Empirical Bayes estimation of parameters in Markov transition probability matrix with computational methods

Babulal Seal* and Sk Jakir Hossain

Department of Statistics, Burdwan University, Burdwan, West Bengal, India

(Received 20 January 2014; accepted 5 September 2014)

Empirical Bayes estimator for the transition probability matrix is worked out in the cases where we have belief regarding the parameters, For example, where the states seem to be equal or not. In both cases, priors are in accordance with our beliefs. Using EM algorithm, computational methods for different hyperparameters of the empirical Bayes are described. Also, robustness of empirical Bayes procedure is investigated.

Keywords: empirical Bayes; robustness; EM algorithm; Dirichlet prior

1. Introduction

Markov model is important in Internal migration [10] behaviour of population. Estimation of parameters of transition probability matrix in classical way is available in literature. When behaviour of the migration pattern is seen, that knowledge is to be employed in finding Bayes procedure [6–8]. For example, if we have a belief that all states are almost equally important, then we should try to add such beliefs in estimation procedures. In this paper, we consider empirical Bayes estimates in the light of such beliefs.

It is also important to suggest a prior distribution for a problem. Actually, the system will speak out the general form of prior. Form of prior can be understood from the form of the problem. What we do, we try to make the prior finer and finer from the data set only. Rather some of us give emphasis on data and not on prior and that is not the spirit of Bayes procedure. So once we have feeling about prior, then rational thing will be taken as general form and then from data, we make this concrete as far as possible. So in our case, that is, in internal migration problem or in transition probability, the general form of prior must be of multivariate Beta and Dirichlet prior and from data we can make this as concrete as possible by Bayes procedure. That is done earlier. For more general forms in terms of variations of hyperparameters of prior

*Corresponding author. Email: babulal_seal@yahoo.com

© 2014 Taylor & Francis
and using data, we get empirical Bayes procedure. So, in Section 2, empirical Bayes estimations with transition data are considered under two different beliefs: (i) When the hyperparameters are the same for different states, that is, it gives knowledge whether the states have equal performance, (ii) When we do not have such beliefs. Then, the next problem is to estimate the hyperparameters if not known. These are included in Sections 3 and 4 under different beliefs as in (i) and (ii) above. The computational methods are worked out using EM algorithm. We suggest a way to get an initial value and give steps to get an accurate value of the hyperparameters. Finally, in Section 5, robustness of such model, that is, model of prior using possible variation of hyperparameters, is given. But it is not given in general form. For some prior hyperparameters, it is seen good. If it is for other also, then we can say that it is more or less close to Bayes estimate. But this numerical work is for clarity. The relevant R-codes used for computations are given in Appendix 2.1 In spite of that, our achievement is to get a general form or model in this context. This is considered in Section 6. There, from a data set, we estimate hyperparameters for the scientific significance of the problem and the relevant R-codes are given in Appendix 3.

2. Empirical Bayes estimation from transition data

In this section, we want to find out the estimators of the unknown parameters involved in the prior distribution. It can be seen that the Bayes estimator $\delta(n_{ij})$ of $p_{ij}$ depends on prior distribution $\pi(p|\alpha)$. When $\alpha_{ij}$’s are unknown, it is not possible to implement $\delta(\cdot)$. The empirical Bayes approach is employed to combine information from observations $n_i$, $i=1,2,\ldots,k$.

If the prior distribution of $p$ is not completely known, the posterior distribution cannot be used to make probability statements about the unknown variables. But, in some situations, it may be possible to compute the posterior distribution of $p$ given $n$ without the complete knowledge of the distribution function of $p$. But the estimate may involve unknown parameters which are characteristics of the marginal distribution of $n$ alone.

Case 1 Priors have the same hyperparameters

Let

$$n_i = (n_{i1}, n_{i2}, \ldots, n_{ik}) \sim \text{Multinomial}(n_i; p_{i1}, p_{i2}, \ldots, p_{ik})$$

and

$$p_i = (p_{i1}, p_{i2}, \ldots, p_{ik}) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_k) \quad \text{for all } i = 1, 2, \ldots, k.$$  

That is, the $k$ groups are tied together by the common prior distribution.

Therefore, the Bayes estimator of $p_{ij}$ under sum of squared error loss is

$$\delta(n_{ij}) = \frac{n_{ij} + \alpha_i}{k\alpha_i + n_i},$$

where

$$n_i = \sum_{j=1}^{k} n_{ij} \quad \text{for all } i = 1, 2, \ldots, k.$$  

But if the Dirichlet parameters are unknown, we are interested in finding out ML estimate of $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ based on available observations $n_i$, for all $i = 1, 2, \ldots, k$. Here, $n_i = \sum_{j=1}^{k} n_{ij}$, for all $i = 1, 2, \ldots, k.$
Given \((n_1, n_2, \ldots, n_k)\), the maximum-likelihood estimators of \((p_1, p_2, \ldots, p_k)\) are the relative frequencies
\[
\hat{p}_{ij} = \frac{n_{ij}}{\sum_{j=1}^{k} n_{ij}} = \frac{n_{ij}}{n_i} \quad \text{for all } i, j = 1, 2, \ldots, k.
\]
If we write
\[
y_{ij} = \frac{n_{ij}}{n_i} \quad \text{for all } i, j = 1, 2, \ldots, k
\]
and
\[
y_i = (y_{i1}, y_{i2}, \ldots, y_{ik})^T \quad \text{for all } i = 1, 2, \ldots, k,
\]
then, the vector \(y_i\) approximately follows the Dirichlet distribution \([2]\).

Without loss of generality, let \(Y_1, Y_2, \ldots, Y_k\) be \(k\) sets of random vector from Dirichlet \((\alpha_1, \alpha_2, \ldots, \alpha_k)\).

Then, the likelihood function is given by
\[
L(\alpha) = \left(\frac{\Gamma(\alpha_i)}{\prod_{j=1}^{k} \Gamma(\alpha_j)}\right)^k \prod_{i=1}^{k} (y_{i1}^{\alpha_i-1} y_{i2}^{\alpha_2-1} \ldots y_{ik}^{\alpha_k-1}),
\]
where \(\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)^T\) and \(\alpha_\cdot = \sum_{j=1}^{k} \alpha_j\)
\[
\frac{\delta \log L(\alpha)}{\delta \alpha_j} = k \psi(\alpha_j) - k \psi(\alpha_\cdot) + \sum_{i=1}^{k} \log y_{ij},
\]
where \(\psi(t) = (\delta/\delta t)\log \Gamma(t)\) is a digamma function.

Then, the maximum-likelihood equations for estimators of \((\alpha_1, \alpha_2, \ldots, \alpha_k)\) are given by Johnson and Kotz \([3]\)
\[
\psi(\hat{\alpha}_j) - \psi(\hat{\alpha}_\cdot) = \frac{1}{k} \sum_{j=1}^{k} \log(y_{ij}) \quad \text{for all } j = 1, 2, \ldots, k.
\]
As the likelihood equation, it clearly does not yield an explicit solution for \(\hat{\alpha}_j\). Thus, Equation (2) must be solved by numerical methods.

**Case 2** Priors have different hyperparameters.

Now, let \(n\) be the random vector corresponding to the observed data having Multinomial distribution with parameter \(p\) (unknown), where
\[
n_i \sim \text{Multinomial}(n_i; p_{i1}, p_{i2}, \ldots, p_{ik}) \quad \text{for all } i = 1, 2, \ldots, k
\]
such that
\[
n_i = \sum_{j=1}^{k} n_{ij}, \quad \sum_{j=1}^{k} p_{ij} = 1 \quad \text{for all } i = 1, 2, \ldots, k
\]
and \(p_i\) have Dirichlet prior distribution, i.e.
\[
p_i \sim \text{Dirichlet}(\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik}) \quad \text{for all } i = 1, 2, \ldots, k.
\]

Then, the Bayes estimator of \(p_{ij}\) \([6–8]\) under squared error loss is
\[
\delta(n_{ij}) = \frac{n_{ij} + \alpha_{ij}}{n_i + \alpha_i} \quad \text{for all } i, j = 1, 2, \ldots, k.
\]
It is noticed that Equation (1) can be used here also, using \(k = 1\). However, in both cases, we need extensive computation.
3. Computational method when hyperparameters are the same for \( k \)-places

If we use the approximation

\[
\psi(t) \equiv \log\left(t - \frac{1}{2}\right)
\]

then from Equation (2) we have

\[
\log\left(\alpha_j - \frac{1}{2}\right) - \log\left(\alpha - \frac{1}{2}\right) = \frac{1}{k} \sum_{i=1}^{k} \log(y_{ij}) \quad \text{for all } j = 1, 2, \ldots, k
\]

\[
\Rightarrow \log\left(\frac{\alpha_j - (1/2)}{\alpha - (1/2)}\right) = \log\left(\prod_{i=1}^{k} y_{ij}^{1/k}\right) \quad \text{for all } j = 1, 2, \ldots, k.
\]

Thus, the approximate values of \((\hat{\alpha}_j - \frac{1}{2})/(\hat{\alpha} - \frac{1}{2})\), for all \( j = 1, 2, \ldots, k \) are given by

\[
\frac{\alpha_j - (1/2)}{\alpha - (1/2)} = \prod_{i=1}^{k} y_{ij}^{1/k} \quad \text{for all } j = 1, 2, \ldots, k
\] (3)

summing over all \( j \), we have

\[
\frac{\alpha - (k/2)}{\alpha - (1/2)} = \sum_{j=1}^{k} \prod_{i=1}^{k} y_{ij}^{1/k}
\]

\[
\Rightarrow 1 - \frac{(k - 1)/2}{\alpha - (1/2)} = \sum_{j=1}^{k} \prod_{i=1}^{k} y_{ij}^{1/k}
\]

\[
\Rightarrow \alpha - \frac{1}{2} = \frac{(k - 1)/2}{1 - \sum_{j=1}^{k} \prod_{i=1}^{k} y_{ij}^{1/k}}.
\]

Thus, from Equation (3) we have the estimators of \( \alpha_j \), which have approximations to \( \hat{\alpha}_j \) given by

\[
\hat{\alpha}_j = \frac{1}{2} + \frac{(k - 1)/2 \prod_{i=1}^{k} y_{ij}^{1/k}}{1 - \sum_{j=1}^{k} \prod_{i=1}^{k} y_{ij}^{1/k}} \quad \text{for all } j = 1, 2, \ldots, k.
\] (4)

Starting from these values of \( \hat{\alpha}_j \) from Equation (4), solutions of Equation (2) can be obtained by an iterative process.

Let us denote the value of \( \alpha \) at the \( r \)th subsequent iteration of the Newton–Raphson method by \( \alpha^{(r)} \). Then, the value of \( \alpha \) in the next iteration is given by

\[
\alpha^{(r+1)} = \alpha^{(r)} + I(\alpha^{(r)})^{-1} S(\alpha^{(r)}),
\]

where \( I(\alpha) = - (\delta^2 / \delta \alpha \delta \alpha') \log L(\alpha) \) is the observed information matrix and \( S(\alpha) = (\delta / \delta \alpha) \log L(\alpha) \) is the score vector.
From the likelihood function (1), the score has entries
\[
\frac{\delta}{\delta \alpha_j} \log L(\alpha) = k \psi(\alpha_j) - k \psi(\alpha) + \sum_{i=1}^{k} \log y_{ij}.
\]
The observed information has entries
\[
-\frac{\delta^2}{\delta \alpha_j \delta \alpha_j'} \log L(\alpha) = k \{1_{[j=j']} \psi'(\alpha_j) - \psi'(\alpha)\},
\]
where \(1_{[j=j']}\) is the indicator function of the event \(\{j=j'\}\), and \(\psi'(t)\) is the trigamma function \((\delta^2/\delta t^2) \log \Gamma(t)\).

The observed information can be summarised in matrix form by
\[
-\frac{\delta^2}{\delta \alpha \delta \alpha'} \log L(\alpha) = I(\alpha) = k(D - c1^T)
\]
where \(D\) is a diagonal matrix with the \(j\)th diagonal entry \(d_j = \psi'(\alpha_j), j = 1, 2, \ldots, k\)

\(c\) is the constant \(\psi'(\alpha)\), and \(1\) is a column vector of all 1’s.

Thus, the limiting value of \(\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)\) as \(\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_k^*)\) gives the empirical Bayes estimate when all prior distributions have the same set of parameters.

4. Computational method when hyperparameters are different

Consider the case where \(\alpha_{ij}\)’s are unknown parameter, for \(i, j = 1, 2, \ldots, k\). We take
\[
\alpha = (\alpha_1^T, \alpha_2^T, \ldots, \alpha_k^T)^T = (\alpha_{11}, \ldots, \alpha_{1k}; \alpha_{21}, \ldots, \alpha_{2k}; \ldots; \alpha_{k1}, \ldots, \alpha_{kk})^T.
\]
We are interested in finding out the estimate of the Dirichlet parameter \(\alpha\) from observed data \(n\). But it is not possible to estimate \(\alpha\) from observed data analytically as if we try separately for each component using Equation (4); then, for \(k = 1\) in case of individual state, it is seen that \(\hat{\alpha}_{ij} = 1/2\) which is absurd. However, in this case, it may be possible to compute iteratively the MLE of \(\alpha\) from observed data by using the quasi-Newton accelerated EM algorithm.

We estimate \(\alpha_i\) individually, since estimates of \(\alpha_i = (\alpha_{i1}, \ldots, \alpha_{ik})\) depend only on \(n_i = (n_{i1}, \ldots, n_{ik})\) and independent of \(n_j = (n_{j1}, \ldots, n_{jk})\), for \(j \neq i\).

Thus, we iterate each \(\alpha_i\) individually and take the limiting value of \(\alpha_i\) as \(\alpha_i^*\), for all \(i = 1, 2, \ldots, k\).

Finally, we have
\[
\alpha^* = (\alpha_1^T, \ldots, \alpha_k^T)^T.
\]

Before that, by \(x\) we denote the vector containing the augmented and by \(p\), we denote the vector containing the additional data, referred to as the unobservable data. Then, the complete-data
vector $\mathbf{x}$ is taken to be
\[
\mathbf{x} = (\mathbf{n}^T, \mathbf{p}^T)^T.
\]
The variables $(p_{i1}, p_{i2}, \ldots, p_{ik})$ are defined so that
\[
f(n_{i1}, n_{i2}, \ldots, n_{ik} | p_{i1}, p_{i2}, \ldots, p_{ik}) = \frac{n_{i1}!}{n_{i1}! n_{i2}! \cdots n_{ik}!} p_{i1}^{n_{i1}} p_{i2}^{n_{i2}} \cdots p_{ik}^{n_{ik}}
\]
for all $i = 1, 2, \ldots, k$ and
\[
\pi(p_{i1}, p_{i2}, \ldots, p_{ik}) = \frac{\Gamma \left( \sum_{j=1}^{k} \alpha_{ij} \right)}{\prod_{j=1}^{k} \Gamma(\alpha_{ij})} p_{i1}^{\alpha_{1i}-1} p_{i2}^{\alpha_{2i}-1} \cdots p_{ik}^{\alpha_{ki}-1} \quad \text{for all } i = 1, 2, \ldots, k.
\]
After integrating w.r.t. $(p_{i1}, p_{i2}, \ldots, p_{ik})$ from joint density function of $(n_{i1}, n_{i2}, \ldots, n_{ik})$ and $(p_{i1}, p_{i2}, \ldots, p_{ik})$ that is formed from Equations (5) and (6), the density function of $(n_{i1}, n_{i2}, \ldots, n_{ik})$, for all $i = 1, 2, \ldots, k$ is given by
\[
f(n_{i1}, n_{i2}, \ldots, n_{ik} | \alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik}) = \frac{n_{i1}!}{\prod_{j=1}^{k} n_{ij}!} \frac{\Gamma \left( \sum_{j=1}^{k} \alpha_{ij} \right)}{\prod_{j=1}^{k} \Gamma(\alpha_{ij})} \int_{S_i} \prod_{j=1}^{k} p_{ij}^{n_{ij}+\alpha_{ij}-1} \prod_{j=1}^{k} dp_{ij},
\]
where the integration is carried out over the region
\[
S_i = \left\{ p_{ij} : p_{ij} \geq 0 \text{ and } \sum_{j=1}^{k} p_{ij} = 1 \right\} \quad \text{for all } i = 1, 2, \ldots, k.
\]
Thus, we get marginal density of $(n_{i1}, n_{i2}, \ldots, n_{ik})$, for all $i = 1, 2, \ldots, k$ is
\[
f(n_{i1}, n_{i2}, \ldots, n_{ik} | \alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik}) = \frac{n_{i1}!}{\prod_{j=1}^{k} n_{ij}!} \frac{\Gamma \left( \sum_{j=1}^{k} \alpha_{ij} \right)}{\prod_{j=1}^{k} \Gamma(\alpha_{ij})} \frac{\prod_{j=1}^{k} \Gamma(n_{ij} + \alpha_{ij})}{\Gamma \left( \sum_{j=1}^{k} (n_{ij} + \alpha_{ij}) \right)},
\]
which is the Multinomial–Dirichlet distribution.

By $(n_{i1}, n_{i2}, \ldots, n_{ik})$, we denote an observed random sample from the Multinomial–Dirichlet distribution with parameters $(\alpha_{i1}, \ldots, \alpha_{ik})$ for all $i = 1, 2, \ldots, k$. Then, for incomplete-data log likelihood is of the form
\[
\log L(\alpha) = \sum_{i=1}^{k} \log L(\alpha_{i}) = \sum_{i=1}^{k} \log(n_{i1}) - \sum_{i=1}^{k} \log(n_{i})! + \sum_{i=1}^{k} \log \Gamma(\alpha_{i})
\]
\[
- \sum_{i=1}^{k} \sum_{j=1}^{k} \log \Gamma(\alpha_{ij}) + \sum_{i=1}^{k} \sum_{j=1}^{k} \log \Gamma(n_{ij} + \alpha_{ij}) - \sum_{i=1}^{k} \log \Gamma(n_{i} + \alpha_{i}),
\]
where $n_{i} = \sum_{j=1}^{k} n_{ij}$, and $\alpha_{i} = \sum_{j=1}^{k} \alpha_{ij}$, for all $i = 1, 2, \ldots, k$.

Because of the conditional structure of the complete-data model specified by Equations (5) and (6), the complete-data likelihood can be factored into product of the conditional density
of \((n_{i1}, n_{i2}, \ldots, n_{ik})\) given \((p_{i1}, p_{i2}, \ldots, p_{ik})\) and the marginal density of \((p_{i1}, p_{i2}, \ldots, p_{ik})\) for all \(i = 1, 2, \ldots, k\). Accordingly, the complete-data log likelihood be written as

\[
\log L_c(\alpha) = \sum_{i=1}^{k} \log L_c(\alpha_i) = \sum_{i=1}^{k} \log(n_{i1}) - \sum_{i=1}^{k} \sum_{j=1}^{k} \log(n_{ij}) + \sum_{i=1}^{k} \log \Gamma(\alpha_i)
\]

\[
- \sum_{i=1}^{k} \sum_{j=1}^{k} \log \Gamma(\alpha_{ij}) + \sum_{i=1}^{k} \sum_{j=1}^{k} (n_{ij} + \alpha_{ij} - 1) \log p_{ij}.
\]

The EM algorithm approaches the problem of solving the incomplete-data likelihood equation (8) indirectly by proceeding iteratively in terms of the complete-data log likelihood function \(\log L_c(\alpha)\). The obstacle due to unobservability is overcome by averaging the complete-data likelihood over its conditional distribution given the observed data \(n\). But in order to calculate this conditional expectation, we have to specify a value for \(\alpha\).

Let \(\alpha_i^{(0)}\) denote the starting value of \(\alpha_i\) and \(\alpha_i^{(r)}\), the value of \(\alpha_i\) on the \(r\)th subsequent iteration of the EM algorithm.

**E-step:** Then, on the first iteration of the EM algorithm, the E-step requires the computation of the conditional expectation of \(\log L_c(\alpha)\) given \(n\), using \(\alpha_i^{(0)}\) for \(\alpha_i\), which can be written as

\[
Q(\alpha_i, \alpha_i^{(0)}) = E_{\alpha_i^{(0)}}[\log L_c(\alpha_i)|n_i],
\]

where Q-function is used to denote the conditional expectation of the complete-data log likelihood function, \(\log L_c(\alpha)\), given the observed data \(n\), using the current fit for \(\alpha\). We have on the \((r + 1)\)th iteration of the E-step

\[
Q(\alpha_i, \alpha_i^{(r)}) = E_{\alpha_i^{(r)}}[\log L_c(\alpha_i)|n_i],
\]

where the expectation operator \(E\) has the subscript \(\alpha_i^{(r)}\) to explicitly convey that this (conditional) expectation is being affected using \(\alpha_i^{(r)}\) for \(\alpha_i\). Therefore, \(Q(\alpha_i, \alpha_i^{(r)})\) can be written as

\[
Q(\alpha_i, \alpha_i^{(r)}) = \log(n_{i1}) - \sum_{j=1}^{k} \log(n_{ij})! + \log \Gamma(\alpha_i) - \sum_{j=1}^{k} \log \Gamma(\alpha_{ij})
\]

\[
+ \sum_{j=1}^{k} (n_{ij} + \alpha_{ij} - 1)E_{\alpha_i^{(r)}}[\log p_{ij}|n_{i1}, \ldots, n_{ik}].
\]

It is seen from Equation (9) that, in order to carry out M-step, we need to calculate the term \(E_{\alpha_i^{(r)}}[\log p_{ij}|n_{i1}, \ldots, n_{ik}]\).

The calculation of the above term can be avoided if one makes use of the identity \([4]\),

\[
S_i(n; \alpha_i^{(r)}) = \left[ \frac{\delta Q(\alpha_i, \alpha_i^{(r)})}{\delta \alpha_i} \right]_{\alpha_i = \alpha_i^{(r)}} \text{ for all } i = 1, 2, \ldots, k,
\]

where \(S_i(n; \alpha_i^{(r)})\) are the score statistics given by

\[
S_i(n; \alpha_i^{(r)}) = \delta \log L(\alpha_i^{(r)}) \text{ for all } i = 1, 2, \ldots, k.
\]
On evaluating $S_{ij}(n; \alpha^{(r)}_i)$, the derivative of Equation (8) with respect to $\alpha_{ij}$ at the point $\alpha_i = \alpha_i^{(r)}$, we have that

$$S_{ij}(n; \alpha^{(r)}_i) = \frac{\delta \log L(\alpha^{(r)}_i)}{\delta \alpha_{ij}^{(r)}} = \frac{\delta \log \Gamma(\alpha_{ij}^{(r)})}{\delta \alpha_{ij}^{(r)}}$$

$$+ \delta \log \frac{\Gamma(n_{ij} + \alpha_{ij}^{(r)})}{\delta \alpha_{ij}^{(r)}} - \delta \log \frac{\Gamma(n_{i} + \alpha_i^{(r)})}{\delta \alpha_{ij}^{(r)}}$$

for all $i, j = 1, 2, \ldots, k$. (10)

On equating $S_{ij}(n; \alpha^{(r)}_i)$ equal to the derivative of Equation (9) with respect to $\alpha_{ij}$ at the point $\alpha_i = \alpha_i^{(r)}$, we obtain

$$E_{\alpha_i^{(r)}} \log(p_{ij}|n_1, \ldots, n_k) = \delta \log \frac{\Gamma(n_{ij} + \alpha_{ij}^{(r)})}{\delta \alpha_{ij}^{(r)}} - \delta \log \frac{\Gamma(n_{i} + \alpha_i^{(r)})}{\delta \alpha_{ij}^{(r)}}$$

$$= \psi(n_{ij} + \alpha_{ij}^{(r)}) - \psi(n_{i} + \alpha_i^{(r)})$$

where

$$\psi(s) = \frac{\delta \log \Gamma(s)}{\delta s}$$

is the digamma function of $s$.

Therefore, from Equation (9) we have

$$Q(\alpha_i, \alpha_i^{(r)}) = \log(n_{i+1}) - \sum_{j=1}^{k} \log(n_{ij}) + \log(\Gamma(\alpha_i) - \log(\Gamma(\alpha_{ij}))$$

$$+ \sum_{j=1}^{k} (n_{ij} + \alpha_{ij} - 1) \left[ \delta \log \frac{\Gamma(n_{ij} + \alpha_{ij}^{(r)})}{\delta \alpha_{ij}^{(r)}} - \delta \log \frac{\Gamma(n_{i} + \alpha_i^{(r)})}{\delta \alpha_{ij}^{(r)}} \right].$$ (11)

M-step: On the M-step at the $(r+1)$th iteration of EM algorithm is to maximise $Q(\alpha_i, \alpha_i^{(r)})$, i.e. $\alpha_i^{(r+1)}$ are obtained after solving

$$\frac{\delta Q(\alpha_i, \alpha_i^{(r)})}{\delta \alpha_{ij}} = 0 \quad \text{for all } i = 1, \ldots, k$$

It follows that $\alpha_i^{(r+1)}$ is a solution of the equation

$$\psi(\alpha_i) - \psi(\alpha_{ij}) + \psi(n_{ij} + \alpha_{ij}^{(r)}) - \psi(n_{i} + \alpha_i^{(r)}) = 0.$$

The E-step and M-step are alternated repeatedly until the difference $L(\alpha^{(r+1)}) - L(\alpha^{(r)})$ changes by an arbitrary small amount in the case of convergence of the sequence of likelihood values $\{L(\alpha^{(r)})\}$. 
Finally, we get
\[ \alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_k^*)^T, \]
where \( \alpha_i^* \) the limiting value of \( \alpha_i \), for all \( i = 1, 2, \ldots, k \). Thus, we have the empirical Bayes estimator of \( p_{ij} \) as
\[ \delta^*(n_{ij}) = \frac{n_{ij} + \alpha_{ij}^*}{n_i + \alpha_i^*} \quad \text{for all } i, j = 1, 2, \ldots, k, \]
where \( \alpha_{ij}^* \) is the limiting value of \( \alpha_{ij} \) for all \( i, j = 1, \ldots, k \), and
\[ \alpha_i^* = \sum_{j=1}^k \alpha_{ij}^* \quad \text{for all } i = 1, \ldots, k. \]

Details of such are included in Appendix 1 and its inspiration is in [1].

Concerning now the M-step, it can be seen that the presence of the terms like \( \log \Gamma(\alpha_{ij}) \) in Equation (11) prevents a closed-form solution for \( \alpha^{(r+1)}_i \).

Now, using [4], the quasi-Newton acceleration procedure in our case becomes
\[ \alpha_i^{(r+1)} = \alpha_i^{(r)} + [I_c(\alpha_i^{(r)}; n_i) + B_i^{(r)}]^{-1} S_i(n_i; \alpha_i^{(r)}) \quad \text{for all } i = 1, \ldots, k. \] (12)

The components of the score statistic \( S_i(n_i; \alpha_i^{(r)}) \) are available from Equation (10), and the information matrix \( I_c(\alpha_i^{(r)}; n_i) \) is given by
\[ I_c(\alpha_i^{(r)}; n_i) = - \frac{\delta^2 Q(\alpha_i^{(r)}; \alpha_i^{(r)})}{\delta \alpha_i \delta \alpha_i^T} \bigg|_{\alpha_i = \alpha_i^{(r)}} = D_i - C_i 11^T \quad \text{for all } i = 1, 2, \ldots, k, \] (13)
where each \( D_i \) is the diagonal matrix with the \( j \)-th diagonal entry
\[ d_{ij} = \psi'(\alpha_{ij}^{(r)}) - \psi'(n_{ij} + \alpha_{ij}^{(r)}) \quad \text{for all } j = 1, 2, \ldots, k, \]
\( C_i \) is the constant \( \psi'(\alpha_{ij}^{(r)}) - \psi'(n_{ij} + \alpha_{ij}^{(r)}) \), and \( 1 \) is a column vector of all 1’s.

where \( \psi'(s) \) is the trigamma function \( \delta^2 \log \Gamma(s)/\delta s^2 \). As trigamma function is decreasing, we have \( d_{ij} > 0 \) when \( n_{ij} > 0 \). For the same reason, \( C_i > 0 \). Since the presentation (13) is preserved under finite sums, it holds, in fact, for the entire sample.

The observed information matrix (13) is the sum of a diagonal matrix, which is trivial to invert, plus a symmetric, rank-one perturbation.

5. Robustness of empirical Bayes estimate

Empirical Bayes estimation, falls outside of the formal Bayesian paradigm. However, it has been proved to be an effective technique of constructing estimators that perform well under both Bayesian and frequentist criteria. One reason for this, as we will see, is that empirical Bayes estimators tend to be more robust against misspecification of the prior distribution.

Bayes risk performance of the empirical Bayes estimator is often robust; i.e. its Bayes risk is reasonably close to that of the Bayes estimator, no matter what values the hyperparameters attain.
We have Bayes risk for unbiased estimator \((n_{ij}/n_i, n_{i2}/n_i, \ldots, n_{ik}/n_i)\), for the \(i\)th state, for \(i = 1, 2, \ldots, k\)

\[
\begin{align*}
    r(\pi, n_i|n_i) &= E_{\pi_i} \left[ \sum_{j=1}^{k} \frac{p_{ij} (1 - p_{ij})}{n_i} \right] = \frac{1}{n_i} E_{\pi_i} \left[ \sum_{j=1}^{k} p_{ij} (1 - p_{ij}) \right] \\
    &= \frac{1}{n_i} E_{\pi_i} \left[ \sum_{j=1}^{k} p_{ij} - \sum_{j=1}^{k} p_{ij}^2 \right] = \frac{1}{n_i} \left[ \sum_{j=1}^{k} \alpha_{ij} - \sum_{j=1}^{k} \frac{\alpha_{ij} (\alpha_{ij} + 1)}{\alpha_i (\alpha_i + 1)} \right] \\
    &= \frac{\alpha_i^2}{n_i} - \sum_{j=1}^{k} \frac{\alpha_{ij}^2}{\alpha_i (\alpha_i + 1)} 
\end{align*}
\]

We have risk function for Bayes estimator for the \(i\)th state

\[
\begin{align*}
    R(\delta, \pi_i) &= E_{n_i|n_i} \left[ \sum_{j=1}^{k} \left( \delta(n_{ij}) - p_{ij} \right)^2 \right] = \sum_{j=1}^{k} \left[ \text{Bias}(\delta(n_{ij})) \right]^2 + \sum_{j=1}^{k} \text{Var}[\delta(n_{ij})] \\
    &= \sum_{j=1}^{k} \left[ p_{ij} - E \left( \frac{\alpha_{ij} + n_{ij}}{\alpha_i + n_i} \right) \right]^2 + \sum_{j=1}^{k} \text{Var} \left[ \frac{\alpha_{ij} + n_{ij}}{\alpha_i + n_i} \right] \\
    &= \sum_{j=1}^{k} \left[ p_{ij} - \left( \frac{\alpha_{ij} + n_{ij}}{\alpha_i + n_i} \right) \right]^2 + \sum_{j=1}^{k} \left[ \frac{n_{ij} p_{ij} (1 - p_{ij})}{(\alpha_i + n_i)^2} \right] \\
    &= \frac{\alpha_i^2 - n_i}{(\alpha_i + n_i)^2} \sum_{j=1}^{k} p_{ij}^2 + \sum_{j=1}^{k} \alpha_{ij}^2 + \sum_{j=1}^{k} \left( n_i - 2 \alpha_i \alpha_{ij} \right) p_{ij} \quad \text{for all } i = 1, 2, \ldots, k
\end{align*}
\]

Bayes risk for Bayes estimator

\[
\begin{align*}
    r(\pi, \delta) &= E[R(\delta, \pi_i)] = E_{\pi_i} E_{n_i|n_i} \left[ \sum_{j=1}^{k} \left( \delta(n_{ij}) - p_{ij} \right)^2 \right] \\
    &= E_{\pi_i} \left[ \frac{\alpha_i^2 - n_i}{(\alpha_i + n_i)^2} \sum_{j=1}^{k} p_{ij}^2 + \sum_{j=1}^{k} \alpha_{ij}^2 + \sum_{j=1}^{k} \left( n_i - 2 \alpha_i \alpha_{ij} \right) p_{ij} \right] \\
    &= \frac{1}{(\alpha_i + n_i)^2} \left[ \sum_{j=1}^{k} \alpha_{ij} (\alpha_{ij} + 1) \right] + \sum_{j=1}^{k} \alpha_{ij}^2 + \sum_{j=1}^{k} \left( n_i - 2 \alpha_i \alpha_{ij} \right) \frac{\alpha_{ij}}{\alpha_i} \\
    &= \frac{1}{(\alpha_i + n_i)^2} \left[ \sum_{j=1}^{k} \alpha_{ij}^2 + \alpha_i \right] + \sum_{j=1}^{k} \alpha_{ij}^2 + n_i - 2 \sum_{j=1}^{k} \alpha_{ij}^2 \\
    &= \frac{\alpha_i^2}{(\alpha_i + n_i) \alpha_i (\alpha_i + 1)} \quad \text{for all } i = 1, 2, \ldots, k.
\end{align*}
\]
Table 1. Bayes risk of Bayes, empirical Bayes and unbiased estimator, for \( k = 3 \), and sample size = 30.

| Prior parameters | Bayes \((\delta^\pi)\) | EB \((\delta^\hat{\pi})\) | Unbiased \((n_{ij}/n_i)\) |
|------------------|-------------------------|--------------------------|--------------------------|
| \((1,1,1)\)      | 0.0151515               | 0.0157934                | 0.0166667                |
| \((1,5,5)\)      | 0.0129342               | 0.0131800                | 0.0176767                |
| \((5,5,10)\)     | 0.01190476              | 0.0119227                | 0.019841                 |
| \((1,5,10)\)     | 0.0106178               | 0.0119566                | 0.013095                 |
| \((1,10,10)\)    | 0.0101860               | 0.0106867                | 0.0159314                |

Table 2. Place of birth by place of enumeration, for Gabon: 1960–1961.

| Place of birth | Woleu N’Tem | Estuaire | Ogooue Maritime | Moyen Ogooue | Ogooue Ivindo | Total Gabon |
|----------------|-------------|----------|----------------|--------------|--------------|-------------|
| Woleu N’Tem    | 74,609      | 526      | 21             | 52           | 562          | 75,770      |
| Estuaire       | 4899        | 1180     | 825            | 1288         | 2070         | 41,770      |
| Ogooue Maritime| 191         | 32,688   | 22,612         | 2623         | 118          | 26,724      |
| Moyen Ogooue   | 121         | 207      | 918            | 24,003       | 659          | 25,908      |
| Ogooue Ivindo  | 113         | 79       | 11             | 124          | 34,160       | 34,487      |
| Total Gabon    | 79,933      | 34,680   | 24,387         | 28,090       | 37,569       | 204,659     |

Table 3. Empirical estimate of Dirichlet parameters.

|                | 48.9648099 | 0.765565 | 0.2469767 | 0.3116376 | 0.7932113 |
|----------------|------------|----------|-----------|-----------|-----------|
| 2.5114255      | 12.599406  | 0.7617918| 0.9788393 | 1.3249833 |
| 0.4603527      | 1.148029   | 12.4422854| 2.0176359 | 0.3874388 |
| 0.3798190      | 0.459585   | 0.9340568| 12.1600461| 0.7760039 |
| 0.5219156      | 0.452650   | 0.2509382| 0.5428788 | 46.3399134 |

Bayes risk for empirical Bayes estimator

\[
\hat{r}(\hat{\pi}, \delta) = E_{\alpha_i = \hat{\alpha}_i} [R(\delta, p_i)] = \frac{\hat{\alpha}_i^2 - \sum_{j=1}^{k} \hat{\alpha}_j^2}{(\hat{\alpha}_i + n_i)\hat{\alpha}_i (\hat{\alpha}_i + 1)} \quad \text{for all } i = 1, 2, \ldots, k.
\]

The Bayes risk of empirical Bayes estimator is only slightly higher than that of the Bayes estimator and is given in Table 1. For this simulation works are done. The important idea in this area is Monte Carlo Simulation [5]. For comparison, we also include the Bayes risk of the unbiased estimator \( n_{ij}/n_i \), where \( n_i = \sum_{j=1}^{k} n_{ij} \) for a particular \( i \). The risk of the empirical Bayes estimator is between that of the Bayes estimator and that of unbiased estimator.

The R-codes to perform these calculations are given in Appendix 2.

6. Working with real data set

Table 2 corresponds to the data of the migration of population of five different states of Gabon from 1960 to 1961 [9]. Here, the States are Woleu NTem, Estuaire, Ogooue Maritime, Mayen Ogooue and Ogooue Ivindo [Source: Partial data in tabular form from Gabon, Service de statistique, and France, I.N.S.E.E., Recensement et enquete demographiques, 1960–1961 (Population
Table 4. Empirical Bayes estimate of transition probability matrix.

|        | 0.984659710 | 0.006947481 | 0.0002802252 | 0.0006899353 | 0.007422648 |
|--------|-------------|-------------|--------------|--------------|-------------|
| 0.117294217 | 0.782532338 | 0.0197606563 | 0.0308455393 | 0.049567250 |
| 0.007159951 | 0.044170826 | 0.8460754186 | 0.0981665257 | 0.004427278 |
| 0.004682374 | 0.008003005 | 0.0354489972 | 0.9264139629 | 0.025451661 |
| 0.003287145 | 0.002300634 | 0.0003257826 | 0.0036062687 | 0.990480169 |

Census and survey, 1960–1961, Paris, p. 116. From the data it is reasonable to consider the transition probabilities according to case 2.

The estimates of Dirichlet parameters and those of transition probability matrix are given in Tables 3 and 4. The R-codes to perform these calculations are given in Appendix 3.

7. Concluding remarks

In this paper, robustness (in Section 5) of EB has been demonstrated, but this should be done elaborately as computations of risk have been done for some hyperparameters. More importantly, we have worked out in Sections 2, 3 and 5 for a class of Dirichlet prior which are reasonable under two cases where hyperparameters are the same for all states and different for all states. But, from other beliefs also, we are to keep this Dirichlet prior, but some other choices which are not like our two cases may be important. There, perhaps computations will have to be modified from this spirit.

Note

1. Supplementary Content may be viewed online at http://dx.doi.org/10.1080/02664763.2014.963525

Acknowledgments

The authors thank the Editor, Prof. Robert G. Aykroyd and the anonymous referees for their constructive comments which were helpful to improve the final version of the paper.

References

[1] W.C. Davidon, Variable metric methods for minimisation, AEC Research and Development Report ANL-5990, Argonne National Laboratory, Chicago, 1959.
[2] N.L. Johnson, S. Kotz, Distributions in Statistics: Discrete Distributions, John Wiley & Sons, New York, 1969.
[3] N.L. Johnson, S. Kotz, N. Balakrishnan, Continuous Univariate Distributions, Vol. 2, John Wiley & Sons, New York, 1994.
[4] K. Lange, A quasi-Newton acceleration of the EM algorithm, Statist. Sin. 5 (1995), pp. 1–18.
[5] G.O. Roberts, J.S. Rosenthal, General state space Markov chains and MCMC algorithms, Probab. Surv. 1 (2004), pp. 20–71 (electronic).
[6] B. Seal, S.J. Hossain, Testing of hypotheses: LRT test of some important hypothesis for internal migration model, Slovak Stat. Demography – Sci. J. 23(2) (2013), pp. 35–47.
[7] B. Seal, S.J. Hossain, Measure of state’s importance and its estimation from migration data, ProbStat Forum, 6 (2013), pp. 10–17.
[8] B. Seal, S.J. Hossain, Bayes and Minimax Estimation of Parameters of Markov Transition Matrix, ProbStat Forum, 6 (2013), pp. 107–115.
[9] H.S. Shryock, S. Siegel, and Associates, The Methods and Materials of Demography, Academic Press, New York, 1976.
[10] United Nations Methods of Measuring Internal Migration, Manuals on Methods of Estimating Population, Manual VI, 1970.