# Motivation

In the field of fisheries stock assessment, biologists model fish populations to estimate sustainable levels of fishing mortality. One important component of these models is the age structure of the fish population, which can be estimated by taking a sample of fish from the population and counting growth layers on a small ear bone called the otolith. This is similar to ageing a tree by counting growth rings.

This is a time consuming process and thus relatively expensive. It is easier to measure fish lengths; however, due to individual variation in growth rates, fish length does not correspond perfectly to fish age. We can consider observed fish lengths as resulting from a mixture of length distributions that correspond to age cohorts.

Density

0.020

### 0 20 40 60 80 100

0.000

Fish Length (cm)

**Problem**

Given a set of observations of fish lengths spanning three age cohorts, estimate the parameters of the Gaussian (normal) distributions of length-at-age, when the age of each length observation is not known.

The probability density function of this mixture distribution is

*f* (*x*) = *λ*1*φ*(*x*|*µ*1*, σ*2) + *λ*2*φ*(*x*|*µ*2*, σ*2) + *λ*3*φ*(*x*|*µ*3*, σ*2)*,*

1

2

3

with *λk* being the probability of an observation coming from component *k* and *φ* is a normal distribution with mean *µk* and variance *σ*2.

*k*

If we knew the component that each observation came from, we would use maximum likelihood to estimate mean length *µ*ˆ*k* and standard deviation *σ*ˆ*k* of each of the *k* ages. We would also know the proportion of observations coming from each age cohort *λk*. However, we do not know which component of the distribution generated each observation (the generation process is “hidden” or “latent”). The key to the estimation problem is therefore to **assign** labels (ages) to observations by some process, and subsequently perform maximum likelihood estimation of *µ*ˆ*k* and *σ*ˆ*k*. This assignment and estimation is performed repeatedly until the parameter estimates cease changing.

# Algorithm to perform the estimation

The expectation maximization (EM) algorithm is the tool we will use for this task. It consists of initialization followed by iteration over expectation and maximization, with test for convergence determining when iteration can stop.

## Background reading

Do, C.B. and S. Batzoglou. 2008. [What is the expectation maximization algorithm?](http://www.nature.com/nbt/journal/v26/n8/full/nbt1406.html) Nature Biotechnology 26:897-899.

## Initialization

1. Using data from the fish of known age as a starting point, assign labels (ages) to each of *N* length observations in the dataset that come from fish of unknown age.
2. Given these initial assignments, compute initial values of *µ*ˆ*k* and *σ*ˆ*k* for each age class *k* = 1*,* 2*,* 3.
3. Compute initial values of *λ*ˆ*k* as *nk/N* .

## Expectation

Calculate the probability that length observation *xi* belongs to component *k* using Bayes Rule:

*P* (*x*

∈ *k*|*x* ) = *P* (*xi*|*xi* ∈ *k*)*P* (*k*) *.*

*i i P* (*xi*)

This *posterior probability* is computed for each observation *xi* based upon our initial guesses about the model parameters. *P* ( ) represents the probability of the event described inside the parenthesis and *xi k* is the event that observation *xi* comes from component *k*.

* ∈

How can we calculate the probability that observation *xi* belongs to age class *k*? If we assume each component of the mixture is Gaussian, then these probabilities are derived from the normal distribution with parameters *µ*ˆ*k* and *σ*ˆ*k*. We compute these probabilities for each observation for each possible component that may have generated each *xi*.

To ensure *P* (*xi k xi*) is a probability, i.e., 0 *P* (*xi k xi*) 1, standardize *P* (*xi k xi*) by dividing by the sum of probabilities across all three components:

∈ | ≤ ∈ | ≤ ∈ |

3

Σ | ∈

*P* (*xi*) = *P* (*xi xi k*)*P* (*k*)

*k*=1

*Hint:* at this point you should have a matrix with rows for each length observation and columns for each age class, where the matrix components are the probability of membership in each age class.

## Maximization

Given the probabilities of membership estimated above, compute another iteration of the estimate *µ*ˆ*k* as

*N*

Σ=

*µ*ˆ *i*=1

Σ

*P* (*xi* ∈ *k*|*xi*)*xi .*

*k N*

*i*=1

*P* (*xi*

∈ *k*|*xi*)

Note that this is a weighted mean, with the weights being the probability of an observation belonging to each age class for each observation.

The new estimate of *σ*ˆ*k* is based on *µ*ˆ*k*:

= ‚., Σ*N*

*i*=1

Σ

*σ*ˆ

*.*

*P* (*xi* ∈ *k*|*xi*) (*xi* − *µ*ˆ*k*)2

Finally,

*k N*

*i*=1

*P* (*xi*

∈ *k*|*xi*)

*N*

1 Σ= (

*λ*ˆ *P x*

*N*

∈ *k*|*x* )*.*

## Likelihood evaluation

*k i i*

*i*=1

We can now calculate the likelihood of our data given these new values of *λ*ˆ, *µ*ˆ, and *σ*ˆ. The likelihood is

*K*

Σ*P* (*X*|*µ, σ, λ*) = *λ* N (*X*|*µ , σ* )*k k*

2

*k*

*k*=1

note N () refers to the normal distribution and *σ*2 refers to the variance (rather than standard deviation) of

*k*

the *kth* component of the mixture.

This likelihood is calculated at each iteration.

**Iteration and the stopping rule**

The process of expectation and maximisation is repeated until the likelihood of data given the estimated parameters in the model changes negligibly. The magnitude of the change in successive likelihoods forms our criterion to determine when iteration should cease. Because likelihoods can become vanishingly small (and lead to underflow problems), it is easier to work with log-likelihoods:

ln(

) = ln( (

)) = Σ

ln ΣΣ*K* (

2)Σ

L

Convergence is reached when

*P X*|*µ, σ, λ*

*n*=1

*N*

*k*=1

*λk*N

*xn*|*µk, σk*

ln(L*i*+1) − ln(L*i*) *< s*

where ln(L*)*+*∞*) is the log-likelihood from the *ith* iteration of the EM algorithm and *s* is a desired tolerance.

## 