Journal of Data Science **2**(2004), 213-230

Factor Effects Testing for Mixture Distributions with Application to the Study of Emergence of *Pontomyia Oceana*

Mong-Na Lo Huang, Chun-Sui Lin and Keryea Soong
National Sun Yat-sen University

Abstract: In this work, testing of factor effects to the observed data from finite mixture distributions are discussed. Likelihood ratio tests are used to test whether factors of interest have significant effects to the mixture distribution model. To carry out the likelihood ratio tests, different methods about the computation algorithm for the maximum likelihood estimation (MLE) of the parameters in the mixture models are studied. These methods are applied to the data obtained from a laboratory study on emergence of *Pontomyia oceana*, where the effects of factors, such as sex and temperature, to the distribution of the dates that *Pontomyia oceana* emerged are investigated. From the results obtained, in some cases, three-component logistic distributions are fitted to the data with two peaks very close to each other. This is somewhat surprising as merely from the histogram, it is not easy to see and usually not expected to say there are two very close peaks. From the practical point of view, as the laboratory conditions excluded the possible effects related to semi-lunar tidal fluctuations that may have a dominating influence in nature. Thus the laboratory results helps to identify all the possible factors that have minor effects. Based on the results of this study, the difference between males and females, nevertheless, suggests that sex hormone may be involved in affecting the emergence dates. The suggestion of a third peak is unexpected from our point of view and it implies that there are factors we never suspected. It is worth noting that through rigorous statistical analysis presented here, it helps to provide an objective estimation on the distribution of the emergence dates as well as the corresponding proportions and the peak synchronous emergence dates in each period under different factor effects. We only start to speculate its possible adaptive meaning after the differences have been established as a true phenomenon. From this study, it reveals some additional biological phenomena worthy of more investigations.

Key words: EM algorithm, likelihood ratio tests, maximum likelihood estimates, mixture of logistic distributions, synchronous emergence of *pontomyia oceana*.

1. Introduction

Pontomyia oceana(*P. oceana* in short) is a kind of marine midges at ocean front. One of the characteristics of this kind of midges is that their life histories consist of a relatively long period of benthic larval development(30-45 days) and followed by a very short adult stage(2 hours). In nature, a semi-lunar rhythm of the midge emergence was observed. The midges emerged around new moons and full moons. It was also noticed that the midges might be concentrated in the first or the last few days of the window suitable for emergence depending on season. We had no idea what environmental factors might be affecting this trait, and we had no idea if an endogenous rhythm was involved. The temperature treatments in this study were an effort to test if an endogenous rhythm was involved. An important characteristic of an endogenous rhythm is temperature-compensation, i.e., the midges are expected to catch up under low temperatures. On the other hand, the difference between males and females in emergence time of the day is known with a convincing adaptive explanation (Soong et al 1999). However difference in dates within a month has been suspected, but never proven. In our study here, as the laboratory conditions excluded the possible effects related to semi-lunar tidal fluctuations that may have a dominating influence in nature. Thus the lab results helped to identify all the possible factors that have minor effects. On the other hand, as an adult remained active for only about 2 hours, during the short adult stage, males must find mates and females besides mating must place fertilized eggs in appropriate habitats. Hence, peak synchronous emergence become very important to these short-lived midges. See Soong et al (1999) for some details on the introduction of life histories about marine midges. In this work, we have investigated whether the factor of temperature and sex indeed have an effect on synchronous emergence of *P. oceana* from statistical point of view with the following laboratory experiment observations, where mixture model describing the emergence distribution and the corresponding proportions as well as peak synchronous dates for each factor combination has been used and estimated. Here the standard ANOVA is not used as we also concern about how many peak periods there are for the emergence date distribution, as well as what are the corresponding peak dates and proportion emergenced during each period. Standard ANOVA will not be able to distinguish all these differences.

The result of this analysis does not support the existence of an endogenous semi-lunar rhythm in controlling the emergence dates of this marine midge. Moreover, the difference between males and females from the analysis, nevertheless, suggests that sex hormone may be involved in affecting the emergence dates. The suggestion of a third peak is unexpected from our point of view and implies that there are factors we never suspected. The detailed modeling and analysis as well

Table 1: Observed frequency counts

Day	20^oC Female	20^oC Male	25^oC Female	25^oC Male	30^oC Female	30^oC Male
23	1	1	6	6	3	3
24	0	0	9	8	3	3
25	1	4	24	10	42	27
26	3	2	37	5	63	65
27	2	2	56	30	34	9
28	2	3	34	24	29	21
29	6	3	7	4	5	0
30	13	9	4	0	7	0
31	10	1	3	0	3	4
32	0	0	2	4	0	0
33	1	0	1	3	3	12
34	0	0	5	6	4	3
35	0	0	8	6	9	4
36	1	1	11	14	6	1
37	1	0	19	21	7	2
38	0	1	26	14	12	7
39	1	0	67	33	13	4
40	0	0	67	33	7	1
41	3	0	64	24	6	4
42	6	7	36	19	1	0
43	5	4	21	22	0	2
44	9	3	20	12	0	0
45	9	6	16	14	0	1
46	5	2	14	6	2	1
47	8	2	11	5	0	0
48	4	5	4	1	0	0
49	21	11	4	6	1	0
50	19	7	6	6	2	0
51	36	13	1	2	0	0
52	23	4	3	0	0	1
53	10	5	7	0	0	0
54	1	0	1	0	0	0
55	1	1	5	0	0	0
56	0	0	0	0	0	0
n	202	97	599	338	262	175

as discussion are presented below.

Now consider the following experiment performed by a laboratory in National Sun Yat-sen University of Taiwan: at first adult *P. oceana* were collected in southern Taiwan and then placed in beakers to collect fertilized eggs. After approximately 20 days of culture, the emergence day observations from female and male *P. oceana* under three different temperatures, say 20^o, 25^o, and 30^o respectively were recorded. See Lee (2000) for a detailed description of the experiments. The sets of observed frequency counts n_{ijk} for each of the factor combination$(i, j), i = 1, 2, j = 1, 2, 3$ on the kth day in histogram form are presented in Table 1 and Figure 1. It can be seen from Figure 1, the time distributions appear to be mixture models with two or three distributions. The effect of temperature seems to be more significant than sex. Meanwhile, it seems that high temperature advances the day of emergence. We discuss all these phenomena by using statistical technique.

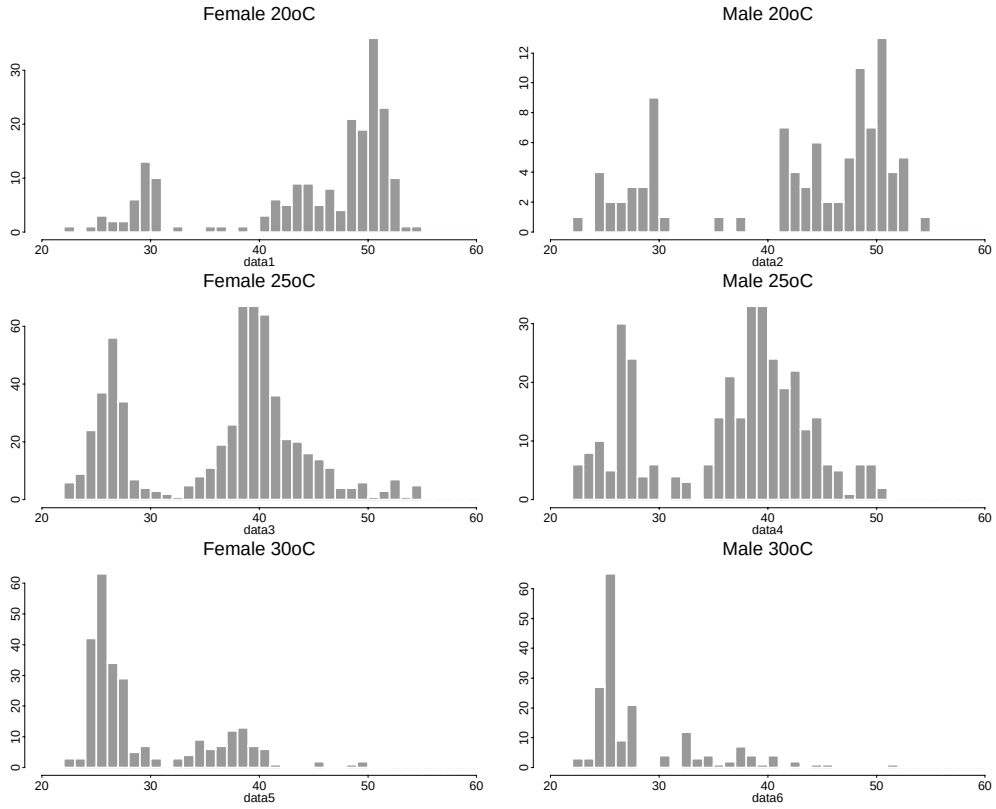

Figure 1: Histograms for quantities of emerged *P. oceana*

In this work, a two-factor factorial design with factors sex and temperature is considered although the response of emergence day for each *P. oceana* is assumed to be a random variable with mixture distribution, where the parameters of the distribution such as the mixture proportions, parameters for each distribution in the mixtures, depend on the factors. In the following, we will first consider the general mixture model and describe the methods we are going to use to do the estimation and perform the appropriate tests.

Let y_{ijk} be the observed response of the emergence day with factor A at the ith level ($i = 1, 2, \ldots, a$) and factor B at the jth level ($j = 1, 2, \ldots, b$) for the kth replicate ($k = 1, 2, \ldots, n_{ij}$). $y_{ijk}, k = 1, 2, \ldots, n_{ij}$ are assumed to be i.i.d. random variables with common mixture distribution function $F(\cdot; \boldsymbol{\theta}_{ij})$, where $\boldsymbol{\theta}_{ij}$ is the unknown parameter vector for $F(\cdot; \boldsymbol{\theta}_{ij})$. We will estimate these parameters first and perform the factor testing later.

In the practical problem here it is of interest to know whether the factors of temperature and sex have effects on the emergence day of *P. oceana*, and understand how those factors would affect the peak periods for the emergence to occur. In this work we first identify the pattern of the emergence time as following a mixture distribution, then distinguish whether there are significant differences on the mixture distributions of emergence day under different factor influences. Later the peak periods with the estimated mixture distributions are estimated through the mean estimate for each component. In order to do that, likelihood ratio tests are used here to accomplish our investigations. To perform the likelihood ratio tests, we have to determine what kind of mixture distribution function are appropriate for the data in hand first. Moreover, among the many popular existing methods for computing the maximum likelihood estimations(MLE), which one is more suitable to be used for our data will be explained in the following sections. The approaches and the methods we adopt here will be briefly described although still keep the presentation clear.

In Section 2 we introduce the definition of a finite mixture distribution and different existing methods to estimate the parameters, such as the method of moments and the maximum likelihood. The computation algorithms for finding MLE of parameter vectors of mixture distributions under different considerations to accommodate the practical situation are also introduced, namely, the EM algorithm proposed first by Demmpster, Laird,and Rubin (1977), the EM algorithm for grouped and truncated data proposed by McLachlan and Jones (1988) and Newton-Raphson iterative scheme by Hasselblad (1966). Later the mixture of normal distributions as well as the mixture of logistic distributions, both will be used to analyze our data. Moreover a method for estimating a probability density function nonparametrically proposed by Minnotte (1998) is considered, so as to recompute data in each interval and estimate mixture distributions afresh for

comparison with the results from parametric models. The Pearson's chi-squared goodness of fit test for ascertaining whether an assumed probability distribution is consistent with a given set of data is also stated. In Section 3 likelihood ratio tests for testing the effects of main factors are formulated, a F-test is also used for testing the factor interaction. In Section 4 we applied all these methods to the emergence data of *P. oceana*, and in Section 5 we conclude with a discussion.

2. Mixture Distributions and Methods of Estimation of Parameters

The probability density function (p.d.f.) of a mixture distribution with finite components is usually expressed as

$$f(x) = \sum_{i=1}^{c} p_i g_i(x; \boldsymbol{\theta}_i). \tag{2.1}$$

where p_i, $i = 1, \ldots, c$, are the mixing proportions satisfing $\sum_{i=1}^{c} p_i = 1$ and $g_i(x; \boldsymbol{\theta}_i)$, $i = 1, \ldots, c$, are the p.d.f. depending on an m-dimensional parameter vector $\boldsymbol{\theta}_i$, see Everitt and Hand (1981, p.4) for more details.

Assume n independent observations, $x_1, \ldots, x_n$, were obtained from a mixture distribution, the parameters of the distribution will be estimated. Let $\boldsymbol{\Delta} = (\boldsymbol{\theta}', \boldsymbol{p}')'$ denote the vector of all unknown parameters of (2.1), where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_c')'$ and $\boldsymbol{p} = (p_1, \ldots, p_c)'$ are the parameter vectors to be estimated. Many methods have been proposed and used for estimating $\boldsymbol{\Delta}$. Two of the well-known methods namely the method of moments and maximum likelihood will be used in this work.

The problem of estimating the parameters in a mixture of two normal distributions was first considered by Karl Pearson (1894) where the method of moments was used. The evaluation of Pearson's estimators involved the solution of a ninth degree polynomial equation. Cohen (1967) presented a procedure with circumvents solution of the nonic equation which reduces the total computational effort required. But it is not difficult to find the solution with the help of computer today.

It is well known that the maximum likelihood estimator (MLE) has a number of desirable statistical properties. In the case of mixture distributions, some computational methods for finding the MLE is described briefly in the following subsection.

2.1 Algorithms for computing the MLE

EM algorithm

EM algorithm is a powerful and useful tool for finding the MLE, which was proposed first by Dempster, Laird, and Rubin (1977). The M-steps and E-steps are repeated iteratively until some convergence criterion is satisfied.

EM algorithm for grouped and truncated data

In practice, data collected on the phenomenon of interest are frequently available only in grouped form and may also be truncated. In our study here, as the emergence number is recorded daily, the data obtained can be regarded as in grouped form, also we find that the incidence of emergence 21 days before and 56 days after the initial emergence time is very rare, we therefore have truncated our observations before 21 days and after 56 days of initial emergence. It seems to be interesting for us to see what influence it would have to our estimates under this kind of consideration. McLachlan and Jones (1988) proposed an EM algorithm for grouped and truncated data. More details can be found in the above paper.

Newton-Raphson iterative scheme

Other than using EM algorithm to obtain the MLE, it is also quite common to find the MLE by using a Newton iterative scheme (Hasselblad (1966)). Estimators obtained by method of moments can be regard as the initial values of the EM algorithm and the Newton's iteration procedure.

2.2 Density estimation with binned data

Other than estimating the p.d.f. parametrically assuming a possible form for the p.d.f., the approach of estimating the p.d.f. nonparametrically is also considered here for comparison. One popular method for estimating a p.d.f. nonparametrically is known as the kernel density estimation, where Minnotte (1998) proposed a method for achieving higher-order convergence rates in nonparametric density estimation. This will also be used for comparison here.

2.3 Pearson's chi-squared goodness-of-fit test

As soon as we have obtained the estimation of parameters, we need to test whether the particular estimated p.d.f. is consistent with those observed data. The *Pearson's chi-squared goodness of fit* test will be used.

3. Effects Testing

In a multi-factors experiment, once a model is fitted to the observed data

in each combination of the treatments of all factors and is not rejected by the goodness of fit test, we may proceed to consider the likelihood ratio test to test the effect of each factor

3.1 Likelihood ratio test

To test $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ versus $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$, the well known likelihood ratio test statistic is defined through

$$\lambda = \frac{\sup\{p(\boldsymbol{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}}{\sup\{p(\boldsymbol{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}}, \tag{3.1}$$

where $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1$. Under some additional regularity conditions, the asymptotic distribution of the statistic $-2\log\lambda$ is χ^2_{r-m}, where $\boldsymbol{\Theta}$ is an r-dimensional subset of R^r and $\boldsymbol{\Theta}_0$ is an m-dimensional subset of $\boldsymbol{\Theta}$. One rejects $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ whenever $-2\log\lambda > C$, where C is determined by the desired level of the test.

For example, in our study here, for the observed data sets of female and male at different temperatures as in Figure 1, let $\boldsymbol{\Delta}_i$, $i = 1, \ldots, 6$, denote the parameter vectors of corresponding distribution, respectively. If we want to test the effect of sex under temperature 20^oC, the test statistic used here is through

$$\lambda_{12} = \frac{L(\hat{\boldsymbol{\Delta}}_{12})}{L(\hat{\boldsymbol{\Delta}}_1)L(\hat{\boldsymbol{\Delta}}_2)}, \tag{3.2}$$

where $L(\hat{\boldsymbol{\Delta}}_i) = \prod_{j=1}^{n} f(x_j; \hat{\boldsymbol{\Delta}}_i)$, $i = 1, 2$ are the maximized likelihood for female and male at 20^oC respectively, $\boldsymbol{\Delta}_{12}$ is the parameter vector of the combined data sets of female and male at temperatures 20^oC, and $L(\hat{\boldsymbol{\Delta}}_{12})$ is the maximized likelihood of the combined data set of female and male at 20^oC; the degrees of freedom equals to the difference between the sum of the number of parameters of $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$ and the number of parameters of $\boldsymbol{\Delta}_{12}$. To test the effect of temperature under female, the test statistic λ_{135} is adopted in the same manner, where

$$\lambda_{135} = \frac{L(\hat{\boldsymbol{\Delta}}_{135})}{L(\hat{\boldsymbol{\Delta}}_1)L(\hat{\boldsymbol{\Delta}}_3)L(\hat{\boldsymbol{\Delta}}_5)}. \tag{3.3}$$

Regardless of different temperatures, the entire effect of sex can be tested by

$$\lambda_s = \frac{L(\hat{\boldsymbol{\Delta}}_{12})L(\hat{\boldsymbol{\Delta}}_{34})L(\hat{\boldsymbol{\Delta}}_{56})}{\prod_{i=1}^{6} L(\hat{\boldsymbol{\Delta}}_i)}. \tag{3.4}$$

The degrees of freedom equal to the difference in dimension between the sum of the number of parameters of $\boldsymbol{\Delta}_{12}$, $\boldsymbol{\Delta}_{34}$ and $\boldsymbol{\Delta}_{56}$ and the sum of the number of parameters of $\boldsymbol{\Delta}_i$, $i = 1, \ldots, 6$. Similarly, the entire effect of temperature can be

tested by

$$\lambda_t = \frac{L(\hat{\mathbf{\Delta}}_{135})L(\hat{\mathbf{\Delta}}_{246})}{\prod_{i=1}^{6} L(\hat{\mathbf{\Delta}}_i)}. \tag{3.5}$$

3.2 F-test

To test the interaction effect between temperature and sex, we use the F-statistics comparing effect of factor A at low level of factor B with respect to effect of factor A at high level of factor B as follows.

$$\begin{aligned} F^{factor} &= \frac{-2\log\lambda_{low}/d.f._{\cdot low}}{-2\log\lambda_{high}/d.f._{\cdot high}} \\ &\sim F_{d.f._{\cdot low},d.f._{\cdot high}} \end{aligned}$$

that is,

$$F^{temp} = \frac{\log\lambda_{ij}/df_{\,ij}}{\log\lambda_{hk}/df_{\,hk}} \tag{3.6}$$

for ij, $hk \in \{12, 34, 56\}$ and

$$F^{sex} = \frac{\log\lambda_{135}/df_{\,135}}{\log\lambda_{246}/df_{\,246}}, \tag{3.7}$$

the critical region would be the tails of the F distribution with corresponding degrees of freedom. The reason for using $F^{temp}, ij = 12, hk = 34$ to test the interaction between sex and temperatures at 20^oC and 25^oC is that we are testing whether the effect of sex under 20^oC would be significantly different from the effect of sex under 25^oC, and so on. Then if the value of F^{temp} or F^{sex} is too small or too large, it means that the differences of variation of sex between different temperatures are significant or the differences of variation of temperature between female and male are significant, then it indicates the effect of interaction does exist.

4. Case Study

We now examine the effects of sex and temperature on the day of emergence of *P. oceana*. In a laboratory trial, the ova fertilized at the same day were collected and kept in constant temperature 20^oC, 25^oC and 30^oC separately. The observed counts of emergence of female and male *P. oceana* at 20^oC, 25^oC and 30^oC from the 21th day to 56th day were recorded every day as are presented in Table 1. The total observed counts of each combination n is listed at the last row of Table 1. The 21th and 56th day were the lower and upper truncated values, and 36 grouped intervals with 1 day equal width.

Table 2: Results of fitting a three(two)-component mixture of logistic distributions to day of emergency (Standard errors are in parentheses)

		$20^oC\ F$ $\hat{\mathbf{\Delta}}_1$	$20^oC\ M$ $\hat{\mathbf{\Delta}}_2$	$25^oC\ F$ $\hat{\mathbf{\Delta}}_3$	$25^oC\ M$ $\hat{\mathbf{\Delta}}_4$	$30^oC\ F$ $\hat{\mathbf{\Delta}}_5$	$30^oC\ M$ $\hat{\mathbf{\Delta}}_6$
	NL	28.85(.06)	27.57(.12)	26.06(.03)	26.04(.05)	25.79(.06)	25.57(.07)
	EM	29.12(.06)	27.87(.13)	26.13(.03)	26.25(.04)	25.78(.06)	25.46(.06)
$\hat{\mu}_1$	GT	29.11(.06)	27.86(.13)	26.13(.03)	26.22(.05)	25.81(.07)	25.50(.06)
	NR	29.12(.06)	27.87(.13)	26.13(.03)	26.22(.05)	25.80(.07)	25.49(.07)
	BD	29.16(.06)	27.66(.12)	26.19(.04)	26.23(.06)	25.88(.08)	25.55(.09)
$\hat{\sigma}_1$	NL	2.06(.19)	2.23(.38)	1.45(.07)	1.66(.12)	1.21(.12)	1.11(.12)
	EM	1.11(.03)	1.45(.07)	.83(.02)	.85(.02)	.70(.03)	.50(.03)
$\hat{\beta}_1$	GT	1.11(.03)	1.44(.07)	.81(.02)	.95(.02)	.73(.03)	.57(.03)
	NR	1.11(.03)	1.45(.07)	.83(.02)	.95(.03)	.76(.04)	.58(.03)
	BD	1.12(.03)	1.34(.06)	.94(.02)	1.08(.03)	.88(.04)	.76(.05)
	NL	44.59(.15)	43.35(.19)	39.50(.05)	39.67(.16)	37.62(.06)	32.44(.03)
	EM	44.08(.11)	43.05(.11)	39.43(.05)	39.21(.16)	36.91(.14)	34.44(.24)
$\hat{\mu}_2$	GT	44.08(.11)	43.05(.11)	39.43(.05)	39.11(.15)	37.03(.13)	35.16(.23)
	NR	44.08(.11)	43.05(.11)	39.43(.05)	40.10(.21)	37.05(.13)	35.11(.23)
	BD	44.31(.14)	42.89(.11)	39.47(.07)	39.80(.17)	37.00(.14)	35.31(.21)
$\hat{\sigma}_2$	NL	3.19(.72)	2.82(.82)	1.46(.12)	3.42(.88)	1.78(.19)	1.14(.06)
	EM	1.56(.06)	1.15(.06)	.93(.03)	1.89(.09)	2.22(.08)	2.88(.12)
$\hat{\beta}_2$	GT	1.55(.06)	1.13(.06)	.90(.03)	1.81(.08)	2.13(.07)	2.83(.12)
	NR	1.56(.06)	1.15(.06)	.93(.03)	2.56(.11)	2.13(.07)	2.86(.12)
	BD	1.83(.08)	1.16(.06)	1.16(.04)	2.10(.09)	2.18(.07)	2.66(.11)
	NL	50.42(.08)	49.94(.13)	41.88(.15)	49.23(.01)	34.90(.22)	38.28(.11)
	EM	50.34(.08)	49.74(.14)	42.05(.15)	43.46(.14)	–	–
$\hat{\mu}_3$	GT	50.34(.08)	49.74(.13)	42.00(.15)	42.04(.18)	–	–
	NR	50.34(.08)	49.74(.14)	42.06(.14)	–	–	–
	BD	50.42(.09)	49.67(.17)	42.70(.15)	49.35(.01)	–	–
$\hat{\sigma}_3$	NL	1.34(.16)	1.70(.33)	5.48(1.30)	.71(.01)	6.17(1.88)	3.18(.52)
	EM	.82(.04)	1.05(.07)	3.10(.08)	2.35(.07)	–	–
$\hat{\beta}_3$	GT	.80(.04)	1.03(.07)	3.09(.08)	2.54(.09)	–	–
	NT	.82(.04)	1.05(.07)	3.10(.07)	–	–	–
	BD	.90(.05)	1.25(.09)	3.11(.08)	.56(.01)	–	–
	NL	.19(.03)	.26(.04)	.29(.02)	.26(.01)	.66(.03)	.74(.03)
	EM	.20(.03)	.27(.04)	.29(.02)	.25(.02)	.69(.003)	.69(.003)
$\hat{p}_1$	GT	.20(.03)	.27(.04)	.29(.02)	.25(.02)	.70(.002)	.70(.004)
	NR	.19(.03)	.26(.04)	.32(.02)	.32(.03)	.70(.002)	.70(.004)
	BD	.19(.03)	.25(.04)	.29(.02)	.27(.01)	.70(.002)	.71(.003)
	NL	.29(.04)	.29(.05)	.35(.03)	.71(.01)	.15(.03)	.11(.02)
	EM	.26(.03)	.24(.04)	.37(.03)	.59(.04)	–	–
$\hat{p}_2$	GT	.26(.03)	.24(.04)	.37(.03)	.49(.06)	–	–
	NR	.25(.03)	.24(.04)	.41(.03)	–	–	–
	BD	.28(.04)	.21(.04)	.42(.03)	.70(.01)	–	–
	NL	-586	-302	-1817	-1048	-676	-429
	EM	-584	-301	-1816	-1057	-681	-429
$\log L$	GT	-584	-301	-1816	-1056	-681	-428
	NR	-584	-301	-1819	-1064	-680	-428
	BD	-585	-302	-1819	-1052	-683	-433

4.1 Mixture of normal distributions

The sets of observed frequency counts in histogram form given in Figure 1

suggests that the time distributions are mixture models, which also seems to be able to explain the real situation reasonably. Since the experimenter collected data in grouped and truncated form, we first fit mixture of three normal distributions to these data via formulas deduced by McLachlan and Jones(1988). The estimate of parameter vectors $\mathbf{\Delta}_i$, including parameters μ_{1i}, σ_{1i}, μ_{2i}, σ_{2i}, μ_{3i}, σ_{3i} and proportions p_{1i} and p_{2i}, $i = 1, \ldots, 6$, are listed at the first row of Table 2 and denoted by NL. But the fits are rejected since the modes of the histograms have higher frequencies and smaller variations than normal distributions do. Hence a mixture of logistic distributions is considered to fit to each data set.

4.2 Mixture of logistic distributions

Since mixture of normal distributions does not work well, we have tried some other distributions. It seems that mixture of logistic distributions is more suitable. In order to obtain the MLE of mixture of logistic distributions, the formulas for performing EM algorithm and EM algorithm for grouped and truncated data are derived and used in our estimates, the derivation can be obtained following the steps illustrated in the corresponding papers and is omitted.

The p.d.f. of logistic distribution is

$$f(x;\mu,\beta) = \frac{e^{-\frac{x-\mu}{\beta}}}{\beta(1+e^{-\frac{x-\mu}{\beta}})^2}. \tag{4.1}$$

To estimate the MLE of parameter vector $\mathbf{\Delta} = (\mu_1, \beta_1, \mu_2, \beta_2, \mu_3, \beta_3, p_1, p_2)$ of a mixture of three logistic distributions four methods were used.

The first is the EM algorithm where the effect of truncation is not considered since the part of the tails seems to be ignorable. Moreover, since the experimental data were collected in grouped form, we regard the individual observations x in each interval as the central point of the interval. For example, at 20^oC we observed that there were 3 female *P. oceana* emerged at the 26th day, then $x_3 = x_4 = x_5 = 25.5$. According to this rule and EM algorithm for mixture of logistic distributions, we obtained the estimates of parameter vectors $\mathbf{\Delta}_i$, $i = 1, \ldots, 6$, listed at row 2 of Table 2 and denoted by EM.

Secondly, we estimate the parameter vectors with consideration of data are initially grouped and truncated or by the Newton-Raphson. The results are listed at row 3 and 4 of Table 2 and denoted by GT and NR respectively.

The last method we used is fitting a nonparametric density, to each set of experimentally observed data in order to smooth the original rugged curve. Then estimate the probable counts at each day by multipling total number to this nonparametric density, and re-estimate the parameter vectors of mixture of logistic distributions. The results are listed at row 5 of Table 2 and denoted by BD.

Compare the results of these four methods presented in Table 2, there is no significant difference between the estimates obtained by EM algorithm, EM algorithm for grouped and truncated data and Newton scheme except for $\hat{\mathbf{\Delta}}_4$. The purpose of estimating a probability function nonparametrically is to smooth down the rough data style initially so as to obtain a better parametric form. But the estimates obtained by BD do not make significant differences here. On the other hand, it seems that the estimates obtained by GT has comparatively larger log-likelihood value listed at the bottom of Table 2 than the others and the curve of the density function with GT estimate also has a better peak estimate at the mode than the others. Hence in the following the discussion will be restricted to the estimates obtained by GT only, and the plots of the estimated density function for each data set are also presented in Figure 2.

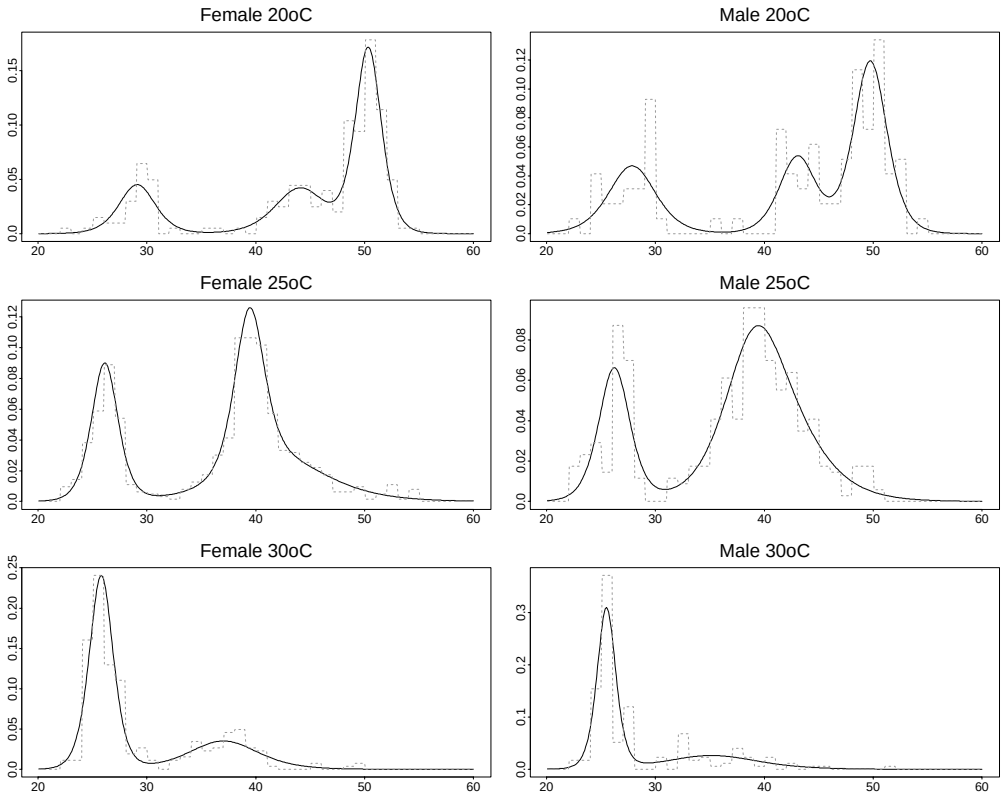

Figure 2: Plots of mixture of logistic distributions

4.3 Distribution fits

To test whether the mixture of logistic distributions is consistent with the observed data, the Pearson's chi-square goodness-of-fit statistic is considered. Before we calculate the test statistics value, we make two adjustments on some intervals. First we have combined several intervals with extremely small probabilities $P_j(\hat{\mathbf{\Delta}})$, since even only one observation occurred in this kind of intervals will cause the test statistics value increased rapidly. Next, we have combined two neighboring intervals when in one interval the observed counts decrease rapidly and in the next interval the response increases to the seemly normal counts. The reason for this adjustment is because no matter what type of distribution was fitted, the unsmoothed rapid increase or decrease would almost always cause the null hypothesis of fit by a distribution with smooth p.d.f. being rejected. Hence combining intervals to alleviate the concussion seems to be a reasonable thing to do. Then the degrees of freedom are reduced accordingly. The results are given in Table 3 for each data set along with the associated p-value. It shows that those fittings are acceptable.

Table 3: Goodness-of-fit test for mixture of logistic distributions by GT (Standard errors are in parentheses)

	20^oC F $\hat{\mathbf{\Delta}}_1$	20^oC M $\hat{\mathbf{\Delta}}_2$	25^oC F $\hat{\mathbf{\Delta}}_3$	25^oC M $\hat{\mathbf{\Delta}}_4$	30^oC F $\hat{\mathbf{\Delta}}_5$	30^oC M $\hat{\mathbf{\Delta}}_6$
$\hat{\mu}_1$	29.11(.06)	27.86(.13)	26.13(.03)	26.22(.05)	25.81(.07)	25.50(.06)
$\hat{\beta}_1$	1.11(.03)	1.44(.07)	.81(.02)	.95(.02)	.73(.03)	.57(.03)
$\hat{\mu}_2$	44.08(.11)	43.05(.11)	39.43(.05)	39.11(.15)	37.03(.13)	35.16(.23)
$\hat{\beta}_2$	1.55(.06)	1.13(.06)	.90(.03)	1.81(.08)	2.13(.07)	2.83(.12)
$\hat{\mu}_3$	50.34(.08)	49.74(.13)	42.00(.15)	42.04(.18)	–	–
$\hat{\beta}_3$	.80(.04)	1.03(.07)	3.09(.08)	2.54(.09)	–	–
$\hat{p}_1$	.20(.03)	.27(.04)	.29(.02)	.25(.02)	.70(.002)	.70(.004)
$\hat{p}_2$	.26(.03)	.24(.04)	.37(.03)	.49(.06)	–	–
χ^2	21.657	22.744	19.360	24.004	14.476	20.731
$d.f.$	20	20	20	20	13	13
p-value	.359	.302	.499	.242	.341	.078

4.4 Effects significance of temperature and sex

From Table 3, it can be seen that three-component logistic distribution fits the data sets of 20^oC and 25^oC and two-component model fits the data sets of 30^oC quite well. The estimates of the parameter vectors at the same temperature are close for female and male. It seems that the effect of temperature is more significant than sex.

Table 4: Results of fitting a three-component mixture of logistic distributions to combined data (Standard errors are in parentheses)

	20^oC $\hat{\mathbf{\Delta}}_{12}$	25^oC $\hat{\mathbf{\Delta}}_{34}$	30^oC $\hat{\mathbf{\Delta}}_{56}$	Female $\hat{\mathbf{\Delta}}_{135}$	Male $\hat{\mathbf{\Delta}}_{246}$
$\hat{\mu}_1$	28.66(.06)	26.17(.03)	25.64(.05)	26.17(.03)	25.88(.04)
$\hat{\beta}_1$	1.31(.03)	.88(.01)	.67(.02)	.93(.02)	.86(.02)
$\hat{\mu}_2$	43.67(.08)	39.33(.05)	36.21(.13)	39.84(.08)	39.55(.14)
$\hat{\beta}_2$	1.40(.05)	1.10(.03)	2.54(.07)	1.92(.04)	2.52(.08)
$\hat{\mu}_3$	50.17(.07)	41.56(.13)	–	50.24(.02)	49.73(.02)
$\hat{\beta}_3$	.90(.04)	3.05(.07)	–	1.08(.01)	.83(.01)
$\hat{p}_1$	.22(.02)	.26(.01)	.69(.01)	.38(.01)	.39(.01)
$\hat{p}_2$	.25(.03)	.34(.03)	–	.48(.01)	.53(.01)
$\log L$	−892	−2885	−1114	−3427	−1966

Table 5: Analysis of interaction between temperature and sex — H_0: no interaction between temperature and sex

	$-2\log\lambda$		$d.f.$	F^{temp}	p-value
20^oC	$-2\log\lambda_{12} =$	13.496	8	$\frac{13.496/8}{25.968/8} = 0.52$	0.813
25^oC	$-2\log\lambda_{34} =$	25.968	8	$\frac{25.968/8}{11.794/5} = 1.38$	0.377
30^oC	$-2\log\lambda_{56} =$	11.794	5	$\frac{11.794/5}{13.496/8} = 1.40$	0.320
	$-2\log\lambda$		$d.f.$	F^{sex}	p-value
$Female$	$-2\log\lambda_{135} =$	691.930	13	$\frac{691.930/13}{361.722/13} = 1.91$	0.128
$Male$	$-2\log\lambda_{246} =$	361.722	13		

In order to test the effect of sex and temperature, we combine the data sets for different combinations and fit mixture of logistic distributions to each combined data set. The MLE for combined data set of sets i and j or sets i , j and k are denoted by $\hat{\mathbf{\Delta}}_{ij}$ and $\hat{\mathbf{\Delta}}_{ijk}$ respectively, which are displayed in Table 4. The maximized likelihood $L(\hat{\mathbf{\Delta}})$ of each data set are listed in Table 3 and Table 4 separately. Consider the likelihood ratio test statistics λ in (3.2) and (3.3), the effect of sex at 20^oC, 25^oC and 30^oC are indicated by λ_{12}, λ_{34} and λ_{56} respectively; the effect of temperature for female and male are indicated by λ_{135} and λ_{246} respectively. All these value are listed in Table 5.

Now, we test the effect of interaction between temperature and sex, i.e., the null hypothesis is H_0: there is no interaction between temperature and sex. We check each of the statistics F^{temp} and F^{sex} as in (3.6) and (3.7) at different

levels of temperature and sex. The test results are shown in Table 5 and we conclude that there is no indication that there is significant interaction between temperature and sex. This seems reasonable since in Figure 2 it shows that the patterns of female and male at the same temperature are similar and in Table 3 the scales and locations of the distributions at the same temperature seem to be quite close.

Next, we are going to test the effect of sex under the consideration of no significant interaction between temperature and sex, i.e., the null hypothesis is H_0: there is no sex effect. The results listed in Table 6 show that at level $\alpha = .05$ the entire effect of sex tested by λ_s as in (3.4) is significant, since p-value of $-2log\lambda_s$ is .0002. The effect of sex at different temperatures are tested by the likelihood ratio test statistic as in (3.2). As shown in Table 6, the effect of sex at 20^oC is less significant and is significant when the temperature is 25^oC or 30^oC. The most significant one is at 25^oC where the p-value is .001.

Table 6: Likelihood ratio test under the hypothesis — H_0: the effect of sex is not significant

	20^oC	25^oC	30^oC	Total Effect
$-2log\lambda$	13.496	25.968	11.794	51.250
$d.f.$	8	8	5	21
p-value	.096	.002	.038	.0002

Finally, we test the effect of temperature, i.e., the null hypothesis is H_0: there is no temperature effect. Through the same argument as testing the effect of sex, the likelihood ratio test statistics λ_t as in (3.5) is used, and it is not surprising that the effect is significant as shown in Table 7. The effects of temperature on female and male are tested by λ_{135} and λ_{246} separately, both show that the effect is significant, since the p-values are almost zero as shown in Table 7.

Table 7: Likelihood ratio test under the hypothesis — H_0: the effect of temperature is not significant

	Female	Male	Total Effect
$-2\log\lambda$	691.930	361.722	1053.652
$d.f.$	13	13	26
p-value	0	10^{-7}	10^{-7}

Compare the results given in Tables 6 and 7, it is clear that the effect of temperature is more significant than sex. The p-values of goodness-of-fit test for the MLE of combined data sets listed in Table 8 which show that the fitting

of $\hat{\mathbf{\Delta}}_{34}$ and $\hat{\mathbf{\Delta}}_{135}$ are not suitable and this is reasonable since we have already shown that the effect of sex at 25^oC and the effect of temperature for female are more significant than others, it indicates that there is an interaction between temperature and sex but it is not as significant enough so as to make the total interaction become significant.

Table 8: Goodness-of-fit test for mixture of logistic distributions of combined data

	20^oC $\hat{\mathbf{\Delta}}_{12}$	25^oC $\hat{\mathbf{\Delta}}_{34}$	30^oC $\hat{\mathbf{\Delta}}_{56}$	Female $\hat{\mathbf{\Delta}}_{135}$	Male $\hat{\mathbf{\Delta}}_{246}$
χ^2	20.901	33.462	30.190	31.700	18.072
$d.f.$	20	20	20	20	20
p-value	.403	.030	.067	.047	.583

4.5 Other related results

In Figure 2 it seems that the curve has only two peaks, but in Table 3, a three-component logistic distribution has been fitted to the data sets of 25^oC. Note that the third peak is estimated to occur at the $42th$ day close to the second one at the $39th$ day, so that the third peak in the curve is not easy to be distinguished just by a rough observation. The suggestion of a third peak is unexpected from our point of view. As the laboratory conditions excluded the possible effects related to semilunar tidal fluctuations that may have a dominating influence in nature. Thus the lab results helps to identify all the possible factors that have minor effects. The result of this analysis does not support the existence of an endogenous semilunar rhythm in controlling the emergence dates of this marine midge. Moreover, the difference between males and females from the analysis, nevertheless, suggests that sex hormone may be involved in affecting the emergence dates. The third peak implies that there are factors we never suspected.

Furthermore, from the analysis it shows that amounts of emergence of *P. oceana* are the highest at 25^oC. Also, it can be seen from the fitted models that the higher the temperature is, the higher proportion of emergence at the first high peak is. Finally, the models show that the day of emergence of *P. oceana* has a high peak first at 30^oC, next at 25^oC and 20^oC the last. That is, high temperature advances the day of emergence. In nature there are cues to constrain which days are allowed to emerge (new moons and full moons). Advanced emergence may be observed within those allowed dates. The temperature effect may determine the proportion to emerge in the first available window allowable

(emergence dates). Moreover from the experimental data and the estimates of parameters $p_1, p_2, \mu_1, \mu_2, \beta_1, \beta_2$ etc., we are able to estimate the number of populations and the corresponding peak date of the emergence for experiments in the laboratory, which is helpful to study the synchronous emergence pattern of *P. oceana*. For the pattern in nature there are still a lot of unanswered questions to be investigated.

5. Discussion

In the process of finding suitable models for the data, some problems have arised. It can be seen that for the male data sets, the fits are not as good for the mode heights using the mixture of normal distributions. Some other distributions have also been used to fit these data sets, but it seems that the mixture of logistic distributions approaches the high peak of the observed data more closely. This is an interesting phenomenon as the logistics seems to be able to provide a steeper pattern for the density estimate than other types of continuous distributions, such as the normal. The major differences in the ability to fit data with steep mode between these two families of distributions are of interest for further investigation. In the case of our study the observed mixture density pattern is not quite smooth due to the nature of the midge, but a mixture of logistic fit does seem to be able to present a reasonable approximation of the practical situation quite well. It is worth noting that through rigorous statistical analysis presented here, it helps to provide an objective estimation on the distribution of the emergence dates as well as the corresponding proportions and the peak synchronous emergence dates in each period under different factor effects. We only start to speculate its possible adaptive meaning after the differences have been established as a true phenomenon. From this study, it reveals some additional biological phenomena worthy of more investigations.

References

Cohen, A. Clifford (1967). Estimation in Mixtures of two Normal Distributions. *Technometrics* **9**, 15-28.

Dempster, A. P., Laird, N. M., and Rubin, D. R. (1977). Maximum Likelihood from Incomplete Data(with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1-38.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions*. Chapman and Hall.

Hasselblad, V. (1966). Estimation of Parameters for a Mixture of Normal Distributions. *Technometrics* **8**, 431-446.

Lee, P. H. (2000). The effect of temperature on synchronized eclosion and the study of eclosion rhythm of the marine midge Pontomyia oceana. Master Thesis, National Sun Yat-sen University.

McLachlan, G. J. and Jones, P. N. (1988). Fitting Mixture Models to Grouped and Truncated Data via the EM Algorithm. *Biometrics* **44**, 571-578.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley.

Minnotte, M. C. (1998). Achieving Higher-Order Convergence Rates for Density Estimation With Binned Data. *Journal of the American Statistical Association*, **93**, 663-672.

Pearson, K. (1894). Contributions to the Mathematical Theory of Evolution. *Philosophical Transactions of the Royal Society* **185**, 71-110.

Soong, K., Chen, G.-F., and Cao, J.-R. (1999). Life history studies on the flightless marine midges *Pontomyia* spp. (Diptera: Chironomidae). *Zoological Studies* **38**, 465-472.

Received February 14, 2003; accepted July 29, 2003.

Mong-Na Lo Huang
Department of Applied Mathematics
National Sun Yat-sen University
Kaohsiung, Taiwan, ROC
lomn@math.nsysu.edu.tw

Chun-Sui Lin
Department of Applied Mathematics
National Sun Yat-sen University
Kaohsiung, Taiwan, ROC
cslin@math.nsysu.edu.tw

Keryea Soong
Institute of Marine Biology
National Sun Yat-sen University
Kaohsiung, Taiwan, ROC
keryea@mail.nsysu.edu.tw