INDIAN JOURNAL OF POPULATION AND DEVELOPMENT
Volume 1(2), July 2021: 285-294

An EM Algorithm Approach to Estimate Parameters of Fluctuating Nature of PPA

Pragya Singh
Kaushalendra K Singh
Brijesh P Singh

Abstract

The Aim of this paper is to use the EM algorithm to estimate parameters of the distribution of the duration of postpartum amenorrhea (PPA). A mixture of the Gumbel distribution has been used to fit the bimodal distribution the duration of postpartum amenorrhea in four states of India - West Bengal, Gujarat, Himachal Pradesh, and Andhra Pradesh. The appropriateness of the model and the estimation technique is examined using data from National Family Health Survey-4 (2015-16) of India.

Introduction

The major determinant that approximates the natural fertility condition is the breastfeeding behaviour. Breastfeeding lengthens the interval between successive births by delaying the resumption of ovulation. Intensive and sustained breastfeeding can result in years of postpartum amenorrhea (PPA). In societies where intensive breastfeeding is the norm, couples tend to have longer intervals between successive births and lower completed fertility (Howie and McNeilly, 1982; Wood, 1994). The length of the duration of postpartum amenorrhoea (PPA), therefore, has been subject of intensive research in the context of fertility transition. The PPA is defined as the period between the termination of the pregnancy and the resumption of the menstruation. Breastfeeding and PPA are closely related. Breast milk production is dependent on the peptide hormone prolactin, which also has a role in inhibiting ovulation after delivery. Prolactin levels depend, in turn, on suckling stimulus. It has been established that more frequent suckling delays the resumption of ovulation and is, therefore, a key factor in variations in the duration of postpartum amenorrhoea between women and between populations (Konner and Worthman, 1980). Different studies have shown the relationship between duration of breastfeeding and duration of PPA (Jones 1989; 1990; Singh and Singh, 1989; Singh et al, 1990; Nath et al, 1993; Nath et al, 1994; Mukherjee et al, 1994; Singh et al, 1999). There are also studies that explain the mechanism through which breastfeeding delays the return of the menstrual cycle after the termination of pregnancy (Lunn et al, 1984).

There are many studies that suggest that the distribution of the duration of postpartum amenorrhoea – the duration between the termination of the pregnancy and the resumption of the menstruation - is bimodal (Ford and Kim, 1987; Huffman et al, 1987; Potter and Kobrin, 1981) which suggests that there are distinct subgroups of women with short and long duration of amenorrhoea. It is argued that the duration of amenorrhoea is short in situations where breastfeeding is absent because of pregnancy loss or infant death, or confusion of postpartum bleeding with resumption of menses (Holman et al, 2006). There are many studies that have attempted to model the distribution of the duration of PPA (Barrette, 1969; Lesthaeghe and Page, 1980; Potter and Kobrin, 1981). Barrette (1969) has used the modified Pascal distribution; Lesthaeghe and Page (1980) have used the logit model; and Potter and Kobrin have used the mixed geometric negative binomial distribution. Ford and Kim (1987), on the other hand, have used the mixture of two extreme value distributions to model the duration of PPA in the presence of censored cases. A multi-centre study conducted by the World Health Organization has found that the distribution of the duration of PPA is bimodal in which one subgroup had a mean time 3-4 months to the resumption of menses after delivery while the other subgroup had a mean time of around 9 months to the resumption of menses (Le Strat and Thalabard, 2001).

In India, like in many other countries, data about the duration of PPA are available from household surveys such as the National Family Health Survey. However, the quality of data obtained from these household surveys is regarded as poor as women do not exactly remember the correct time of the return of menses after the termination of pregnancy which results in gross heaping in the distribution. In order to address the problems of data quality associated with the data on the duration of PPA available from the household surveys, different models have been developed to study the distribution of the duration of PPA including finite mixture probability models. In the present study, we have used the mixture Gumbel distribution (Jhonson and Kotz, 1970) to model the distribution of the duration of PPA. In previous studies, non-linear maximisation, or minimisation (NLM) techniques have been used to model the distribution of the duration of PPA. These techniques are based on an assumed probabilistic distribution. In this paper, we attempt to fit a model to the distribution of the duration of PPA using the expectation-maximisation (EM) technique which is particularly suited to deal with the problem of missing data. The EM technique provides an iterative solution to obtain the density estimate of a dataset by searching across different probability distributions and their parameters.

Mixture Model and Estimation

The model used in the present paper is based on the extreme value distribution. The distribution function of the model is given by:

$$F(x) = \exp\left(-\exp\left(-\frac{(x-\alpha)}{\beta}\right)\right), \alpha, \beta > 0 \qquad (1)$$

and the density function is given by:

$$f(x;\alpha,\beta)=\frac{1}{\beta}\exp\left(-\frac{x-\alpha}{\beta}-\exp\left(-\frac{(x-\alpha)}{\beta}\right)\right) \tag{2}$$

The model has two parameters α and β that need to be estimated. The parameter α is the location parameter while the parameter β is the scale or dispersion parameter. The model contains a mixing parameter *a* which lies between 0 and 1.

We write the mixture model as:

$$f(x)=af_1(x)+(1-a)f_2(x) \tag{3}$$

for simplifying the calculations, we put $a=w_1$ and $1\text{-}a=w_2$ so that w_1 and w_2 are the mixture weights of (3). In equation (3) the first extreme value distribution is denoted by f_1(x) and the second is denoted as f_2(x). The required density function is given by

$$f_m(x)=\frac{a}{\beta_1}\exp\left(-\frac{(x-\alpha_1)}{\beta_1}-\exp\left(-\frac{(x-\alpha_1)}{\beta_1}\right)\right)+\frac{((1-a))}{\beta_2}\exp\left(-\frac{(x-\alpha_2)}{\beta_2}-exp\left(-\frac{(x-\alpha_2)}{\beta_2}\right)\right) \tag{4}$$

The mixture distribution can be written as

$$F_m(x)=a\exp\left(-\exp\left(-\frac{x-\alpha_1}{\beta_1}\right)\right)+(1-a)\exp\left(-\exp\left(-\frac{x-\alpha_2}{\beta_2}\right)\right) \tag{5}$$

The parameter *a* of equation (4) is the proportion of women having short duration of PPA (less than 6 months) so that 1-*a* is the proportion of women having long duration of PPA. The other parameters (α_1,β_1), (α_2,β_2) denote the mean and variance of the distribution of short duration PPA and long duration PPA, respectively. These parameters can be estimated using the maximum likelihood estimation methods if the value of *a* is known. When the parameter *a* is unknown, the parameters are estimated by maximising the likelihood through an iterative procedure by using the EM algorithm which was developed for situation when the data are incomplete or missing (Dempster et al, 1997). The basic idea behind the EM procedure is to relate the missing data problem with a complete data problem for which maximising the likelihood is the simplest way. The EM algorithm is widely used in the estimation of mixture models (Meng and Pedlow, 1992). The general formulation of the algorithm is similar to the one proposed by Hindenes (2017), McLachlan and Krishnan (1996) and Otiniano (2017).

Let us consider a random sample $x=x_1,\ldots,x_n$ from the observed variable X, with the distribution function $f_x(x;\theta)$. Here $\theta=(\theta_1,\ldots,\theta_n)\in\Omega$ are the parameters to be estimated and Ω denotes the parameter space. Further assume that there is some unobservable data y, with random variable Y, such that z=(x,y) denotes the complete data. Let $f_{xy}(x;y;\theta)$ is the distribution of the complete data. The estimation task, then, is to maximise the likelihood associated with the complete data, $L_c(z;\theta)$ or maximising the log likelihood of the complete data $l_c=\log L_c(z;\theta)$. However, the log likelihood of the complete dataset is unknown. Therefore, the expectation of the complete data may be obtained from the observed data. The $\theta^{(k)}$, are the current parameter estimates that we have used to evaluate the expectation and θ and new parameters that we optimize to increase Q. Let

$$\left(\theta, \theta^{(k)}\right) = E_{\theta^{(k)}}[l_c(z;\theta)|x] \tag{6}$$

denotes the expectation of the observed data which is incomplete. Here $E_\theta(k)$ denotes the expectation of the observed data with parameters $\theta^{(k)}$. There are two steps in each iteration of the algorithm. The first step is the expectation (E) step, and the second step is the maximisation (M) step. For each iteration, we first compute the expected likelihood of the complete dataset using equation (6) and then maximise the expected likelihood such that

$$Q\left(\theta^{k+1};\theta^{k}\right) \geq Q\left(\theta;\theta^{k}\right)\forall\theta\epsilon\,\Omega \tag{7}$$

The two steps are repeated until the convergence is achieved.

In the present case, let us consider mixture of two Gumbel distributions,

$$f_X(x;\theta) = \sum_{j=1}^{2}(w_j\ \alpha_j\beta_j), w_1 + w_2 = 1 \tag{8}$$

Here θ=(w_1, w_2, α_1, β_1, α_2, β_2); w_1,w_2>0 and x=(x_1, …., x_n) is a vector of observed duration of PPA so that fitting of the model can be formulated as an incomplete data problem. Let Y is a latent variable such that y= (y_1, …., y_n) denotes the missing data vector. Here, the variable y_i is treated as a two-dimensional indicator variable with first and second variable equal to either 1 or 0 if the observation $x_i$either did or did not arise from the first and second mixture component, respectively.

$y_{ij} = 1$, if x_i belongs to the jth component,

$y_{ij} = 0$, if x_i does not belong to the jth component.

We can write the log likelihood of the complete dataset, z=(x,y) as

$$l_c(\theta, y) = \sum_{i=1}^{n}\log f_{x,y}(x_i, y_i\,;\theta) = \sum_{i=1}^{n}\log(f_y(y_i\,;\theta) f_{x|y}(x_i|y_i;\theta)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{2}\log(w_j\ f_j(x_j;\alpha_j,\beta_j))^{y_{ij}} = \sum_{i=1}^{n}\sum_{j=1}^{2} y_{ij}\log\left(\ w_j\ f_j(x_j;\alpha_j,\beta_j)\right) \tag{9}$$

As y is unknown, the log likelihood of the complete dataset cannot be computed. Instead, the EM algorithm considers the conditional expectation of $l_c(\theta,y)$ given the complete dataset and the values of current parameter $\theta^{(k)}$ where Y is now treated as a random variable. The conditional expectation is given by

$$Q(\theta, \theta^{(k)}) = E_{\theta^k}[l_c(z;\theta)\,|\,x,\theta^{(k)}]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{2} P(y_{ij} = 1|y_i, \theta^{(k)})\log w_j f_j\left(x_i;\alpha_j,\beta_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{2} h_{ij}^{k}\log w_j\ f_j(x_j;\alpha_j,\beta_j) \tag{10}$$

here $h_{ij}^{k}$is defined as the probability that $x_i$belongs to component j, given the current estimates. That is h_{ij}^{k}= $P(y_{ij} = 1|y_i, \theta^{(k)})$. We need to compute h_{ij}^{k} in the E-step in order to obtain the expected complete log-likelihood while in the M-step we maximize it with respect to θ.

Since$f_j(x_j;\alpha_j,\beta_j)$is the Gumbel distribution function given by equation (2), we have

$$Q(\theta,\theta^{(k)}) = \sum_{i=1}^{n}\sum_{j=1}^{2} h_{ij}^{(k)} \log\left[\frac{w_j}{\beta_j}\exp\left[-\left(\frac{x_i-\alpha_j}{\beta_j}+\exp\left(-\frac{x_i-\alpha_j}{\beta_j}\right)\right)\right]\right] \qquad (11)$$

In order to maximise this with respect to θ, we can write

$$\arg\max Q(\theta,\theta^{(k)}) = \arg\max \sum_{i=1}^{n}\sum_{j=1}^{2} h_{ij}^{(k)}\left[\log\frac{w_j}{\beta_j}-\left(\frac{x_i-\alpha_j}{\beta_j}+\exp\left(-\frac{x_i-\alpha_j}{\beta_j}\right)\right)\right] \qquad (12)$$

The Lagrangian function for the above equation will be

$$L = \sum_{i=1}^{n}\sum_{j=1}^{2} h_{ij}^{(k)}\left[\frac{x_i-\alpha_j}{\beta_j}+\exp\left(-\frac{x_i-\alpha_j}{\beta_j}\right)+\log w_j-\log\beta_j\right]-\lambda\left(\sum_{j=1}^{2} w_j-1\right), \qquad (13)$$

here λ is the Lagrange multiplier. By computing the partial derivatives of L with respect to parameters α_j, β_j, w_j and λ, expressions for parameters are obtained which optimise the expected value of log-likelihood for the complete dataset.

Duration of PPA in Selected States of India

We have applied the above model to examine the pattern of the duration of PPA in selected states of India based on the data available through the National Family Health Survey (NFHS-4) 2015-2016 (Government of India, 2017). The NFHS-4 provides data about the duration of PPA in completed months only. However, we have considered the duration of PPA as a continuous variable because a continuous variable can be mathematically treated in an easier way than a discrete variable. Because of the poor quality of the data about the duration of PPA available from the National Family Health Survey, it is quite difficult to model the distribution of the duration of PPA. This problem can be addressed to some extent using the approach outlined in this paper. We restrict the analysis to bimodal pattern of the distribution of the duration of PPA since the model incorporates only the bimodal pattern of distribution.

The parameters of the model have been obtained by solving the EM algorithm described above. The model fitting is based on the information about the duration of PPA after the last-but-one birth as reported by ever-married women in the age group 15-49. The analysis has been carried out for four states: West Bengal, Gujarat, Himachal Pradesh, and Andhra Pradesh.

The estimates of the parameters of the model are given in table 1. The present model is only limited to incorporate up to two modes only. The goodness of fit of the model has been tested by Kolmogorov Smirnov (K-S) test statistic. Table 2 shows the observed and expected values of the mean and standard deviation of the observed distribution of the duration of PPA along with the mean and standard deviation of the distribution derived from the model. It is obvious from table 2 that based on the K-S test statistic, the proposed modelling approach provides good fit to the data on the duration of PPA in four states available from the National Family Health Survey 2015-16. It is also clear from table 1 that the distribution of the duration of PPA in all the four states is bimodal.

Table 1 :Estimated values of parameters of the model

States	Parameters				
	α_1	β_1	α_2	β_2	a
West Bengal	0.37	1.69	11.79	2.67	0.66
Gujarat	3.18	3.89	12.21	4.91	0.78
Himachal Pradesh	4.13	5.10	17.01	5.63	0.79
Andhra Pradesh	4.03	4.78	11.88	4.05	0.81

Source: Authors' calculations

Table 1 suggests that the distribution of the duration of PPA in the four states is different. In Gujarat, the first mode of the distribution of the duration of PPA is at around 3-4 months while the second mode is at around 12-13 months. The proportion of women with short duration PPA in the state is around 80 per cent (parameter *a*). The observed mean and median duration of PPA is about 7 months which means that 50 per cent of the women in the state resume their menses around 7 months after the termination of pregnancy. In Himachal Pradesh and Andhra Pradesh also, the first mode is at around 4 months but the second mode in Himachal Pradesh is at around 17 months whereas it is at around 12 months in Andhra Pradesh. The proportion of women with short duration of PPA is around 80 per cent in both states. The observed mean and median duration of PPA in Andhra Pradesh are 7 months and 6 months respectively but 8 months and 7 months in Himachal Pradesh.

Table 2: Observed and estimated values of parameters of the model

States	Observed values		Estimated values		'p'
	Mean	SD	Mean	SD	
West Bengal	5.59	2.18	5.42	2.72	0.988
Gujarat	7.15	5.17	7.53	5.19	0.883
Himachal Pradesh	8.53	6.01	9.8	6.68	0.897
Andhra Pradesh	7.00	3.99	7.26	4.17	0.961

Source: Authors' calculations

In West Bengal, the distribution of the duration of PPA is significantly different from the distribution of the duration of PPA in the other three states. Around two-third of the women in West Bengal start menstruating within one month of the termination of pregnancy so that the mean duration of PPA in the state is the lowest amongst the four states included in the analysis. Since the primary determinant of the duration of PPA is the duration and pattern of breasting including suckling frequency, it appears that in around two third women of the state, breastfeeding duration and patterns have virtually little impact on the duration of PPA. This also implies that, that the duration and the pattern of breastfeeding, in these women, has negligible impact on their fertility. At least half of the women in West Bengal start menstruating within six months of the termination of their pregnancy compared to more than 8 months in Himachal Pradesh and around 7 months in Andhra Pradesh and Gujarat. Table 1 and Table 2 suggest that the distribution of the duration of PPA in the four states is essentially different. There appear state-specific factors that influence the duration of PPA which has implications for fertility regulation.

Conclusions

The postpartum infecundity, primarily due to postpartum amenorrhoea, is one of the proximate determinants of fertility (Bongaarts, 1978). Information about the duration of postpartum amenorrhoea is usually collected through retrospective enquiries during household surveys and the quality of the information so collected is of poor quality to analyse the patterns of duration of postpartum amenorrhoea. In this paper, we have proposed a modelling approach based on the extreme value mixture distributions along with the use of EM algorithm that can be used to analyse the pattern of the duration of postpartum amenorrhoea even if the data quality is poor. The EM algorithm is one of the commonly used techniques for determining parameters of mixture models when there are missing values in the data.

Using the data available from the National Family Health Survey 2015-2016, the present analysis also suggests that the distribution of the duration of PPA in four states of India – West Bengal, Gujarat, Himachal Pradesh, and Andhra Pradesh – is different, although bimodal. The bimodal distribution of the duration of PPA can be attributed to numerous factors including the poor quality of data. There is extensive literature that suggests that breastfeeding is an important reason for the delay in ovulation after the termination of pregnancy (McNeilly, 1977; Billewicz, 1979; Habicht et al, Mishra et al, 2021) and discontinuation of breastfeeding is a crucial factor in the short duration of PPA as observed in four states. Breastfeeding, particularly, sucking exerts pressure which raises the level of the hormone prolactin and results in suppressing the ovarian activity and delay in the return of menses. It has been reported the longer the duration of breastfeeding the longer the period of the return of menstruation after the termination of pregnancy (Singh et al, 2012).

The nutritional status of woman also plays a key role in decided the length of PPA. It has been observed in a study based on the data from the National Family Health Survey 1998-1999 that the duration of PPA is significantly longer in under-nourished women as compared to the duration of PPA in well-nourished women (Dwivedi, 2010). Our analysis, however, suggests that, in the four states, a small proportion of woman have long duration of PPA. In almost 80 per cent women in three of the four states, the duration of PPA is short. It is only in West Bengal where around 33 per cent of the women are estimated to be having long duration of PPA according to the data available through NFHS-4. At the same time, around two-third women in West Bengal have very short duration PPA.

References

Barrett JC (1969) A monte carlo simulation of human reproduction. *Genus* 22: 1-22.

Billewicz WZ (1979) The timing of post-partum menstruation and breast feeding: a simple formula. *Journal of Biosocial Science* 11(2): 141-151.

Bongaarts J (1978) A framework for analyzing the proximate determinants of fertility. *Population and Development Review* 4(1): 105-132.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39(1): 1-22.

Dwivedi LK (2010) Impact of maternal nutritional status on lactational amenorrhoea in India: a regional analysis. Paper presented at the Annual Conference of Population Association of America.

Ford K, Kim Y (1987) Distributions of postpartum amenorrhea: some new evidence. *Demography* 24(3): 413-430.

Government of India (2017) *National Family Health Survey (NFHS-4), 2015-16*. Mumbai, International Institute for Population Sciences.

Habicht J-P, Davanzo J, Butz WP, Meyers L (1985) The contraceptive role of breastfeeding. *Population Studies* 39(2): 213-232.

Hindenes S (2017) Mixture models for flood frequency analysis - A case study for Norway (Master's thesis, NTNU).

Holman DJ, Grimes MA, Achterberg JT, Brindle E, O'Connor KA (2006) Distribution of postpartum amenorrhea in rural Bangladeshi women. *American Journal of Physical Anthropology* 129(4): 609-619.

Howie PW, McNeilly AS (1982) Effect of breast-feeding patterns on human birth intervals. *Reproduction* 65(2): 545-557.

Huffman SL, Chowdhury A, Chakraborty J, Simpson NK (1980) Breast-feeding patterns in rural Bangladesh. *The American Journal of Clinical Nutrition* 33(1): 144-154.

Jones RE (1989) Breast-feeding and post-partum amenorrhoea in Indonesia. *Journal of Biosocial Science* 21(1): 83-100.

Jones RE (1990) The effect of initiation of child supplementation on resumption of post-partum menstruation *.Journal of Biosocial Science* 22(2): 171.

Johnson NL, Kotz S (1970) *Continuous Multivariate Distributions*. New York, Wiley.

Konner M, Worthman C (1980) Nursing frequency, gonadal function and birth spacing among Kung hunter-gatherers. *Science* 207: 788-791.

Lesthaeghe RJ, Page HJ (1980) The post-partum non-susceptible period: development and application of model schedules. *Population Studies* 34(1): 143-169.

Le Strat Y, Thalabard JC (2001) Analysis of postpartum lactational amenorrhoea in relation to breast-feeding: some methodological and practical aspects. *Journal of Biosocial Science* 33(4): 529-549.

Lunn PG, Austin S, Prentice AM, Whitehead RG (1984) The effect of improved nutrition on plasma prolactin concentrations and postpartum infertility in lactating Gambian women. *The American Journal of Clinical Nutrition* 39(2): 227-235.

McNeilly AS (1977) Physiology of lactation. *Journal of Biosocial Science* 9(S4): 5-21.

McLachlan GJ, Peel D (1996) An algorithm for unsupervised learning via normal mixture models. In DL Dowe, KB Korb, JJ Oliver (Eds) *ISIS: Information, Statistics and Induction in Science*: 354-363.

Meng X-L, Pedlow S (1992) EM: a bibliographic review with missing articles. *Proc. Statist. Comput. Sect. Am. Statist. Ass.* 39: 1992.

Mishra R, Singh KK, Singh Brijesh P (2021) Duration of post-partum amenorrhoea: a model-based approach. *Indian Journal of Population and Development* 1(1): 41-50.

Mukherjee S, Singh KK, Bhattacharya BN (1991) Breastfeeding in eastern Uttar Pradesh, India :Differentials and determinants. *Janasamkhya* 9(1-2): 25-41.

Nath DC, Singh KK, Land KC, Talukdar PK (1993) Breastfeeding and postpartum amenorrhea in a traditional society: a hazards model analysis. *Social Biology* 40(1-2): 74-86.

Nath DC, Land KC, Singh KK (1994) The role of breast-feeding beyond postpartum amenorrhoea on the return of fertility in India: a life table and hazards model analysis. *Journal of Biosocial Science* 26(2): 191-206.

Otiniano CEG, Gonçalves CR, Dorea CCY (2017) Mixture of extreme-value distributions: identifiability and estimation. *Communications in Statistics - Theory and Methods* 46(13): 6528-6542.

Potter RG, Kobrin FE (1981) Distributions of amenorrhoea and anovulation. *Population Studies* 35(1): 85-99.

Ramachandran P (1987) Breast-feeding and fertility: sociocultural factors. *International Journal of Gynecology & Obstetrics* 25(1):

Singh KK, Suchindran CM, Singh K (1999) Breast-feeding and post-partum amenorrhoea: an Indian experience. *Demography India* 28: 1-12.

Singh NS, Singh NS, Narendra RK (2012) Postpartum amenorrhoea among Manipuri women: a survival analysis. *Journal of Health, Population, and Nutrition* 30(1): 93-98.

Singh SN, Singh KK (1989) Life table analysis of censored data on post-partum amenorrhoea period. *Demography India* 18: 27-38.

Singh SN, Singh KK, Singh K, Singh S (1990) Socio-economic development and transition in the duration of post-partum amenorrhoea: survival function analysis of data. *Janasamkhya* 8(1): 41-54.

Wood JW (1994) *Dynamics of Human Reproduction*. New York, Aldine.