

STOCHASTIC MODELING OF ENVIRONMENTAL TIME SERIES

Richard W. Katz

LECTURE 2

- (1) Background on EM Algorithm
- (2) Description & Properties of EM Algorithm
- (3) Environmental Applications of EM Algorithm
- (4) Maximum Likelihood Estimation for Mixtures

(1) Background on EM Algorithm

Expectation-Maximization (EM) Algorithm

- **Basic Idea of EM Algorithm**
- **Types of Estimation**
- **Incomplete Data**

Basic Idea of EM Algorithm

- **Observed mixture**

- **Maximum likelihood estimation of parameters straightforward if case for component distributions**

- **Hidden mixture**

- **Parameter estimation much more complex**

- **Exploit simpler structure of observed mixture (via conditional expectation)**

Types of Estimation

More approach than “algorithm”

- **Maximum likelihood estimation (MLE)**
 - Primary use
- **Bayesian**
 - Bayesian orientation (“precursor” of MCMC)
 - Maximize posterior density estimation
 - Maximum penalized likelihood estimation

Incomplete Data

EM algorithm applies to data that are “incomplete”

- **Mixtures**
- **Missing data/ “imputation”**
- **Grouped data**
- **Censored or truncated data**
- **Random effects**
- **Hidden/Latent variables**

(2) Description & Properties of EM Algorithm

- **Description of EM Algorithm**
 - Expectation Step (**E-Step**)
 - Maximization Step (**M-Step**)
- **Properties of EM Algorithm**
 - Monotonicity
 - Convergence

Description of EM Algorithm

- **Conceptual example**

Mixture of two distributions with density function

$$f(x) = (1 - w) f_0(x) + w f_1(x)$$

For now, assume that only mixing proportion w is unknown

– i.e., conditional densities $f_i(x)$, $i = 0, 1$, completely specified

- **Observed data**

Given observed random sample from density f : x_1, x_2, \dots, x_n

Log likelihood for observed data, $\log L(w)$:

$$\log L(w) = \log[(1 - w) f_0(x_1) + w f_1(x_1)] + \dots + \log[(1 - w) f_0(x_n) + w f_1(x_n)]$$

Differentiating log likelihood function with respect to w and equating result to zero does not yield explicit solution for mixing proportion w

- **Unobservable data**

Let $i_j = 1$ if x_j sampled from density f_1 ($i_j = 0$ otherwise)

If i_1, i_2, \dots, i_n were observable, then **MLE** of w would be:

$$(i_1 + i_2 + \dots + i_n) / n$$

“Complete-data” log likelihood function, **$\log L_C(w)$** :

$$\log L_C(w) = (1 - i_1) [\log(1 - w) + \log f_0(x_1)] + i_1 [\log(w) + \log f_1(x_1)]$$

$$+ \dots + (1 - i_n) [\log(1 - w) + \log f_0(x_n)] + i_n [\log(w) + \log f_1(x_n)]$$

- **Eliminating unobservables**

Let I_1, I_2, \dots, I_n denote random variables corresponding to unobservable data i_1, i_2, \dots, i_n

To eliminate unobservables, take conditional expectation of I_1, I_2, \dots, I_n given observed data x_1, x_2, \dots, x_n

In case of mixture, complete data log likelihood function is linear in unobservable data i_1, i_2, \dots, i_n

So only need to calculate:

$$\begin{aligned} E(I_j | x_1, x_2, \dots, x_n) &= \Pr\{I_j = 1 | x_1, x_2, \dots, x_n\} \\ &= [w f_1(x_j)] / [(1 - w) f_0(x_j) + w f_1(x_j)] \end{aligned}$$

[Bayes' Theorem: posterior probability j th member of sample from density f_1]

- **EM algorithm**

(k+1)th iteration, $k = 0, 1, 2, \dots$

-- Expectation Step (E-Step)

$$i_j^{(k)} = w^{(k)} f_1(x_j) / [(1 - w^{(k)}) f_0(x_j) + w^{(k)} f_1(x_j)]$$

-- Maximization Step (M-Step)

$$w^{(k+1)} = (i_1^{(k)} + i_2^{(k)} + \dots + i_n^{(k)}) / n$$

Initial step of algorithm: Need to specify starting value for w , say $w^{(0)}$

Properties of EM Algorithm

- **Monotonicity**

Likelihood cannot decrease after an iteration of **EM** algorithm

-- Proof (concavity of logarithm & Jensen's inequality)

- **Convergence**

Under general conditions, convergence of **EM** algorithm to MLE

-- Issue of choosing starting values

- **Standard errors**

Not automatically produced (as with Newton-Raphson)

Approaches to obtain Hessian/observed information matrix

- **Direct numerical differential (if likelihood not too complicated)**

- **Extraction of observed information matrix from **EM** algorithm (e.g., making use of “Missing Information principle”)**

- **Resampling**

- **Rate of Convergence**

- **Methods for accelerating [e.g., generalized **EM** (**GEM**) algorithm]**

(3) Environmental Applications of EM Algorithm

- **Example**

Mixture of two exponential distributions

-- **Chico daily precipitation intensity**

- **Example**

Mixture of two normal distributions

-- **Yellowstone Geyser time between eruptions**

Example: Mixture of two exponential distributions

- EM algorithm for mixture of two exponentials (parameters w, σ_0, σ_1)

E-step: Unchanged (same for any mixture)

M-step [Solution for $(k+1)$ th step]: ($w^{(k+1)}$ as before)

$$\sigma_1^{(k+1)} = [i_1^{(k)} x_1 + i_2^{(k)} x_2 + \cdots + i_n^{(k)} x_n] / [i_1^{(k)} + i_2^{(k)} + \cdots + i_n^{(k)}]$$

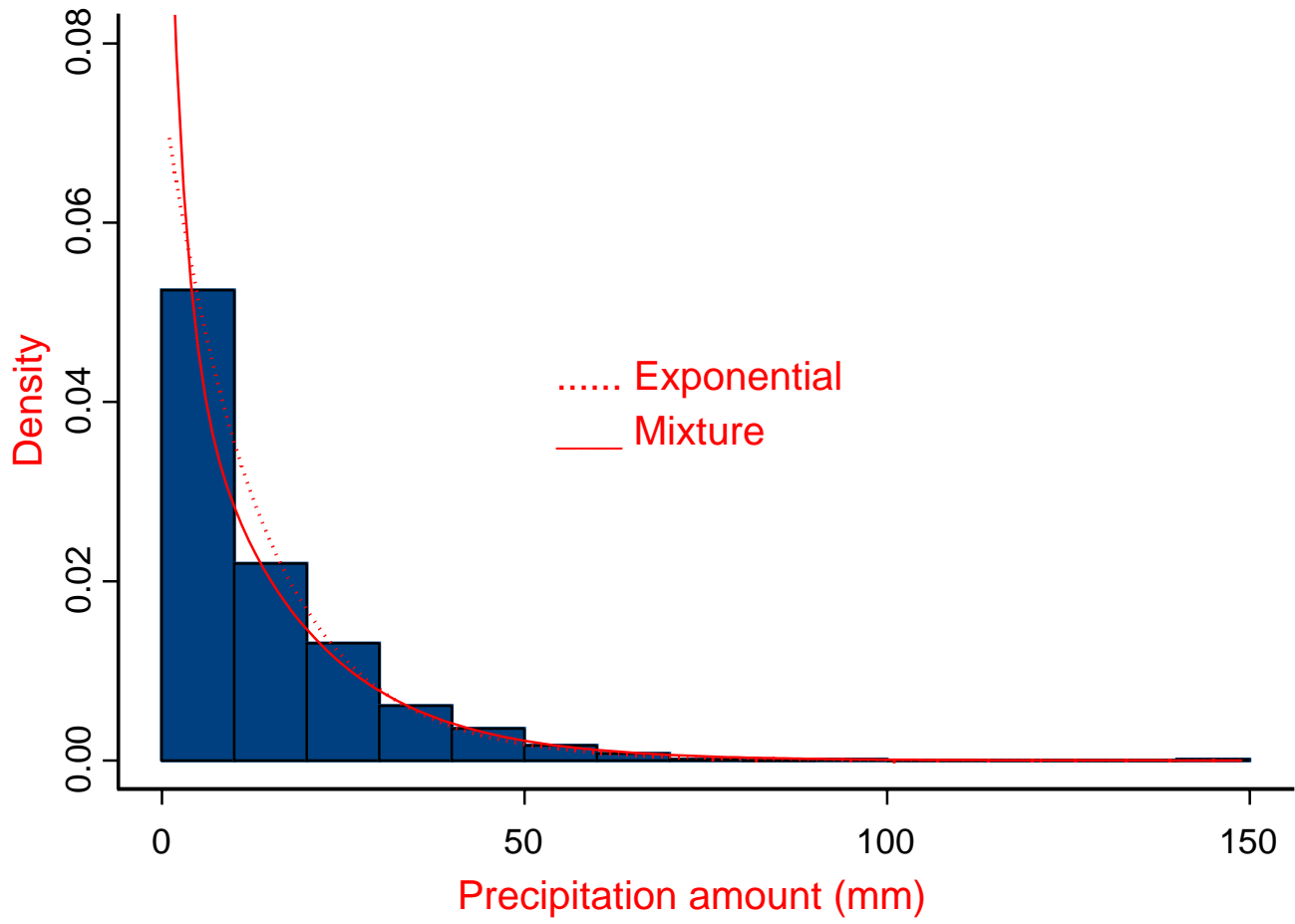
$$\sigma_0^{(k+1)} = [(1 - i_1^{(k)}) x_1 + \cdots + (1 - i_n^{(k)}) x_n] / [(1 - i_1^{(k)}) + \cdots + (1 - i_n^{(k)})]$$

- Chico, CA, USA, daily precipitation intensity (January, 78 yrs.)

$n = 787$, sample mean = 13.3596 mm

k	$w^{(k)}$	$\sigma_0^{(k)}$	$\sigma_1^{(k)}$	$\log L$
0	0.50000	6.6798	20.0394	-2822.50817
1	0.51290	6.7995	19.5896	-2821.75749
.
.
.
86	0.81735	2.1704	15.8600	-2806.58896
87	0.81736	2.1703	15.8599	-2806.58896
SE's	(0.03296)	(0.3778)	(0.7757)	

Chico precipitation intensity



Example: Mixture of Two Normal Distributions

- EM algorithm for mixture of two normals (parameters $w, \mu_0, \mu_1, \sigma_0^2, \sigma_1^2$)

E-step: Unchanged

M-step [Solution for $(k+1)$ th step]: ($w^{(k+1)}$ as before)

$$\mu_1^{(k+1)} = [i_1^{(k)} x_1 + i_2^{(k)} x_2 + \dots + i_n^{(k)} x_n] / [i_1^{(k)} + i_2^{(k)} + \dots + i_n^{(k)}]$$

$$[\sigma_1^{(k+1)}]^2 = [i_1^{(k)} (x_1 - \mu_1^{(k+1)})^2 + \dots + i_n^{(k)} (x_n - \mu_1^{(k+1)})^2] / [i_1^{(k)} + i_2^{(k)} + \dots + i_n^{(k)}]$$

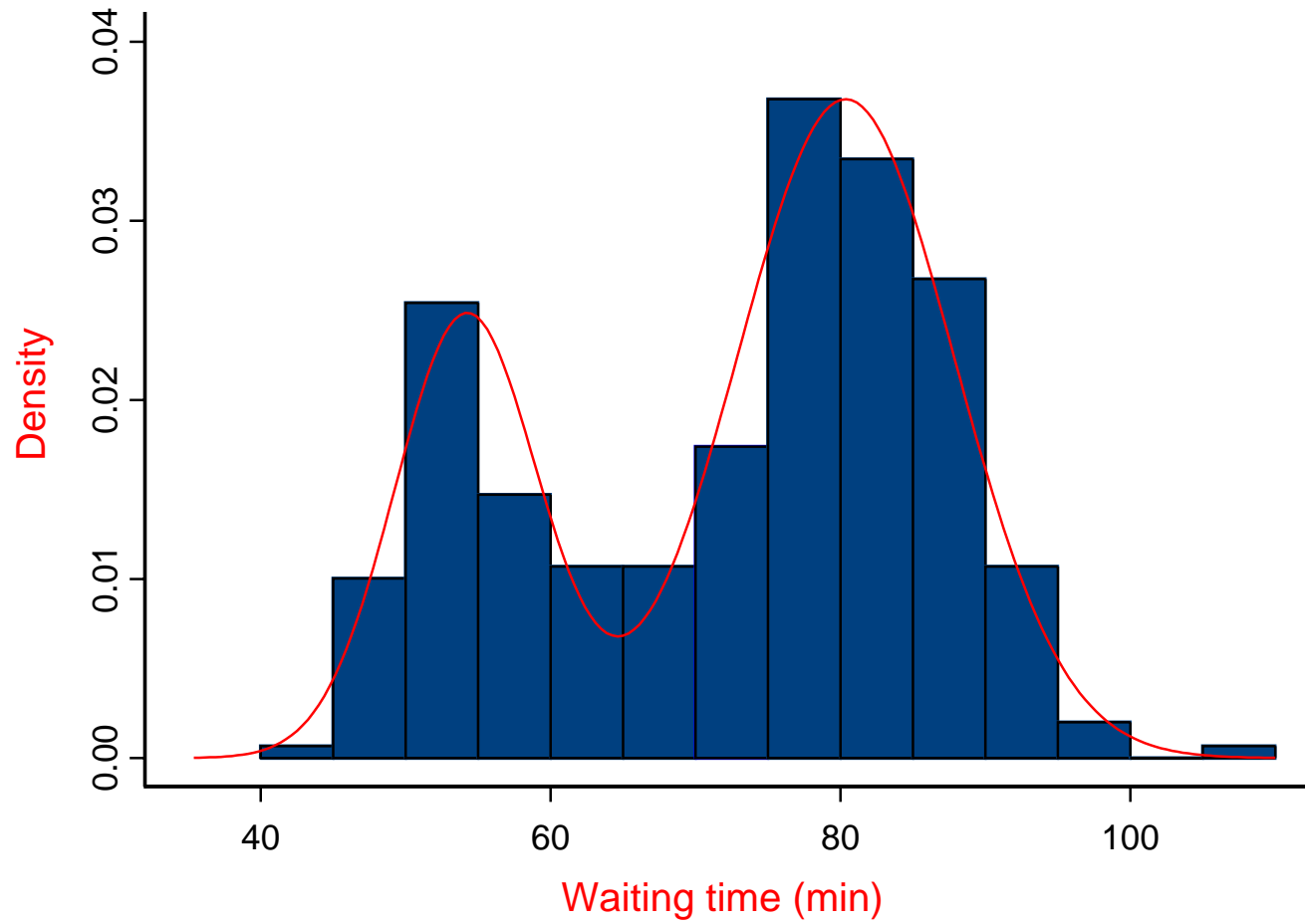
(similarly for $\mu_0^{(k+1)}$ & $[\sigma_0^{(k+1)}]^2$)

- Yellowstone Geyser (Waiting Time Between Eruptions)**

$n = 299$, sample mean = 72.3144 min, standard dev. = 13.8671 min

k	$w^{(k)}$	$\mu_0^{(k)}$	$\mu_1^{(k)}$	$\sigma_0^{(k)}$	$\sigma_1^{(k)}$	$\log L$
0	0.50000	65.3808	79.2479	12.0092	12.0092	-1208.30926
1	0.51087	64.8022	79.5069	13.0968	10.3160	-1200.61590
.
.
.
40	0.69238	54.2036	80.3611	4.9528	7.5069	-1157.54202
41	0.69238	54.2034	80.3609	4.9526	7.5071	-1157.54202
SE's	(0.03438)	(0.6831)	(0.6333)	(0.5183)	(0.5070)	

Waiting time between eruptions



(4) Maximum Likelihood Estimation for Mixtures

- **Point Estimation**

- “Nonstandard” case
- Likelihood can be infinite
- Maximize largest local likelihood

- **Tests of Significance**

- Likelihood ratio test (**LRT**)

No longer asymptotic chi-square distribution

Recalibrate (e.g., adjust degrees of freedom or resample)

- **Model Selection Criteria**

- Akaike's Information Criterion (**AIC**) or Bayesian Information Criterion (**BIC**) still appear to be valid

Choose model for which AIC (or BIC) is minimum

$$\text{AIC} = -2 \log L + 2p$$

$$\text{BIC} = -2 \log L + p \log n$$

where p is number of parameters estimated,

L is maximized likelihood function,

n is sample size

-- Chico Precipitation Intensity ($n = 787$)

Model	p	Log Likelihood	AIC	BIC
Single exponential	1	-2827.089	5656.177	5660.846
Mixture of two exponentials	3	-2806.589	5619.178	5633.183

-- Yellowstone Geyser Waiting Time ($n = 299$)

Model	p	Log Likelihood	AIC	BIC
Single normal	2	-1210.488	2424.977	2432.378
Mixture of two normals	5	-1157.542	2325.084	2343.586