Relabelling in Bayesian mixture models by pivotal units

Una procedura di relabelling in modelli mistura Bayesiani basata su unità pivotali

Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli

Abstract A simple procedure based on relabelling to deal with label switching when exploring complex posterior distributions by MCMC algorithms is proposed. Although it cannot be generalized to any situation, it may be handy in many applications because of its simplicity and low computational burden. A possible area where it proves to be useful is when deriving a sample for the posterior distribution arising from finite mixture models when no simple or rational ordering between the components is available.

Abstract *Si propone una strategia di 'relabelling' per il problema del 'label switching' nell'esplorazione di distribuzioni a posteriori con algoritmi di tipo MCMC. Nonostante non sia possibile generalizzare tale metodo in ogni situazione, esso si dimostra adatto in molte applicazioni in virtù della sua semplicità e della complessità computazionale relativamente bassa. In particolare, l'approccio proposto si rivela utile nel caso si voglia simulare un campione da una mistura con un numero finito di componenti.*

Key words: Mixture model, Relabelling, Label switching, MCMC.

1 Introduction

The label switching problem arises in Markov Chain Monte Carlo (MCMC) exploration of the posterior distribution of a Bayesian finite mixture model [2] because the

Roberta Pappadà, Francesco Pauli, Nicola Torelli
Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche, 'Bruno de Finetti', Università degli Studi di Trieste, Via Tigor 22, 34124 Trieste, Italy, e-mail: rpappada@units.it, e-mail: francesco.pauli@deams.units.it, e-mail: nicola.torelli@deams.units.it

Leonardo Egidi
Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Via Cesare Battisti 241, 35121 Padova, Italy, e-mail: egidi@stat.unipd.it

likelihood of the model is invariant to the relabelling of mixture components. Since there are as many maxima of the likelihood as there are permutations of the indices which denote the G elements of the mixture, this is a minor problem in the context of classical inference where any maximum, regardless of which one is chosen, leads to a valid solution and equivalent inferential conclusions. On the contrary, invariance with respect to labels is a major problem when Bayesian methods are used. In fact, if the prior distribution is invariant with respect to the labelling as well as the likelihood, then the posterior distribution is multimodal.

To make inference on a parameter specific of a component of the mixture a sample from the posterior that represent different modes would be inappropriate. An actual MCMC sample may or may not switch labels depending on the efficiency of the sampler. If the raw MCMC sampler randomly switch labels, then it is unsuitable for exploring the posterior distributions for component related parameters.

A range of solutions has been proposed, depending on the objective of the inference. See, e.g., [5] and [3], where the relabelling involves imposing identifiability constraints. However, such methods may not be applicable when an obvious constraint does not exist or when the components are not well separated.

The general problem is introduced in Sect. 2. Besides the extreme case of label invariant quantities, in Sect. 3 we illustrate a method which, starting from a clustering of the samples, performs a relabelling with the purpose of obtaining an MCMC sample suitable to infer on the characteristics of the clustering in terms of both probabilities of each unit being in each group and the group parameters. The simulation study is discussed in Sect. 4 and Sect. 5 concludes.

2 The relabelling problem

Prototypical models in which the labelling issue arises are mixture models, where, for a sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ we assume $(Y_i|Z_i = g) \sim f(y; \mu_g, \phi)$, with Z_i, $i = 1, \ldots, n$, i.i.d. random variables and $Z_i \in \{1, \ldots, G\}$, $P(Z_i = g) = \pi_g$. The likelihood of the model is

$$L(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\pi}, \phi) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f(y_i; \mu_g, \phi), \tag{1}$$

and it is invariant under a permutation of the indices of the groups.

As a consequence, the model is not identified with respect to an arbitrary permutation of the labels. When Bayesian inference for the model is performed, if the prior distribution $p_0(\boldsymbol{\mu}, \boldsymbol{\pi}, \phi)$ is invariant under a permutation of the indices, that is $p_0(\boldsymbol{\mu}, \boldsymbol{\pi}, \phi) = p_0(\boldsymbol{\mu}', \boldsymbol{\pi}', \phi)$, then so is the posterior $p(\boldsymbol{\mu}, \boldsymbol{\pi}, \phi|\boldsymbol{y}) \propto p_0(\boldsymbol{\mu}, \boldsymbol{\pi}, \phi) L(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\pi}, \phi)$, which is then multimodal with (at least) $G!$ modes.

In what follows, we assume that a sample is obtained from the posterior distribution for model (1) with a labelling invariant prior. Let $\{[\theta]_h : h = 1, \ldots, H\}$ denote the MCMC sample for the parameter $\theta = (\boldsymbol{\mu}, \boldsymbol{\pi}, \phi)$ and $\{[Z]_h : h = 1, \ldots, H\}$ the

sample for the Z variable. In principle, a perfectly mixing chain should visit the points $(\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\phi})$ and $(\boldsymbol{\mu}', \boldsymbol{\pi}', \boldsymbol{\phi})$ with the same frequency. A chain with a less than perfect mixing may either concentrate on one mode of the posterior distribution or exhibit random switches.

A naive, but effective, solution to the relabelling issue is to use a sampler which is inefficient with respect to the labelling – that is, it is unlikely to switch labels – but otherwise efficient [4]. A limitation of such an approach is the fact that, in practice, it is difficult to tune a sampler so that it is inefficient enough to avoid label switches but not too inefficient.

Note that the presence of label switches is totally not relevant if the quantities we are interested in are invariant with respect to the labels, as is the case of a prediction for $\boldsymbol{y}$, or the inference for the parameter $\boldsymbol{\phi}$; moreover, a partition of the observations in $\hat{G}$ groups can also be easily obtained by employing some clustering technique.

A particularly relevant example of invariant quantity is the probability of two units being in the same group, $c_{ij} = P(Z_i = Z_j|\mathscr{D})$, $i, j = 1, \ldots, n$ whose estimate based on the sample is

$$\hat{c}_{ij} = \frac{1}{H} \sum_{h=1}^{H} \left| [Z_i]_h = [Z_j]_h \right|. \tag{2}$$

The $n \times n$ matrix C with elements $\hat{c}_{ij}$ can be seen as a similarity matrix between units and S, with elements $\hat{s}_{ij} = 1 - \hat{c}_{ij}$, as a dissimilarity matrix.

Relabelling becomes relevant when we are interested in the features of the G groups, as the difference $\mu_2 - \mu_1$ or the probability of each unit belonging to each group, $q_{ig} = P(Z_i = g|\mathscr{D})$.

In order to study the posterior distributions of component-related quantities such as μ_g, we need to define a suitable method to permute the labels at each iteration of the Markov chain. Then, the new labels are such that different labels do refer to different components of the mixture.

3 Pivotal method

Consider a partition of the observations in $\hat{G}$ groups, $\mathscr{G}_1, \ldots, \mathscr{G}_{\hat{G}}$, obtained with a suitable clustering technique. Suppose we are interested in the probabilities $P(Z_i = g)$ and in the posteriors for groups parameters μ_g. Assume that we can find $\hat{G}$ units, $i_1, \ldots, i_{\hat{G}}$, one for each group, which are (pairwise) separated with (posterior) probability one (that is, the posterior probability of any two of them being in the same group is zero). In terms of the matrix C, the $\hat{G} \times \hat{G}$ submatrix with only the row and columns corresponding to $i_1, \ldots, i_{\hat{G}}$ will be the identity matrix. Such units, called pivots in what follows, can be used to identify the groups and to relabel the chains: for each $h = 1, \ldots, H$ and $g = 1, \ldots, \hat{G}$

$$[\mu_g]_h = [\mu_{[Z_{i_g}]_h}]_h; \qquad [Z_i]_h = g \ \text{ for } i : [Z_i]_h = [Z_{i_g}]_h. \tag{3}$$

The availability of $\hat{G}$ perfectly separated units is crucial to the procedure, and it can not always be guaranteed. In particular, there exist three different circumstances under which the relabelling procedure is unsuitable: the number of actual groups in the MCMC sample is (i) higher than $\hat{G}$, (ii) lower than $\hat{G}$, (iii) equal to $\hat{G}$ but the pivots are not perfectly separated, where the actual number of groups is the number of non empty groups, denoted by G_0. It is then clear that the Markov chain does not have informations on more than G_0 groups.

Now, let $[G]_h = \#\{g : [Z_i]_h = g \text{ for some } i\}$. Consider now the set $\mathscr{H}_1 \subset \{1,\dots,H\}$ of iterations where $[G]_h > \hat{G}$; some units and groups will then have no available pivot. For these units

$$\sum_{g=1}^{\hat{G}} \hat{P}(Z_i = g) = \sum_{g=1}^{\hat{G}} \hat{q}_{ig} = \sum_{g=1}^{\hat{G}} \frac{1}{H} \sum_{i=1}^{H} |[Z_i]_h = g| < 1 \tag{4}$$

We suggest cancelling those iterations of the chains where this occur, that is, the final –partial– chain is a sample from the posterior conditional on having at most $\hat{G}$ non empty groups. Consider now the set $\mathscr{H}_2 \subset \{1,\dots,H\}$ of iterations where $[G]_h < \hat{G}$; if $h \in \mathscr{H}_2$, then $[Z_{i_k}]_h = [Z_{i_s}]_h$ for some pivots i_k, i_s. As a consequence, the pivots are not perfectly separated: $\hat{c}_{hk} > 0$. The procedure in (3) can not be performed so also in this case we proceed cancelling that part of the chain. Finally, consider the set of iterations $\mathscr{H}_3 = \{h : \exists k,s \text{ s.t. } [Z_{i_k}]_h = [Z_{i_s}]_h\}$, where (at least) two pivot are put in the same group. As above, we need to get rid of this part of the chain.

In the end, we will relabel the chain with iterations $\mathscr{H}_0 = \{1,\dots,H\} \setminus (\mathscr{H}_1 \cup \mathscr{H}_2 \cup \mathscr{H}_3)$ which can be considered a sample from the posterior distribution conditional on (i) there being exactly $\hat{G}$ non empty groups, (ii) the pivots falling into different groups. The extent to which this conditioning is restrictive is measured by its probability, whose estimated based on the (original, raw) MCMC sample is

$$\frac{1}{H} \# \mathscr{H}_0. \tag{5}$$

A relevant issue is how to identify the pivots, noting that perfectly separated pivots may not exist and that, even if they exist, we may not be able to find them. A unit for each group can be selected according to some criterion, for instance for group g containing units $\mathscr{G}_g$ we choose $\bar{i} \in \mathscr{G}_g$ that maximizes one of the quantities

$$\max_{j \in \mathscr{G}_g} c_{\bar{i}j}, \quad \sum_{j \in \mathscr{G}_g} c_{\bar{i}j}, \quad \sum_{j \in \mathscr{G}_g} c_{\bar{i}j} - \sum_{j \notin \mathscr{G}_g} c_{\bar{i}j}. \tag{6}$$

Notice that other criteria can be considered as well, for instance one can choose $\bar{i} \in \mathscr{G}_g$ which minimizes $\sum_{j \notin \mathscr{G}_g} c_{\bar{i}j}$. The quality of the choice is then measured by the probability of the conditioning event (5).

The idea of solving the relabelling issue by fixing the group for some units dates back to [1]. More recently, [6] proposed an approach similar to the one we present, where each iteration is relabelled by minimizing some distance from a reference

labelling. Such a method avoids the need to condition on the pivots being separated, but it can be computationally demanding because of the required minimizations.

4 Some preliminary results from a simulation exercise

A simulation exercise is performed in order to allow a preliminary evaluation of the general behaviour of the pivotal method explained in Sect. 3. In the simplest scenario we could simulate data which exhibit a natural groups separation and thus the pivots detection is easy to achieve. For instance, when considering data concerning the urban traffic between cities of a specific region, the pivots are likely to coincide with the main urban centrers. Then a more complex framework is considered in order to induce the label switching problem and to make the pivots choice and the relabelling issue not trivial. The simulation scheme consists in the following steps.

(i) Simulate N values from a mixture of bivariate normal distributions with G components $\sum_{j=1}^{G} p_j N_2(\mu_j, \Sigma)$, where $p_j,\ j = 1, ..., G$ are the weights assigned to the mixture components.
(ii) Set up a Bayesian model in `JAGS` with non-informative and weakly informative prior distributions and run the MCMC algorithm with M iterations.
(iii) If necessary, discard some of the iterations of the chains according to the criterion discussed in Sect. 3.
(iv) Perform a post-process relabelling of the chains after the MCMC sampling.
(v) Apply a suitable clustering technique to the similarity matrix C of the MCMC sample in order to detect the pivots.

Different clustering algorithms have been considered, including hierarchical and non-hierarchical techniques. Then, the pivots are identified according to one of the criteria suggested in Eq. (6). An alternative approach consists in performing a cluster analysis via non-parametric density estimation and then select those units whose estimated density is maximum within each cluster.

To the end of a first exploration of the performances of the proposed method, a bivariate normal distribution for the mixture components seems appropriate. To start, G equally weighted components are considered, where G is a small number (e.g. G=3, 4). The input means are allowed to vary in the bi-dimensional space, while the variance components are fixed. As a result, we found that the 'closer' the means are, the harder is to obtain perfectly separated pivots, as expected. In this situation the submatrix of C corresponding to the selected pivots is far to be the identity matrix.

Being interested in the ability of our method in detecting the pivots, we need to define a more challenging scenario in terms of both relabelling issue and pivotal choice. To do this, a mixture of mixtures is considered in step (i), meaning that conditionally of being in one of the G groups, the value of the single $\mathbf{y}_i$ is picked out from one of two possible normal distributions, one more 'peaked' (subgroup 1), and one more 'flat' (subgroup 2), with weights $\pi_{j1}, \pi_{j2},\ j = 1, ..., G$. In this

case we held the means fixed, by varying the subgroups covariance matrices and the $\pi_{js},\ j = 1,...,G,\ s = 1,2$. The simulation results show that the greater is the difference between π_{j2} and π_{j1}, the more overlapped are the groups and the harder is to satisfactorily pick out the pivots.

Preliminary results seem to confirm that the proposed approach may represent a valid solution to the label switching problem, giving overall good results in detecting the pivotal units for the mixture components even when no trivial partition is derivable from the data structure.

5 Conclusions

A methodology for dealing with the relabelling issue in Bayesian mixture models is proposed. It requires the identification of the pivotal units of the model components and consists in post-processing the MCMC chains in such a way it enables us to draw inference on both probabilities of each unit being in each group and the group parameters. The performance of the method is explored via a simulation exercise, in terms of the ability in identifying the pivots according to different criteria. Possible applications of the proposed approach and the comparison with other available methods will be discussed.

References

1. Chung, H., Loken, E., Schafer, J.L.: Difficulties in drawing inferences with finite-mixture models. The American Statistician **58(2)** (2004)
2. Frühwirth-Schnatter, S.: Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association **96(453)**, 194–209 (2001)
3. Jasra, A., Holmes, C.C., Stephens, D.A.: Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. Statistical Science 50–67 (2005)
4. Puolamäki, K., Kaski, S.: Bayesian solutions to the label switching problem. In: Advances in Intelligent Data Analysis VIII, pp. 381–392. Springer (2009)
5. Stephens, M.: Dealing with label switching in mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **62(4)**, 795–809 (2000)
6. Yao, W., Li. L.: An online Bayesian mixture labelling method by minimizing deviance of classification probabilities to reference labels. Journal of Statistical Computation and Simulation **84(2)**, 310–323 (2014)