Estimating dependent Binomial mixture models through reversible jump MCMC

Daiane Aparecida Zuanetti
Luis Aparecido Milan

Departamento de Estatística, UFSCar - Brazil

Abstract: We present a hidden Markov model of Binomial variables as a dependent mixture model and propose the reversible jump procedure to estimate the number of components and parameters of the model and identify from which component each observation comes from. We present some simulations to illustrate the performance of the method and apply to diabetes data.

Key words: Binomial mixture models, dependence structure, reversible jump.

1 Introduction

Hidden Markov models (HMM) have been applied in many research areas. A few examples are econometrics (Hamilton, 1989; Biblio *et al.*, 1998), finance (Rydén *et al.*, 1998), speech recognition (Rabiner, 1989), biology and genetic (Churchill, 1989 and 1992; Boys *et al.*, 2000 and 2002).

HMM is compound by two sequences of random variables, one is observable and another is unobservable. Usually, the unobservable sequence is a (finite-state) Markov chain and define the probability distribution of each variable of the observable sequence. Therefore, HMM can be defined as a mixture model with dependent variables.

Inferences for HMM was firstly studied by Baum and Petrie (1966) and Baum *et al.* (1970) that propose maximum likelihood estimators for the case that K, number of components, is known. These estimators are obtained using EM algorithm. When K is unknown, Bayesian approach is an attractive alternative to estimate the model since it allows to choose the most probable model and estimate the parameters jointly. Robert *et al.* (2000) propose the reversible jump for a HMM of normal variables.

We consider a mixture model with first-order dependence of Binomial variables and propose the Bayesian approach reversible jump procedure, RJ, to estimate the model. It does model selection over all possible models with different number of components.

The article is organized as follows: Section 2 presents HMM as a dependent mixture model, Section 3 details the Bayesian methodology and the reversible jump MCMC algorithm. In Section 4, we analyze the performance of the method on simulated and real data sets. We conclude with a discussion in the Section 6.

2 HMM as a dependent mixture model

Let $\mathbf{S} = \{S_1, S_2, ..., S_T\}$ be a Markov chain, where $S_t \in \{1, 2, ..., K\}$, for $t = 1, 2, ..., T$, $Pr\left(S_1 = s_1\right) = p_{0s_1}$ and $Pr\left(S_t = s_t | S_{t-1} = s_{t-1}, ..., S_1 = s_1\right) = Pr\left(S_t = s_t | S_{t-1} = s_{t-1}\right) = p_{s_{t-1}s_t}$, for $t = 2, 3, ..., T$. Let $\mathbf{Y} = \{Y_1, Y_2, ..., Y_T\}$ be a sequence of Binomial random variables with density given by $f_{Y_t|S_t=k}(y_t) = f_{Y_t}(y_t|\theta_k)$, for $t = 1, 2, ..., T$, and $k \in \{1, 2, ..., K\}$, where θ_k is the unknown success probability associated to the k-th component.

Suppose K known and let

1. $P = \{p_{jk}\}$ be the transition matrix of $\mathbf{S}$, $p_{jk} = Pr(S_{t+1} = k|S_t = j)$, for $j, k \in \{1, 2, ..., K\}$;

2. $\mathbf{p}_0 = (p_{01}, ..., p_{0K})$ be the initial probability of S_1, with $p_{0k} = Pr(S_1 = k)$; and

3. $\theta = (\theta_1, ..., \theta_K)$ be the success probabilities of Binomial distributions associated to states of S_t.

The joint distribution of $\mathbf{Y}$ and $\mathbf{S}$ is $f_{\mathbf{Y},\mathbf{S}}(\mathbf{y}, \mathbf{s}|\theta) = \prod_{t=1}^{T} p_{s_{t-1}s_t} f_{Y_t}\left(y_t|\theta_{s_t}\right)$, where $p_{s_0 s_1} = p_{0s_1}$. As sequence $\mathbf{S}$ is non-observable, the probability distribution of $\mathbf{Y}$ can be written as

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}|\theta) &= \sum_{\mathbf{s}} f_{\mathbf{Y},\mathbf{S}}(\mathbf{y}, \mathbf{s}|\theta) = \sum_{s_1=1}^{K} p_{s_0 s_1} f_{Y_1}\left(y_1|\theta_{s_1}\right) \ldots \sum_{s_T=1}^{K} p_{s_{T-1}s_T} f_{Y_T}\left(y_T|\theta_{s_T}\right) \\ &= \prod_{t=1}^{T} \sum_{s_t=1}^{K} p_{s_{t-1}s_t} f_{Y_t}\left(y_t|\theta_{s_t}\right) \end{aligned} \quad (1)$$

which characterizes the observable variable Y_t, $t = 1, ..., T$, as a mixture of K distributions.

The augmented likelihood function for θ, P, $\mathbf{p}_0$ and K is defined as

$$L(\theta, \mathbf{p}_0, P, K|\mathbf{y}, \mathbf{s}, \mathbf{m}) = \left\{ \prod_{t=1}^{T} \left(\begin{array}{c} m_t \\ y_t \end{array} \right) \right\} \left\{ \prod_{k=1}^{K} \theta_k^{\sum\limits_{t:s_t=k} y_t} \left(1 - \theta_k\right)^{\sum\limits_{t:s_t=k} (m_t - y_t)} \left(\prod_{j=0}^{K} p_{jk}^{n_{jk}} \right) \right\}, \quad (2)$$

where $\mathbf{m} = \{m_1, ..., m_T\}$, $n_{0k} = I\left(S_1 = k\right)$ and $n_{jk} = \sum_{t=2}^{T} I\left(S_t = j, S_{t+1} = k\right)$, for $j, k = 1, ..., K$.

3 Bayesian approach

Consider K is unknown and $\mathbf{p}_j$'s and θ_k's, for $j = 0, 1, ..., K$ and $k = 1, ..., K$, to be independent *a priori*. The joint *a priori* distribution for parameters θ, $\mathbf{p}_0$, P and K is

$$f(\theta, \mathbf{p}_0, P, K) = f(K) \left(\prod_{j=0}^{K} f(\mathbf{p}_j|K) \right) \left(\prod_{k=1}^{K} f(\theta_k) \right) \quad (3)$$

where we assume

1. $K \sim \text{Uniform}(1, 2, ..., K_{max})$, where K_{max} indicates the maximum value of K;

2. $\mathbf{p}_j|K \sim \text{Dirichlet}(\gamma_{j1}, ..., \gamma_{jK})$, for $j = 0, ..., K$ and $\gamma_{jK} > 0$ are known hyper-parameters; and

3. $\theta_k \sim \text{Beta}(\alpha_k, \beta_k)$, for $k = 1, ..., K$ and $\alpha_k > 0$ and $\beta_k > 0$ are known hyper-parameters.

Combining the likelihood function in eq. (2) with the *a priori* distribution, we obtain the following conditional *a posteriori* distributions

1. $\mathbf{p}_0|\left(\theta, P, K, \mathbf{y}, \mathbf{s}, \mathbf{m}\right) \sim \text{Dirichlet}(n_{01} + \gamma_{01}, ..., n_{0K} + \gamma_{0K})$;

2. $\mathbf{p}_k|\left(\theta, \mathbf{p}_0, P_{-\mathbf{p}_k}, K, \mathbf{y}, \mathbf{s}, \mathbf{m}\right) \sim \text{Dirichlet}(n_{k1} + \gamma_{k1}, ..., n_{kK} + \gamma_{kK})$, $k = 1, ..., K$;

3. $\theta_k|\left(\theta_{-\theta_k}, \mathbf{p}_0, P, K, \mathbf{y}, \mathbf{s}, \mathbf{m}\right) \sim \text{Beta}\left(\sum\limits_{t:s_t=k} y_t + \alpha_k, \sum\limits_{t:s_t=k} (m_t - y_t) + \beta_k \right)$, $k = 1, ..., K$; and

4. $S_t|\left(\theta, \mathbf{p}_0, P, K, \mathbf{y}, \mathbf{s}_{-s_t}, \mathbf{m}\right) \sim \text{Multinomial}(1, (\delta_{t1}, ..., \delta_{tK}))$, where $\delta_{tk} = \frac{p_{s_{t-1}k} f_{Y_t}(y_t|\theta_k) p_{ks_{t+1}}}{\sum_{k=1}^{K} p_{s_{t-1}k} f_{Y_t}(y_t|\theta_k) p_{ks_{t+1}}}$, $p_{kS_{T+1}} = 1$ and $f_{Y_t}(y_t|\theta_k)$ is the Binomial probability function of the k-th component.

3.1 Reversible jump algorithm

One of the goals in this model is to estimate K and it can be done through MCMC reversible jump methodology. In a mixture model, the movements that suggest dimension change are called split or merge moves. A split breaks a component in two components increasing K in one and a merge joins two components decreasing K in one.

To describe the merge move, consider that the current state of MCMC algorithm is $\overline{x} = (\overline{P}, K+1, \overline{\theta}, \overline{\pi}, \overline{\mathbf{p}}_0)$, where $\pi = (\pi_1, ..., \pi_K)$ represents the stationary distribution of P, such that $\pi = \pi P$. We choose randomly one pair of components (j_1, j_2), combine them in one single component j_* and create a new state $x = (P, K, \theta, \pi, \mathbf{p}_0)$. This movement can be created through the following steps:

1. do $s_t = j_*$, if $\overline{s}_t = j_1$ or $\overline{s}_t = j_2$, or $s_t = \overline{s}_t$, otherwise, for $t = 1, 2, ..., T$;

2. do $\theta_{j_*} = (\overline{\theta}_{j_1} + \overline{\theta}_{j_2})/2$ and remaining θ_k's are copied from $\overline{x}$;

3. do $p_{j_*k} = \frac{\overline{\pi}_{j_1}}{\overline{\pi}_{j_1}+\overline{\pi}_{j_2}} \overline{p}_{j_1k} + \frac{\overline{\pi}_{j_2}}{\overline{\pi}_{j_1}+\overline{\pi}_{j_2}} \overline{p}_{j_2k}$ for $k \neq j_*$, $p_{kj_*} = \overline{p}_{kj_1} + \overline{p}_{kj_2}$ for $k \neq j_*$ and remaining p_{kj} are copied from $\overline{x}$ or obtained by equating row sums to one; and

4. do $p_{0j_*} = \overline{p}_{0j_1} + \overline{p}_{0j_2}$ and remaining p_{0k} are copied from $\overline{\mathbf{p}}_0$.

To describe the split move, consider the current state of MCMC algorithm $x = (P, K, \theta, \pi, \mathbf{p}_0)$. We choose randomly one component j_* to break into two new components (j_1, j_2) and create a new state $\overline{x} = (\overline{P}, K+1, \overline{\theta}, \overline{\pi}, \overline{\mathbf{p}}_0)$. According to Robert *et al.* (2000), this movement need to preserve the stationary probabilities of P, i.e., $\overline{\pi}_k = \pi_k$ for $k \neq j_1, j_2$, $\overline{\pi}_{j_1} = u_0 \pi_{j_*}$ and $\overline{\pi}_{j_2} = (1-u_0)\pi_{j_*}$, for $0 < u_0 < 1$. This can be done through the following steps

1. do $\overline{p}_{j_1k} = \frac{\mu_j}{u_0} p_{j_*k}$, $\overline{p}_{j_2k} = \frac{(1-\mu_k)}{(1-u_0)} p_{j_*k}$, for $k \neq j_1, j_2$;
   $\overline{p}_{kj_1} = \nu_k p_{kj_*}$, $\overline{p}_{kj_2} = (1-\nu_k) p_{kj_*}$, for $k \neq j_1, j_2$; $\overline{p}_{j_1j_2} = u_1 \left(1 - \sum_{k \neq j_*} \frac{\mu_k}{u_0} p_{j_*k}\right)$;
   $\overline{p}_{j_2j_1} = \left\{(1-u_1) \sum_{k \neq j_*} \mu_k p_{j_*k} - u_0 u_1 - \sum_{k \neq j_*} \frac{\pi_k}{\pi_{j_*}} \nu_k p_{kj_*}\right\} / (1-u_0)$;
   $\overline{p}_{j_1j_1}$ and $\overline{p}_{j_2j_2}$ are obtained by equating row sums to one and remaining $\overline{p}_{kj}$'s are copied from P and $0 < u_i, \mu_k, \nu_k < 1$, for $i = 0, 1$ and $k = 1, 2, ..., K-2$. We assume $u_i, \mu_k, \nu_k \sim$ Beta$(2,2)$, all independent. All elements of the new matrix $\overline{P}$ need to be a value between zero and one. If it doesn't happen, the step is rejected.

2. do $\overline{\theta}_{j_1} = \theta_{j_*} - \omega$, $\overline{\theta}_{j_2} = \theta_{j_*} + \omega$ and remaining $\overline{\theta}_k$'s are copied from θ. We assume $\omega \sim$ Uniform$(0, min(\theta_{j_*}, 1-\theta_{j_*}))$;

3. do $\overline{p}_{0j_1} = \lambda p_{0j_*}$, $\overline{p}_{0j_2} = (1-\lambda) p_{0j_*}$ and remaining $\overline{p}_{0i}$'s are copied from $\mathbf{p}_0$. We assume that $\lambda \sim$ Beta$(2,2)$; and

4. since $\overline{P}$, $\overline{\theta}$ and $\overline{\mathbf{p}}_0$ are already created, do $\overline{S}_t = s_t$, if $s_t \neq j_*$, or simulate $\overline{S}_t$ from its conditional *a posteriori* distribution considering only components j_1 and j_2, if $s_t = j_*$, for $t = 1, 2, ..., T$.

The acceptance probabilities for the split and merge moves are, respectively, $min(1, A)$ and $min(1, 1/A)$, where

$$A = \frac{L\left(\overline{\theta}, \overline{\mathbf{p}}_0, \overline{P}, K+1 | \mathbf{y}, \overline{\mathbf{s}}, \mathbf{m}\right)}{L\left(\theta, \mathbf{p}_0, P, K | \mathbf{y}, \mathbf{s}, \mathbf{m}\right)} \frac{f(K+1) f(\overline{P}|K+1) f(\overline{\theta}|K+1) f(\overline{\mathbf{p}}_0|K+1)}{f(K) f(P|K) f(\theta|K) f(\mathbf{p}_0|K)} \frac{d_{j_1j_2}}{b_{j_*} P_{alloc}} \frac{J}{g(\mathbf{u})}, \qquad (4)$$

$d_{j_1j_2}$ is the probability of choosing the merge movement between the components j_1 and j_2, b_{j_*} is the probability of choosing the split movement of the component j_*, P_{alloc} is the probability of a specific allocation of $\overline{S}_t$ defined as the product of conditional *a posteriori* probabilities used to allocate the observations, $g(\mathbf{u})$ is the joint distribution of $(u_0, u_1, \mu\text{'s}, \nu\text{'s}, \omega, \lambda)$ and J is the Jacobian determinant of the transformations used to complete the dimension of $\overline{P}$, $\overline{\theta}$ and $\overline{\mathbf{p}}_0$ in split movement. The calculation of Jacobian determinant is better described in Robert *et al.* (2000).

4 Examples

We apply the proposed method to simulated data sets and a real data set to estimate the parameters of the model. We consider a mixture of Binomial distributions and set hyper-parameters $\gamma_{jk} = 1$, $\gamma_k = 1$, $\alpha_k = 1$ and $\beta_k = 1$, for $k = 1, ..., K + 1$ and $j = 0, 1, ..., K + 1$.

4.1 Simulated data sets

We consider a mixture of $k = 2$ Binomial distributions in a sequence with $T = 100$ observations and parameters $\theta_1 = 0.48$ and $\theta_2 = 0.38$. We simulate 12 distinct situations with different m_t's, for $t = 1, ..., T$. We vary m_t between $[20, 30]$, $[50, 60]$, $[80, 90]$ and $[110, 120]$, and fix three ways to define the Binomial distribution in each position t. First, we simulate independent S_t's (situation called $CH1$). Second, we define four change points in $\mathbf{S}$ at positions 21, 41, 61 and 81, where 20 consecutive observations are generated from the same Binomial distribution and then we change to other Binomial distribution (situation called $CH2$). Third, we define just one change point in $\mathbf{S}$ at position 51, where the first 50 observations are generated from the Binomial(m_t, θ_1) and the last 50 observations are generated from the second Binomial(m_t, θ_2) (situation called $CH3$).

We run RJ $B = 200000$ iterations, discard the $L = 1000$ first iterations and consider one draw for every 5 element of the chain to result in 39800 simulated values for each parameter. The chains are always initialized with $K = 1$. We analyze the convergence and conclude the number of iterations is enough for reliable results.

Table 1 shows the *a posteriori* probability for K in each situation. The highest *a posteriori* probability for each situation is in bold. We note RJ has a good performance to estimate K when $m_t \geq 80$ or in first-order cases. However, for independent model (CH1) and small m_t's ($m_t < 60$) we do not have a good performance. The performance is expected to be better if the θ's are more apart.

Table 1: *A posteriori* probability of number of components K

K	m_t											
	$[20, 30]$			$[50, 60]$			$[80, 90]$			$[110, 120]$		
	$CH1$	$CH2$	$CH3$	$CH1$	$CH2$	$CH3$	$CH1$	$CH2$	$CH3$	$CH1$	$CH2$	$CH3$
1	0.245	**0.652**	0.001	**0.786**	0.006	0.001	0.000	0.000	0.000	0.000	0.000	0.000
2	**0.477**	0.265	**0.898**	0.181	**0.852**	**0.851**	**0.697**	**0.869**	**0.970**	**0.620**	**0.857**	**0.942**
3	0.207	0.068	0.095	0.028	0.127	0.132	0.240	0.128	0.029	0.306	0.137	0.055
4	0.055	0.012	0.005	0.003	0.014	0.014	0.053	0.003	0.001	0.062	0.005	0.003
5	0.012	0.002	0.000	0.000	0.001	0.002	0.008	0.000	0.000	0.010	0.001	0.000
≥ 6	0.003	0.001	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.001	0.000	0.000

Although the estimates and credibility intervals are not shown here, the method estimates precisely the parameters θ and $\mathbf{S}$ when $K = 2$. We also note that algorithm converges to $K = 2$, the real value of K, very quickly (before of the first 50 iterations) and hardly moves to another value of K.

The acceptance rates are represented at Table 2. They are lower than desired, however, since the moves involve only a change of model and all other parameters are updated in each sweep, it is not a problem, but an evidence that when the procedures go to the suitable model is very difficult to find another model that has similar probability of it and make the change. This fact is confirmed by the reduction of the acceptance rate when m_t or the expected time to make a transition between different Binomial distribution increase. In these cases, the suitable models are more evident and the probability of changing the model is lower.

Table 2: Acceptance rates of split/merge moves

m_t	$CH1$	$CH2$	$CH3$
$[20, 30]$	2.10%	2.31%	0.32%
$[50, 60]$	1.57%	0.52%	0.32%
$[80, 90]$	0.65%	0.30%	0.08%
$[110, 120]$	0.49%	0.36%	0.11%

4.2 Number of diagnosed diabetes

The incidence of diabetes has increased in all over the world, mainly because of sedentary lifestyles and different eating habits. From 1980 through 2011, the percentage of people in USA diagnosed with diabetes increased 167%. We analyze the USA number of diagnosed diabetes by age available in website 'www.cdc.gov/ diabetes/statistics/prev/national/figbyage.htm'. We present here the range of $65 - 74$ ages. The sequence of yearly number of diagnosed diabetes, from 1980 through 2011 ($T = 32$), is a realization of $\mathbf{Y}$, where $Y_t|S_t = k \sim \text{Binomial}(m_t, \theta_k)$, for $t = 1, 2, ..., T$ and $k \in \{1, 2, ..., K\}$, m_t is known.

We run RJ $B = 101000$ iterations, discard the $L = 1000$ first iterations and consider one draw for every 10 element of the chain to result in 10000 simulated values for each parameter. The chain are initialized with $K = 1$. We analyze the convergence and the number of iterations is sufficient for results to be considered reliable.

Table 3 shows the *a posteriori* probabilities for K. As we observe, the maximum probability is for $K = 3$, *i.e.*, in analyzed period of time there are three different rates of diagnosing diabetes for ages between 65 and 74.

Table 3: *A posteriori* probability for the number of components K

K	estimate
≤ 2	0.000
3	0.921
4	0.078
5	0.001

The three rates and their 95% credibility intervals are presented in Table 4. The procedure identifies the component from each observation comes from and the change points are located at $t = 17$ (1996 year) and $t = 23$ (2002 year). We also observe that θ_1 estimate (average of simulated values) is almost the double of θ_3 estimate, *i.e.*, diagnosed diabetes rate nearly doubled between 1996 and 2002.

Table 4: Estimates and 95% credibility intervals for diagnosed diabetes rates

	Estimate
θ_1	$0.19(0.18 - 0.20)$
θ_2	$0.14(0.13 - 0.16)$
θ_3	$0.10(0.09 - 0.10)$

5 Conclusion

We present the HMM as a dependent mixture model and propose a reversible jump procedure to estimate models with unknown number of components. The method presents a good performance under tested situations and, for the Binomial mixture model, we conclude that it is more precise when the number of trials or the expected permanence time of the sequence in the same component increases.

References

[1] Baum, L.E. e Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist* **37**, 1554-1563.

[2] Baum, L.E., Petrie, T., Soules, G. e Weiss, N. (1970). A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist* **41**, 164-171.

[3] Biblio, M., Monfort, A. e Robert, C.P. (1998). Bayesian estimation of switching ARMA models. *J. Econometrics* (to appear).

[4] Boys, R. e Henderson, D. (2002). On determining the order of Markov dependence of an observed process governed by a hidden Markov model. *Scientifc Programming* **10**, 241-251.

[5] Boys, R., Henderson, D. e Wilkinson, D. (2000). Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **49**,269-285.

[6] Churchill, G. (1992). Hidden Markov chains and the analysis of genome structure. *Computers and Chemistry* **16**,107-115.

[7] Churchill, G. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51**, 79-94.

[8] Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrika* **57**, 357-384.

[9] Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**, 257-285.

[10] Robert, C.P., Rydén, T. e Titterington, D.M. (2000). Bayesian inference in hidden Markov models through reversible jump Markov chain Monte Carlo. *Statistical Methodology* **62**, 57-75.

[11] Rydén, T., Teräsvirta, T. e Asbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model. *J. Appl. Econometrics* (to appear).