecting the fact that the former is a harder problem. In particular, variable selection and dimension reduction in the context of model-based clustering have not received much attention.

In a model-based clustering framework, work has been done on this problem by Law, Jain, and Figueiredo (2002), Vaithyanathan and Dom (1999), Liu, Zhang, Palumbo, and Lawrence (2003), Ding, He, Zha, and Simon (2002), Chakrabarti and Mehrotra (2000), Mitra, Murthy, and Pal (2002), and Talavera (2000). In the non–model-based clustering context, work on variable selection has been done by Dy and Brodley (2000), Lazzeroni and Owen (2002), Getz, Levine, and Domany (2000), McLachlan, Bean, and Peel (2002), McCallum, Nigam, and Ungar (2000), Brusco and Cradit (2001), Devaney and Ram (1997), Friedman and Meulman (2004), Gnanadesikan, Kettenring, and Tsao (1995), and Desarbo, Carroll, Clarck, and Green (1984).

Our examples have involved continuous data modeled by mixtures of normal distributions. But the same basic ideas can be applied to variable selection in other clustering contexts, including clustering multivariate discrete data using latent class models (Clogg and Goodman 1984; Becker and Yang 1998) and, more generally, Bayesian graphical models with a hidden categorical node (Chickering and Heckerman 1997). When the present approach is adapted to these other clustering problems, it should retain the aspects that make it flexible, especially its ability to simultaneously estimate the number of clusters and group structure, as well as to select the clustering variables.

APPENDIX: VARIABLE SELECTION AND CLUSTERING ALGORITHM

Here we give a more complete description of the variable selection and clustering algorithm for the case of continuous data modeled by multivariate normal groups. This version allows for choosing the number of clusters and also the model parameterizations, if required; otherwise, one can simply alter the following steps slightly to choose only the number of clusters:

- Choose $G_{\max}$, the maximum number of clusters to be considered for the data.
- *First step*. The first clustering variable is chosen to be the one that gives the greatest difference between the BIC for clustering on it (maximized over the number of clusters from two up to $G_{\max}$ and different parameterizations) and the BIC for no clustering (a single group structure maximized over different parameterizations) on it, where each variable is considered separately. We do not require that the greatest difference be positive at this point, because in certain cases there is no evidence of univariate clustering in data where multivariate clustering may be present, and to find this clustering, we need a starting variable.

  Specifically, we split $Y^{(3)} = Y$ into its variables, $y^1, \ldots, y^{D_1}$. For all j in $1, \ldots, D_1$ we compute the approximation to the Bayes factor in (6) by

$$BIC_{\text{diff}}(y^j) = BIC_{\text{clust}}(y^j) - BIC_{\text{not clust}}(y^j),$$

  where $BIC_{\text{clust}}(y^j) = \max_{2\le G\le G_{\max}, m\in\{E,V\}}\{BIC_{G,m}(y^j)\}$, with $BIC_{G,m}(y^j)$ the BIC given in (1) for the model-based clustering model for y^j with G clusters and model m either the one-dimensional equal variance (E) or the unequal variance model (V), and $BIC_{\text{not clust}}(y^j) = BIC_{\text{reg}}$ as given in (A.1) (for a regression model with constant mean) with $\dim(Y^{(1)}) = 0$.

  We choose the best variable, y^{j_1}, such that

$$j_1 = \arg\max_{j:y^j\in Y}(BIC_{\text{diff}}(y^j))$$

and create

$$Y^{(1)} = (y^{j_1})$$

and

$$Y^{(3)} = Y \setminus y^{j_1},$$

where $Y \setminus y^{j_1}$ denotes the set of variables in Y excluding variable y^{j_1}.

- *Second step.* Next, the set of clustering variables is chosen to be the pair of variables, including the variable selected in the first step, that gives the greatest difference between the BIC for clustering on both variables (maximized over the number of clusters from two up to $G_{\max}$ and different parameterizations) and the sum of the BIC for the univariate clustering of the variable chosen in the first step and the BIC for the linear regression of the new variable on the variable chosen in the first step. Note that we do not assume that the greatest difference is positive, because the only criterion that the variables need to satisfy is being the best initialization variables.

  Specifically, we split $Y^{(3)}$ into its variables $y^1, \ldots, y^{D_2}$. For all j in $1, \ldots, D_2$ we compute the approximation to the Bayes factor in (6) by

$$BIC_{\text{diff}}(y^j) = BIC_{\text{clust}}(y^j) - BIC_{\text{not clust}}(y^j),$$

  where $BIC_{\text{clust}}(y^j) = \max_{2 \le G \le G_{\max}, m \in M}\{BIC_{G,m}(Y^{(1)}, y^j)\}$, with $BIC_{G,m}(Y^{(1)}, y^j)$ the BIC given in (1) for the model-based clustering model for the dataset including both the previously selected variable [contained in $Y^{(1)}$] and the new variable y^j with G clusters and model m in the set of all possible models M, and $BIC_{\text{not clust}}(y^j) = BIC_{\text{reg}} + BIC_{\text{clust}}(Y^{(1)})$, where BIC_{reg} is given in (7) [the regression model with independent variable $Y^{(1)}$ and dependent variable y^j] when $\dim(Y^{(1)}) = 1$ (the number of variables currently selected) and $BIC_{\text{clust}}(Y^{(1)})$ is the BIC for the clustering with only the currently selected variable in $Y^{(1)}$.

  We choose the best variable, y^{j_2}, with

$$j_2 = \arg \max_{j: y^j \in Y^{(3)}} (BIC_{\text{diff}}(y^j))$$

  and create

$$Y^{(1)} = Y^{(1)} \cup y^{j_2}$$

  and

$$Y^{(3)} = Y^{(3)} \setminus y^{j_2},$$

  where $Y^{(1)} \cup y^{j_2}$ denotes the set of variables including those in $Y^{(1)}$ and variable y^{j_2}.
- *General step* (inclusion part). The proposed new clustering variable is chosen to be the variable that gives the greatest difference between the BIC for clustering with this variable included in the set of currently selected clustering variables (maximized over numbers of clusters from two up to $G_{\max}$ and different parameterizations) and the sum of the BIC for the clustering with only the currently selected clustering variables and the BIC for the linear regression of the new variable on the currently selected clustering variables.
- If this difference is positive, then the proposed variable is added to the set of selected clustering variables. If it is not, then the set remains the same. Specifically, at step t we split $Y^{(3)}$ into its variables $y^1, \ldots, y^{D_t}$. For all j in $1, \ldots, D_t$, we compute the approximation to the Bayes factor in (6) by

$$BIC_{\text{diff}}(y^j) = BIC_{\text{clust}}(y^j) - BIC_{\text{not clust}}(y^j), \qquad \text{(A.1)}$$

  where $BIC_{\text{clust}}(y^j) = \max_{2 \le G \le G_{\max}, m \in M}\{BIC_{G,m}(Y^{(1)}, y^j)\}$, with $BIC_{G,m}(Y^{(1)}, y^j)$ being the BIC given in (1) for the model-based clustering model for the dataset including both the previously selected variables [contained in $Y^{(1)}$] and the new variable y^j with G clusters and model m in the set of all possible models M, and $BIC_{\text{not clust}}(y^j) = BIC_{\text{reg}} + BIC_{\text{clust}}(Y^{(1)})$, where BIC_{reg} is given in (7) [the regression model with independent variables $Y^{(1)}$ and dependent variable y^j] when $\dim(Y^{(1)}) =$ (the number of variables currently selected) and $BIC_{\text{clust}}(Y^{(1)})$ is the BIC for the clustering with only the currently selected variables in $Y^{(1)}$.

  We choose the best variable, y^{j_t}, with

$$j_t = \arg \max_{j: y^j \in Y^{(3)}} (BIC_{\text{diff}}(y^j))$$

  and create

$$Y^{(1)} = Y^{(1)} \cup y^{j_t} \quad \text{if } BIC_{\text{diff}}(y^{j_t}) > 0$$

  and

$$Y^{(3)} = Y^{(3)} \setminus y^{j_t} \quad \text{if } BIC_{\text{diff}}(y^{j_t}) > 0;$$

  otherwise, $Y^{(1)} = Y^{(1)}$ and $Y^{(3)} = Y^{(3)}$.
- *General step* (removal part). The proposed variable for removal from the set of currently selected clustering variables is chosen to be the variable from this set that gives the smallest difference between the BIC for clustering with all currently selected clustering variables (maximized over number of clusters greater than two up to $G_{\max}$ and different parameterizations) and the sum of the BIC for clustering with all currently selected clustering variables except for the proposed variable and the BIC for the linear regression of the proposed variable on the other clustering variables.
- If this difference is negative, then the proposed variable is removed from the set of selected clustering variables. If it is not, then the set remains the same.

  In terms of equations for step $t+1$, we split $Y^{(1)}$ into its variables $y^1, \ldots, y^{D_{t+1}}$. For all j in $1, \ldots, D_{t+1}$, we compute the approximation to the Bayes factor in (6) by

$$BIC_{\text{diff}}(y^j) = BIC_{\text{clust}} - BIC_{\text{not clust}}(y^j),$$

  where $BIC_{\text{clust}} = \max_{2 \le G \le G_{\max}, m \in M}\{BIC_{G,m}(Y^{(1)})\}$, with $BIC_{G,m}(Y^{(1)})$ the BIC given in (1) for the model-based clustering model for the dataset including the previously selected variables [contained in $Y^{(1)}$] with G clusters and model m in the set of all possible models M, and $BIC_{\text{not clust}}(y^j) = BIC_{\text{reg}} + BIC_{\text{clust}}(Y^{(1)} \setminus y^j)$, where BIC_{reg} is given in (A.1) [the regression model with independent variables being all of $Y^{(1)}$ except y^j and dependent variable y^j] when $\dim(Y^{(1)}) =$ (the number of variables currently selected) $- 1$ and $BIC_{\text{clust}}(Y^{(1)} \setminus y^j)$ is the BIC for the clustering with all of the currently selected variables in $Y^{(1)}$ except for y^j.

  We choose the best variable, $y^{j_{t+1}}$, with

$$j_{t+1} = \arg \min_{j: y^j \in Y^{(1)}} (BIC_{\text{diff}}(y^j)),$$

  and create

$$Y^{(1)} = Y^{(1)} \setminus y^{j_{t+1}} \quad \text{if } BIC_{\text{diff}}(y^{j_{t+1}}) \le 0$$

  and

$$Y^{(3)} = Y^{(3)} \cup y^{j_{t+1}} \quad \text{if } BIC_{\text{diff}}(y^{j_{t+1}}) \le 0;$$

  otherwise, $Y^{(1)} = Y^{(1)}$ and $Y^{(3)} = Y^{(3)}$.
- After the first and second steps, the general step is iterated until consecutive inclusion and removal proposals are rejected. At this point the algorithm stops, because any further proposals will be the same ones already rejected.

[Received June 2004. Revised April 2005.]

REFERENCES

Anderson, E. (1935), "The Irises of the Gaspé Peninsula," *Bulletin of the American Iris Society*, 59, 2–5.

——— (1936), "The Species Problem in *Iris*," *Annals of the Missouri Botanical Garden*, 23, 457–509.

Badsberg, J. H. (1992), "Model Search in Contingency Tables by CoCo," in *Computational Statistics*, Vol. 1, eds. Y. Dodge and J. Whittaker, Heidelberg: Physica-Verlag, pp. 251–256.

Banfield, J. D., and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 48, 803–821.

Becker, M. P., and Yang, I. (1998), "Latent Class Marginal Models for Cross-Classifications of Counts," *Sociological Methodology*, 28, 293–326.

Brodatz, P. (1966), *Textures: A Photographic Album for Artists and Designers*, New York: Dover.

Brusco, M. J., and Cradit, J. D. (2001), "A Variable Selection Heuristic for *k*-Means Clustering," *Psychometrika*, 66, 249–270.

Campbell, N. A., and Mahon, R. J. (1974), "A Multivariate Study of Variation in Two Species of Rock Crab of Genus *Leptograpsus*," *Australian Journal of Zoology*, 22, 417–425.

Celeux, G., and Govaert, G. (1995), "Gaussian Parsimonious Clustering Models," *Pattern Recognition*, 28, 781–793.

Chakrabarti, K., and Mehrotra, S. (2000), "Local Dimensionality Reduction: A New Approach to Indexing High-Dimensional Spaces," *The VLDB Journal*, 89–100.

Chang, W. C. (1983), "On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions," *Applied Statistics*, 32, 267–275.

Chickering, D. M., and Heckerman, D. (1997), "Efficient Approximations for the Marginal Likelihood of Bayesian Networks With Hidden Variables," *Machine Learning*, 29, 181–212.

Clogg, C. C., and Goodman, L. A. (1984), Latent Structure Analysis of a Set of Multidimensional Contingency Tables," *Journal of the American Statistical Association*, 79, 762–771.

Desarbo, W. S., Carroll, J. D., Clarck, L. A., and Green, P. E. (1984), "Synthesized Clustering: A Method for Amalgamating Clustering Bases With Differential Weighting of Variables," *Psychometrika*, 49, 57–78.

Devaney, M., and Ram, A. (1997), "Efficient Feature Selection in Conceptual Clustering," in *Machine Learning: Proceedings of the Fourteenth International Conference*, Nashville, TN, pp. 92–97.

Ding, C., He, X., Zha, H., and Simon, H. D. (2002), "Adaptive Dimension Reduction for Clustering High-Dimensional Data," in *Proceedings of the IEEE International Conference on Data Mining*, Maebashi, Japan, pp. 147–154.

Dy, J. G., and Brodley, C. E. (2000), "Feature Subset Selection and Order Identification for Unsupervised Learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, San Francisco, CA, pp. 247–254.

Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188.

Fraley, C., and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631.

——— (2003), "Enhanced Software for Model-Based Clustering," *Journal of Classification*, 20, 263–286.

Friedman, J. H., and Meulman, J. J. (2004), "Clustering Objects on Subsets of Attributes" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 66, 1–25.

Getz, G., Levine, E., and Domany, E. (2000), "Coupled Two-Way Clustering Analysis of Gene Microarray Data," *Proceedings of the National Academy of Sciences USA*, 97, 12079–12084.

Gnanadesikan, R., Kettenring, J. R., and Tsao, S. L. (1995), "Weighting and Selection of Variables for Cluster Analysis," *Journal of Classification*, 12, 113–136.

Green, P. E., and Krieger, A. M. (1995), "Alternative Approaches to Cluster-Based Market Segmentation," *Journal of the Market Research Society*, 37, 221–239.

Guérin-Dugué, A., and Avilez-Cruz, C. (1993), "High-Order Statistics From Natural Textured Images," in *ATHOS Workshop on System Identification and High-Order Statistics*, Sophia-Antipolis, France.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Keribin, C. (1998), "Consistent Estimate of the Order of Mixture Models," *Comptes Rendues de l'Academie des Sciences, Série I-Mathématiques*, 326, 243–248.

Law, M. H., Jain, A. K., and Figueiredo, M. A. T. (2002), "Feature Selection in Mixture-Based Clustering," in *Proceedings of Conference of Neural Information Processing Systems*, Vancouver.

Lazzeroni, L., and Owen, A. (2002), "Plaid Models for Gene Expression Data," *Statistica Sinica*, 12, 61–86.

Liu, J. S., Zhang, J. L., Palumbo, M. J., and Lawrence, C. E. (2003), "Bayesian clustering With Variable and Transformation Selections," in *Bayesian Statistics*, Vol. 7, eds. J. M. Bernardo, M. J. Bayarri, A. P. Dawid, J. O. Berger, D. Heckerman, A. F. M. Smith, and M. West, Oxford, U.K.: Oxford University Press, pp. 249–275.

McCallum, A., Nigam, K., and Ungar, L. (2000), "Efficient Clustering of High-Dimensional Data Sets With Application to Reference Matching," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 169–178.

McLachlan, G. J., Bean, R., and Peel, D. (2002), "A Mixture Model-Based Approach to the Clustering of Microarray Expression Data," *Bioinformatics*, 18, 413–422.

McLachlan, G. J., and Peel, D. (1998), "Robust Cluster Analysis via Mixtures of Multivariate *t*-Distributions," in *Lecture Notes in Computer Science*, Vol. 1451, eds. P. P. A. Amin, D. Dori, and H. Freeman, Berlin: Springer-Verlag, pp. 658–666.

——— (2000), *Finite Mixture Models*, New York: Wiley.

Miller, A. J. (1990), *Subset Selection in Regression*, London: Chapman & Hall.

Mitra, P., Murthy, C. A., and Pal, S. K. (2002), "Unsupervised Feature Selection Using Feature Similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 301–312.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge, U.K.: Cambridge University Press.

Talavera, L. (2000), "Dependency-Based Feature Selection for Clustering Symbolic Data," *Intelligent Data Analysis*, 4, 19–28.

Vaithyanathan, S., and Dom, B. (1999), "Generalized Model Selection for Unsupervised Learning in High Dimensions," in *Proceedings of Neural Information Processing Systems*, eds. S. A. Solla, T. K. Leen, and K. R. Muller, Cambridge, MA: MIT Press, pp. 970–976.

Wolfe, J. H. (1963), "Object Cluster Analysis of Social Areas," master's thesis, University of California, Berkeley.

Yeung, K. Y., and Ruzzo, W. L. (2001), "Principal Component Analysis for Clustering Gene Expression Data," *Bioinformatics*, 17, 763–774.