Bayesian Variable Selection and Cluster Analysis

Vasiliki Dimitrakopoulou
Supervisor: Prof. Phil Brown

Institute of Mathematics, Statistics and Actuarial Science
University of Kent

Greek Stochastics Meeting α' :
Monte Carlo: Probability and Methods

August 28-31, 2009

Table of contents

Introduction

Technological Advances $\longrightarrow$ large amount of data with the number of variables being substantially larger than the number of observations: $n \prec\prec p$.

Our Aim:

- Uncover the group structure of the observations.

  But dealing with high-dimensional datasets $\longrightarrow$ The cluster structure is often confined to a small subset of variables.

  And the inclusion of unnecessary covariates could complicate or mask the cluster structure *(Fowlkes, Gnanadesikan and Kettering 1998; Milligan 1989; Gnanadesikan, Kettering and Tao 1995; Brusco and Cradit 2001)*.

- Identify the discriminating variables.

The clustering task

Reduce the amount of data by grouping similar data items together.

Some methods:

- Hierarchical clustering.
- K-means.
- Model-based clustering.

The identification task

Identify the variables that define the true cluster structure.

differentially weighting the covariates

selecting the discriminating ones

Variable selection procedures:

- outperform the variable weighting schemes *(Gnanadesikan et al. (1995))*
- improve the prediction of cluster membership
- reduce the measurement and storage requirements for future samples
- provide more cost-effective predictors

The main idea of our approach

Simultaneously select the discriminating variables and uncover the structure of the data/cluster the samples into G groups.

Clustering $\longrightarrow$ formulation of a multivariate mixture model with G components (Model-based clustering).

Variable selection $\longrightarrow$ use of latent indicators to identify the discriminating variables.

Model-based Clustering

n: number of observations
p: number of variables
G: number of groups
n$\prec\prec$ p

Data are viewed as coming from a mixture of distributions, each of which represents a different cluster.

$X = (x_1, x_2, \ldots, x_n)$: n independent p-dimensional observations from G populations.

Model-based Clustering

Cluster the n samples using a mixture of G probability functions:

$$f(x_i|w,\theta) = \sum_{k=1}^{G} w_k f(x_i|\theta_k)$$

- $f(x_i|\theta_k)$: density of an observation x_i from the k^{th} component.
- $w = (w_1, \ldots, w_G)^T$: component weights with $w_k \geq 0$ and $\sum_{k=1}^{G} w_k = 1$
- $y = (y_1, \ldots, y_n)^T$: latent variables to identify the cluster from which each observation is drawn.
- $y_i = k$: the i^{th} observation comes from the k^{th} component.
- y_i's are i.i.d with probability mass function: $p(y_i = k) = w_k$.

Variable Selection

Introduction of a γ latent p-vector with binary entries.

$$\gamma_j = \begin{cases} 1, & \text{the } j^{th} \text{ variable defines a mixture distribution for the data} \\ 0, & \text{the } j^{th} \text{ variable favors a single multivariate normal density} \end{cases}$$

γ : index for the discriminating variables.

γ^c: index for the non-discriminating variables.

$p_\gamma = \sum_{j=1}^{p} \gamma_j$: number of discriminating variables.

Model

The case : $x_i|y_i = k, w, \theta \sim N(\mu_k, \Sigma_k)$

We keep the number of clusters fixed.

We have our likelihood $L(G, \gamma, w, \mu, \Sigma|X, y)$ proportional to :

- $\eta_{(\gamma^c)}$: the mean for the non discriminating variables,
- $\Omega_{(\gamma^c)}$: the covariance matrix for the non discriminating variables,
- $\mu_{k(\gamma)}$: the mean of cluster k for the discriminating variables,
- $\Sigma_{k(\gamma)}$: the covariance matrix for the discriminating variables.

The algorithm is much more efficient if we integrate out the parameters $\mu, \Sigma, \eta, \Omega$.

Prior Assumptions

The integration can be facilitated by taking the conjugate priors.

- $\mu_{k(\gamma)}|\Sigma_{(\gamma)}, G \sim N(\mu_{0(\gamma)}, h_1\Sigma_{(\gamma)})$
- $\eta_{(\gamma^c)}|\Omega_{(\gamma^c)} \sim N(\mu_{0(\gamma^c)}, h_0\Omega_{(\gamma^c)})$
- $\Sigma_{(\gamma)}|G \sim IW(\delta; Q_{1(\gamma)})$
- $\Omega_{(\gamma^c)} \sim IW(\delta; Q_{0(\gamma^c)})$

μ_0: p-vector of the midpoints of the variables.

h_0, h_1: arbitrarily large, between 10-1000.

δ: shape parameter, small $\rightarrow$ weak prior.

$Q_1 = (1/k_1)I_{(p\times p)}$.

$Q_0 = (1/k_0)I_{(p\times p)}$.

k_1: 1 - 10 % of the upper decile of the n-1 non-zero eigenvalues.

k_0: 1 - 10 % of the lower decile of the n-1 non-zero eigenvalues.

- w: symmetric Dirichlet: $w|G \sim$ Dirichlet(a,...,a)

Full Conditionals

Since we have integrated out the μ, Σ,η, Ω parameters we need to sample only from the joint posterior of (G,y,γ,w):

$$f(X, y|G, w, \gamma)$$

The full conditionals are:

- $f(y|G, w, \gamma, X) \propto f(X, y|G, w, \gamma)$
- $w|G, \gamma, y, X \sim \text{Dirichlet}(\alpha + n_1, \ldots, \alpha + n_G)$

Updating Steps

- Update w using a Gibbs sampler.
- Update y one element at a time using a sub-Gibbs strategy.
- Update γ using simulated annealing.

Update γ

We generate a new candidate γ^{new} by randomly choosing one of the following three transition moves with probability 1/3:

1. **Add**: Randomly choose a 0 in γ^{old} and change it to a 1.
2. **Delete**: Randomly choose a 1 in γ^{old} and change it to a 0.
3. **Swap**: Choose independently and at random a 0 and a 1 in γ^{old}, and switch their values.

Update γ

We have a cost function: $C(\gamma) = loss\ function + c_\gamma$

At each step we calculate: $d = C(\gamma^{new}) - C(\gamma^{old})$

- If d$\prec$ 0 $\rightarrow$ γ^{new} is always accepted
- Otherwise we accept γ^{new} with probability $exp(-d/T)$

T is the control parameter temperature.

Idea: Start with a high T so that all steps are accepted.

Gradually reduce T so that eventually only steps that improve the algorithm are accepted.

Stop the algorithm after M steps of no accepted moves, for large M.

Application

- $n = 15$ observations.
- $G = 3$ groups.
- $p = 100$ variables.
- $p_\gamma = 10$ discriminating variables.

We applied

- Simulated annealing.
- A simplified version of the method discribed in the paper "Bayesian Variable Selection in Clustering High-Dimensional Data" by Tadesse, M.G., Sha, N., and Vannucci, M.

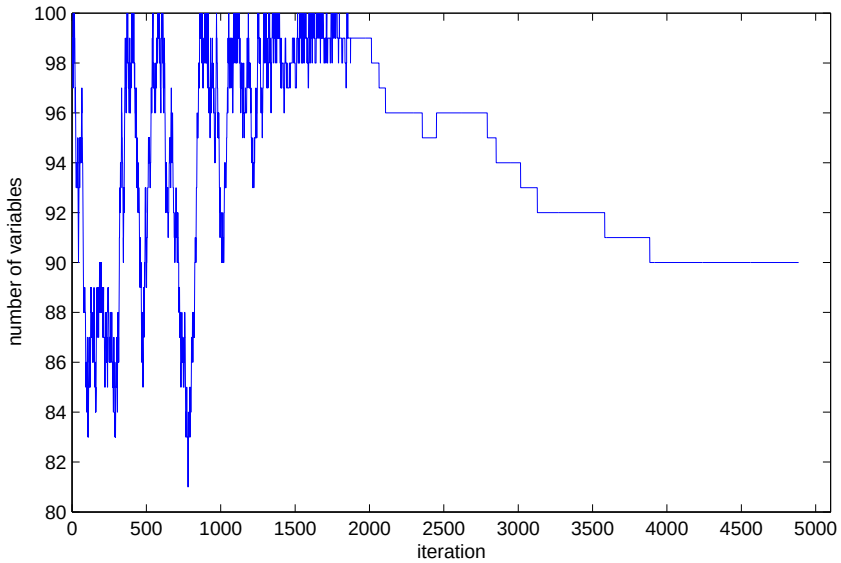
keeping the number of clusters fixed.

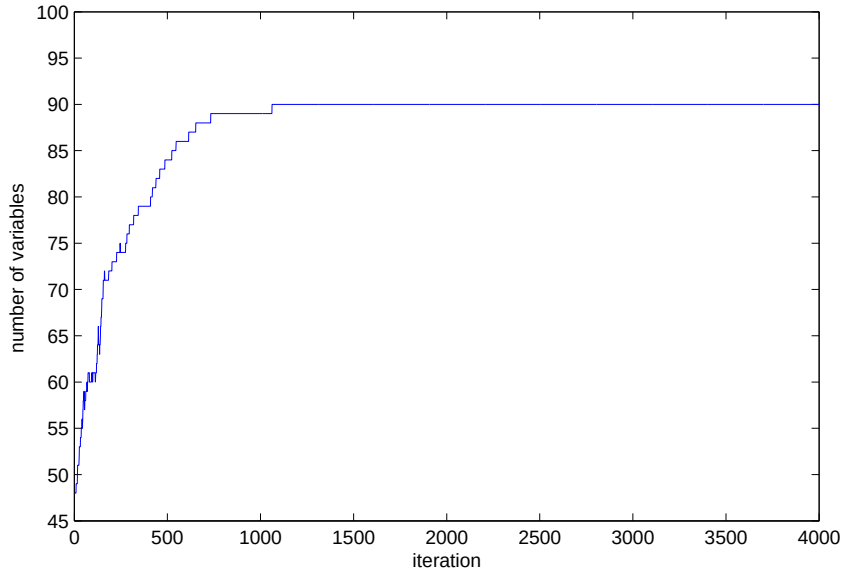
Application

We tried different values for the hyperparameters.

After a few runs on each case we noticed that both methods have the same behaviour with an "interesting" feature.

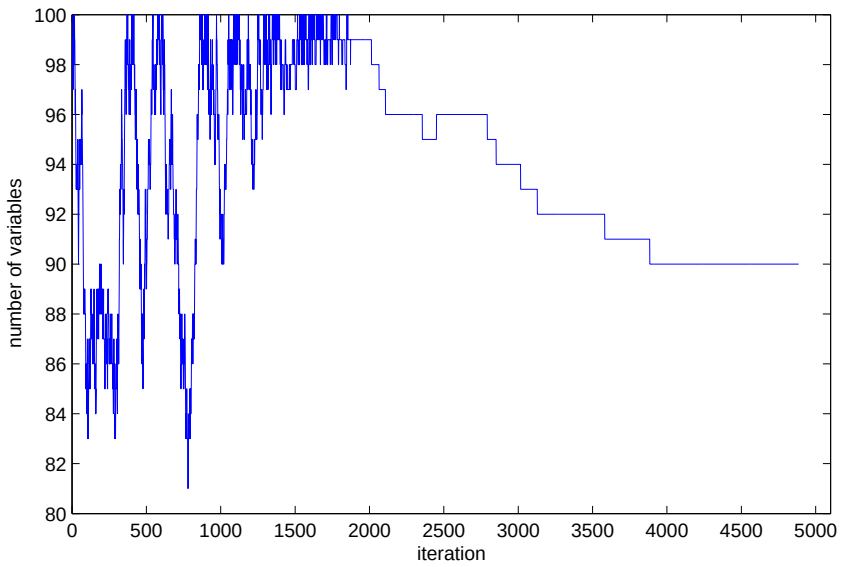
Under specific values of the hyperparameters both methods converge to the subset of the 10 variables.

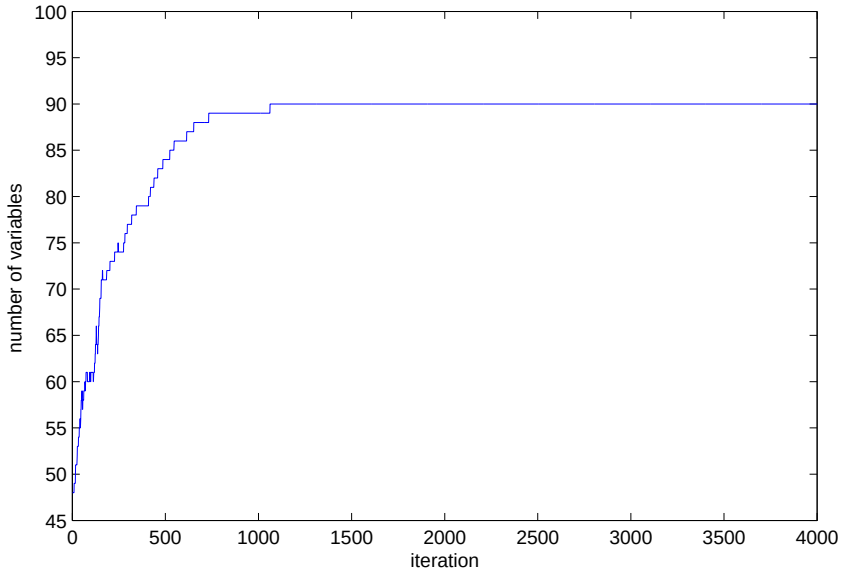




Application

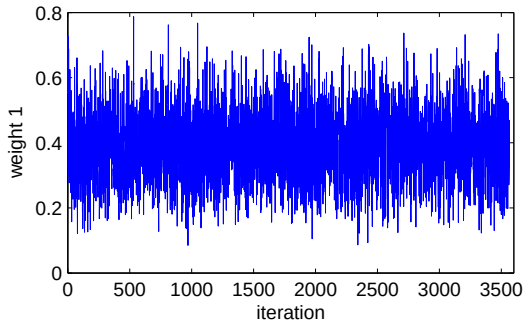
However under other hyperparameterisation, the methods choose the complement subset of variables: 90 variables are chosen as the discriminating ones.

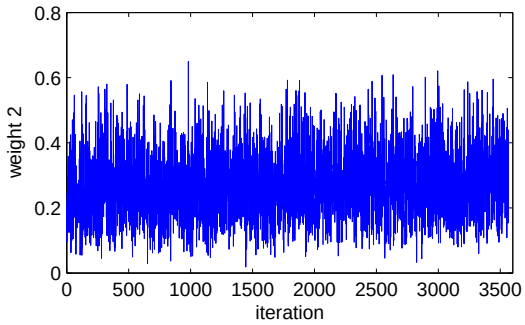


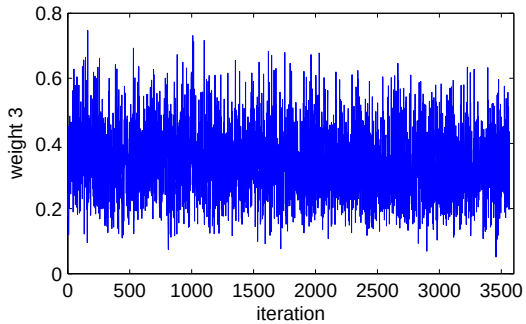


Application

The clustering works properly for both methods uncovering the true group structure of the data.







References

- Brusco, M.J., and Cradit, J.D.(2001), "A Variable Selection Heuristic for k-Means Clustering", *Psychometrika*, 66, 249-270.
- Fowlkes, E.B., Ganadesikan, R.m and Kettering, J.R. (1998), "Variable Selection in Clustering", *Journal of Classification*, 5, 205-228.
- Friedman, J.H., and Meulman, J.J. (2003), "Clustering Objects on Subsets of Attributes", technical report, Stanford University, Dept of Statistics and Stanford Linear Accelerator Center.
- Gnanadesikan, R., Kettering, J.R., and Tao, S.L (1995), "Weighting and Selection of Variables for Cluster Analysis", *Journal of Classification*, 12, 113-136.
- Gosh, D., and Chinnaiyan, A.M. (2002), "Mixture Modelling of Gene Expression Data From Microarray Experiments", *Bioinformatics*, 18, 275-286.
- McLachlan, G.J., Bean, R.W., and Peel, D. (2002), "A mixture Model-Based Approach to the Clustering of Microarray Expression Data", *Bioinformatics*, 18, 413-422.
- Raftery, A.E., and Dean, N. (2006), "Variable Selection for Model-Based Clustering", *Journal of the American Statistical Association*, 66, 101, 168-178
- Tadesse, M.G., Sha, N., and Vannucci, M. (2005), "Bayesian Variable Selection in Clustering High-Dimensional Data", *Journal of the American Statistical Association*, Vol.100, No.470.