MCMC Clustering and Its Convergence Assessment

JSM 2013

Namdar Homayounfar, Vahid Partovi Nia, Masoud Asgharian

McGill

Introduction

Clustering can be described as the partitioning of data into homogeneous groups. The modern clustering approaches such as the EM algorithm or the k-means are sensitive to initial values. In order to make the clustering algorithm insensitive to starting values, we consider the data groupings as an unobserved random variable and sample from its closed form posterior distribution using an MCMC method such as the Gibbs sampler.

The Markov Chain samples are used to estimate the *maximum a posteriori* (MAP) grouping after the chain has converged.

The space of groupings is a nominal finite state space. We propose a quantitative convergence criterion for MCMC algorithms run on nominal state spaces. More precisely, we define a one-dimensional statistic of fit and present its distribution. This statistic is used to assess the convergence of a Markov Chain via a formal statistical significance test.

We apply this clustering methodology to the genetic mutants of the flowering plant *Arabidopsis thaliana*.

Bayesian Model for Clustering

In Bayesian clustering, the labeling of the observations is regarded as a random variable and has a probability distribution. Therefore a Bayesian model with a likelihood function and prior distribution is adopted for the clusters. Let

$c \in \{1,\dots,C\}$ $\rightarrow$ cluster label

T_c $\rightarrow$ number of observations in cluster c and $T=\sum_{c=1}^{C} T_c$ the total number of observations

$v \in \{1,\dots,V\} \rightarrow$ subscript of independent continuous variables

y_{vct} $\rightarrow$ data of clustering individual $t \in \{1,\dots,T_c\}$ in cluster $c \in \{1,\dots,C\}$ from variable $v \in \{1,\dots,V\}$

$\vec{c}=\{c_t\}_{t=1}^{T} \rightarrow$ is a grouping such that $c_t = c \in \{1,\dots,C\}$ if the ith observation is allocated to cluster c

In order to impose uniqueness in cluster labeling, we assume that grouping parameters in $\vec{c}$ appear in an increasing order.

The state space of interest is that of all possible allocations under the marginal posterior distribution $\pi(\vec{c}|\vec{y}) \propto \pi(\vec{y}|\vec{c})\pi(\vec{c})$ where $\pi(\vec{c})$ is the prior distribution on the allocations and

$$\pi(\vec{y}|\vec{c}) = \int \Big\{ \prod_{c=1}^{C} \prod_{\{t;c_t=c\}} \pi(y_t|\theta,\vec{c}) \Big\} \pi(\theta|\vec{c}) d\theta$$

We assume that, conditional on $\vec{c}$ and the model parameters, the observations are independent within and across clusters. This assumption allows us to derive a closed form for $\pi(\vec{y}|\vec{c})$. As a result, we can further compute $\pi(\vec{c}|\vec{y})$ up to a constant and construct a Markov Chain of groupings using a reversible Gibbs sampler. We estimate the MAP grouping by considering the most frequent labeling in the sample.

Data Analysis

Messerli et al. (2007) study the metabolic pattern of 14 genetic mutants of the plant Arabidopsis thaliana from measurements of 43 metabolites (mostly sugars, sugar alcohols, amino acids and organic acids), obtained by the method of gas chromatography mass spectrometry. Figure 1 presents the data, where mutants are represented by integer labels, and four replicates are are available for each mutant; exceptionally, for mutant 1 only 3 replicates exists.

In our analysis, for each mutant type we considered the mean of the replicates. Also, we centered all the metabolite variables around the median. The goal is to a perform metabolomic characterization of these mutants via clustering.

We fit the following hierarchical Bayesian model. Given the data allocation vector $\vec{c}$

$$y_{vct} \sim N(\theta_{vc}+\epsilon_{vct}, \sigma_\theta^2+\sigma_\epsilon^2), \qquad \theta_{vc} \sim N(0,\sigma_\theta^2), \qquad \epsilon_{vct} \sim N(0,\sigma_\epsilon^2)$$

The subscripts v, c and t denote respectively variable, cluster and mutant in cluster. θ_{vc} represents the cluster mean and ϵ_{vct} denotes the measurement and experimental errors between mutants. We assume the uniform multinomial-Dirichlet distribution as the prior:

$$\pi(\vec{c}) \propto \frac{(C-1)!T_1!\dots T_c!}{T(T+C-1)!}$$

The MAP estimation using the Gibbs sampler after 50000 iterations yields the grouping {1,8,11,12,13,14}, {6,7}, and {2,3,4,5,9,10} with the estimated probability $\hat{\pi} = 0.4092$.

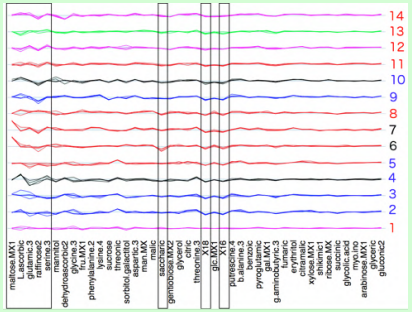

Figure 1: Plot of the log spectra (solid lines) of the matabolite data. Different colors indicate the category of the mutant: black for those defective in starch biosynthesis, red for those defective in starch degradation, green for the comparative plant, blue for the uncharacterized mutants, and magenta for the wild types. The estimated MAP grouping is represented by three different colors on the mutant labels.

Convergence Assessment

We can be more certain of the accuracy of our MAP estimation if we can assess if our Markov Chain has actually converged.

We assess convergence to equilibrium through the ratio of the empirical pmf to the true pmf known up to a normalizing constant. Averaged over states under consideration, this gives an intuitive, variance-like, one-dimensional statistic of fit. Under the hypothesis of stationarity, we expect this statistic to be small, so we propose to reject the null hypothesis for large values. In practice, we estimate the asymptotic variance by regenerative simulation.

$\{X_t\}_{t\geq 1}$ $\rightarrow$ irreducible, aperiodic Markov chain with discrete space S_M of cardinality M. The value of X_t is an integer that refers to a distinct grouping.

$\Pi = \{\Pi_i, i \in S_M\} \rightarrow$ the unique stationary distribution associated to $\{X_t\}_{t\geq 1}$ assume Π_i is known only up to a normalizing constant Z, i.e. $\Pi_i = \pi_i \backslash Z$

$X = \{X_t, t=1,\dots,n\} \rightarrow$ our finite length ergodic Markov chain

$m \leq \min(n, M)$ $\rightarrow$ the number of unique visited states by our Gibbs sampler

$S \rightarrow$ state space of X

$\hat{\pi}_i$ $\rightarrow$ proportion of X_i in the sample. Consistent estimator for Π_i

The method of regenerative simulation identifies random times at which the Markov chain probabilistically restarts itself. Let

$R \rightarrow$ the total number of regeneration tours in a chain of length n

Define the variance test statistic

$$V_n = \frac{R}{m}\sum_{i\in S}(f_i - \bar{f})^2$$

where $f_i = \hat{\pi}_i \backslash \pi_i$ and $\bar{f} = m^{-1}\sum_{j\in S} f_j$

- For large R we expect $f_i \approx Z^{-1}$ for all $i \in S$
- For large R we have the approximation

$$V_n \approx \frac{R}{m}\sum_{i\in S}(f_i - Z^{-1})^2 = \frac{R}{m}\frac{1}{Z^2}\sum_{i\in S}\frac{(O_i - E_i)^2}{E_i^2}$$

where $O_i = n\hat{\pi}_i$ and $E_i = n\Pi_i$. Similar to Pearson goodness of fit statistic

Implementation of Regenerative Sampling

In order to improve the performance of the regenerative sampling, we restrict our attention to k high mass states and merge and rename all the remaining states as the new state $k+1$

1. Set $t = n$
2. Run the Gibbs sampler for t iterations, and let i be the most frequently visited state. Split the chain into R regeneration tours defined by return visits to state i.
3. Compute the statistic V_t and the p-value p_t. If at significance level α, $p_t \leq \alpha$, reject the null hypothesis that the chain is in equilibrium. Continue for further n iterations, i.e. set $t = t+n$ and return to step 2.
   If $p_t > \alpha$, there is no evidence against the null hypothesis that the chain is in equilibrium by iteration t.

Is Our Gibbs Markov Chain in Equilibrium?

Let $k=4$, i.e. we relabel our chain $\{X_t\}_{t\geq 1}$ such that the top the top most 4 frequently visited states have lables 1 to 4 in a decreasing manner and all the other states have label 5. At 0.95 confidence level, Figure 2 below suggest that the Gibbs sampler has indeed converged

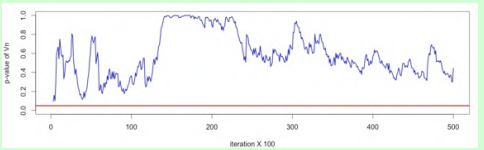

Figure 2: Plot of the p-value of V_n vs. the number of iterations y_1. A horizontal line is drawn at 0.05 as a threshold for the p-values. The samples are converged if the curve falls *above* the threshold.

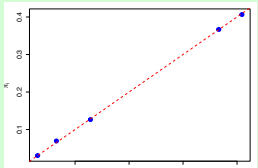

Figure 3: Plot of $\hat{\pi}_i$ vs. π_i suggests that the sample Markov chain represents the groupings well.

Namdar Homayounfar, McGill University, Montreal, QC, Canada
E-mail: namdar.homayounfar@mail.mcgill.ca
Professor Vahid Partovi Nia, Ecole Polytechnique de Montreal, QC, Canada
E-mail: vahid.partovinia@polymtl.ca
Professor Masoud Asgharian, McGill University, Montreal, QC, Canada
E-mail: masoud@math.mcgill.ca