MODELS FOR TAGGING AND BANDING OF ANIMAL POPULATIONS

There is a large class of models for the analysis of bird banding (Europeans refer to this as bird ringing) or fish tagging data. Typically a number of animals are marked each year at about the same time of the year and during a relatively short period of time (3-4 weeks, perhaps). These animals are subject to harvest following banding and some proportion of the harvested individuals have their band or tag reported to a central location or agency. The harvest could be over a short time period or extend over nearly all of the year. The purpose of these sampling programs to most often to estimate the probability of survival.

The periods do not have to be years nor do the periods have to be of equal length. The primary reference to this class of models is Brownie et al. (1985). Program *MARK* allows a very general analytical approach to the estimation of parameters in this setting.

Known Constants:

R_i The number of animals banded or tagged in year i.

Random Variables:

m_{ij} The number of bands or tags reported in year j from releases in year i A matrix of recoveries.

The word "encountered" is a useful, general term for recovered, reported, resighted, sampled, etc.

Parameters:

S_j Conditional probability of survival in year j, <u>given</u> the animal is alive at the beginning of year j.

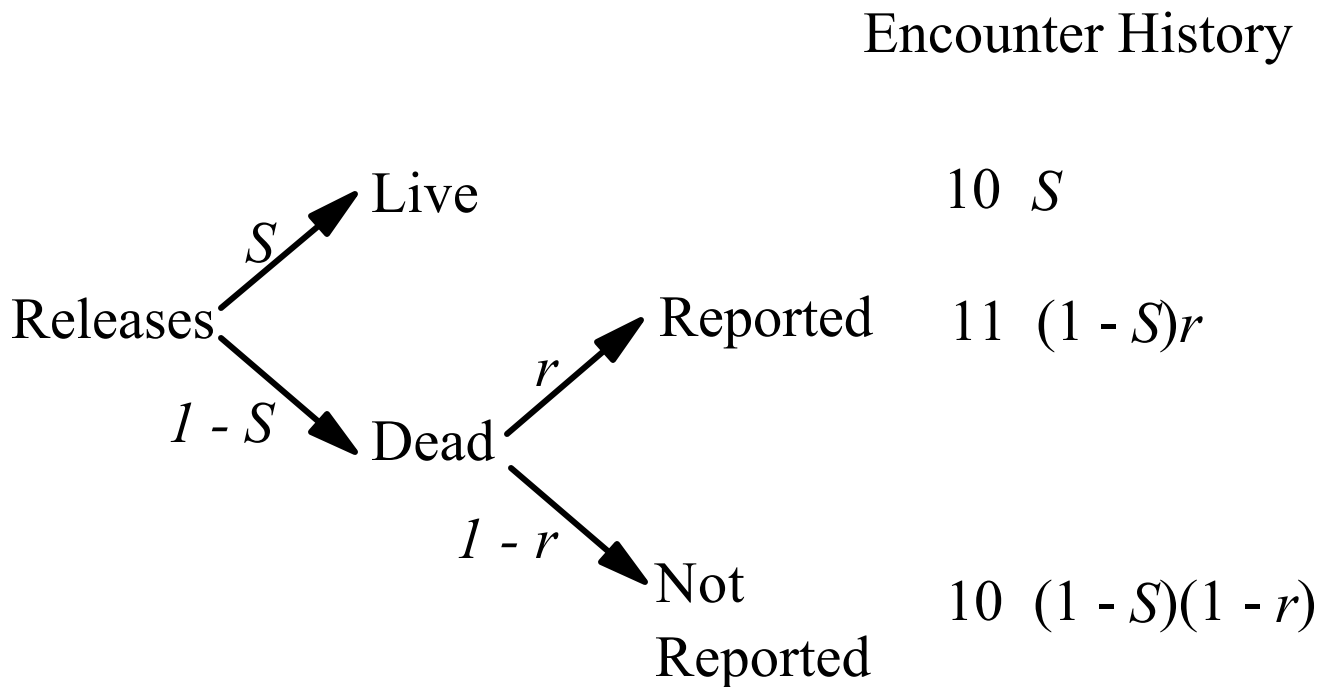S_j Relates to the interval between banding/tagging periods!

r_j Conditional probability of being reported in year j, <u>given</u> the fish died in year j.

Program *SURVIV* and publications such as Brownie et al. (1985) used f_j as a "recovery probability." Here, we use a "reporting probability" r_j,

$$r_j = \frac{f_j}{(1-S_j)} \; ,$$

This parameterization allows banding or tagging data and capture-recapture data to be modeled under a common framework and allows the recovery process to be separated from the survival process.

Tag Recovery Data Can be Viewed as;

Encounter History

Live   10 S

Releases

    S (Live)

    $1 - S$ (Dead)

Reported   11 $(1 - S)r$

    r

Dead

    $1 - r$

Not Reported   10 $(1 - S)(1 - r)$

Why Model?

The model links, in a formal manner, the data $\{R_i, m_{ij}\}$, assumptions, unknown parameters (the S_j and r_j) and allows rigor in making inductive inferences via likelihood and information theory.

Models are an essential component of science.

Problems occur if an estimate $\hat{S}_j > 1$; this is a diagnostic indicating an over-parameterized model. The model assuming the survival and reporting probabilities $\{S, r\}$ are constant always have estimates in the $\{0, 1\}$ range.

Note, of course, $S = 1 - M$ where M is total mortality rate; a finite rate (not instantaneous).

Modeling band recovery data involves *reparameterizing* the multinomial likelihood and log-likelihood. First, we look at the structure of the parameters that might be hypothesized to underlying recovery data that we observe.

Consider the small data set on (simulated) striped bass –

i	R_i	m_{ij}				
		$j=1$	2	3	4	5
1	2000	30	70	114	**43**	15
2	2000		80	97	55	19
3	2000			169	46	10
4	2000				72	24
5	2000					34

Notice that 43 fish tagged in year 1 were harvested by anglers and reported in year 4; m_{14}, shown in bold.

Each row of the table (recovery matrix) is a multinomial distribution (under certain assumptions). Animals banded or tagged in year 1 are reported in year 1, 2, ..., 5 or "never."

The **Saturated Model** has as many parameters as there are cells. Here the subscripts i and j denote year-specific parameters (the j), specific to each banded cohort (the i).

Number Tagged   **Expected Number of Tags Reported (dead)**

	$j=1$	2	3	4	5
R_1	$R_1\theta_{11}$	$R_1\theta_{12}$	$R_1\theta_{13}$	$R_1\theta_{14}$	$R_1\theta_{15}$
R_2		$R_2\theta_{22}$	$R_2\theta_{23}$	$R_2\theta_{24}$	$R_2\theta_{25}$
R_3			$R_3\theta_{33}$	$R_3\theta_{34}$	$R_3\theta_{35}$
R_4				$R_4\theta_{44}$	$R_4\theta_{45}$
R_5					$R_5\theta_{55}$

Thus, $K = 15$ as there are 15 cells.

Model $\{S_t, r_t\}$ Here the subscript t denotes year-specific parameters.

Number Tagged   **Expected Number of Tags Reported (dead)**

	$j=1$	2	3	4	5
R_1	$R_1(1\text{-}S_1)r_1$	$R_1S_1(1\text{-}S_2)r_2$	$R_1S_1S_2(1\text{-}S_3)r_3$	$\mathbf{R_1S_1S_2S_3(1\text{-}S_4)r_4}$	$R_1S_1S_2S_3S_4(1\text{-}S_5)r_5$
R_2		$R_2(1\text{-}S_2)r_2$	$R_2S_2(1\text{-}S_3)r_3$	$R_2S_2S_3(1\text{-}S_4)r_4$	$R_2S_2S_3S_4(1\text{-}S_5)r_5$
R_3			$R_3(1\text{-}S_3)r_3$	$R_3S_3(1\text{-}S_4)r_4$	$R_3S_3S_4(1\text{-}S_5)r_5$
R_4				$R_4(1\text{-}S_4)r_4$	$R_4S_4(1\text{-}S_5)r_5$
R_5					$R_5(1\text{-}S_5)r_5$

Model $\{S_t, r\}$ Note the r appears without a subscript, meaning this parameter is constant.

Number Tagged   **Expected Number of Tags Reported (dead)**

	$j=1$	2	3	4	5
R_1	$R_1(1\text{-}S_1)r$	$R_1S_1(1\text{-}S_2)r$	$R_1S_1S_2(1\text{-}S_3)r$	$\mathbf{R_1S_1S_2S_3(1\text{-}S_4)r}$	$R_1S_1S_2S_3S_4(1\text{-}S_5)r$
R_2		$R_2(1\text{-}S_2)r$	$R_2S_2(1\text{-}S_3)r$	$R_2S_2S_3(1\text{-}S_4)r$	$R_2S_2S_3S_4(1\text{-}S_5)r$
R_3			$R_3(1\text{-}S_3)r$	$R_3S_3(1\text{-}S_4)r$	$R_3S_3S_4(1\text{-}S_5)r$
R_4				$R_4(1\text{-}S_4)r$	$R_4S_4(1\text{-}S_5)r$
R_5					$R_5(1\text{-}S_5)r$

Model $\{S, r_t\}$

In this case, survival is constant, but reporting probability is time-specific.

Number Tagged | **Expected Number of Tags Reported (dead)**

Number Tagged	$j=1$	2	3	4	5
R_1	$R_1(1-S)r_1$	$R_1 S(1-S)r_2$	$R_1 SS(1-S)r_3$	$\mathbf{R_1 SSS(1-S)r_4}$	$R_1 SSSS(1-S)r_5$
R_2		$R_2(1-S)r_2$	$R_2 S(1-S)r_3$	$R_2 SS(1-S)r_4$	$R_2 SSS(1-S)r_5$
R_3			$R_3(1-S)r_3$	$R_3 S(1-S)r_4$	$R_3 SS(1-S)r_5$
R_4				$R_4(1-S)r_4$	$R_4 S(1-S)r_5$
R_5					$R_5(1-S)r_5$

Model $\{S, r\}$

Note, here neither parameter is subscripted, implying they are constants.

Number Tagged | **Expected Number of Tags Reported (dead)**

Number Tagged	$j=1$	2	3	4	5
R_1	$R_1(1-S)r$	$R_1 S(1-S)r$	$R_1 SS(1-S)r$	$\mathbf{R_1 SSS(1-S)r}$	$R_1 SSSS(1-S)r$
R_2		$R_2(1-S)r$	$R_2 S(1-S)r$	$R_2 SS(1-S)r$	$R_2 SSS(1-S)r$
R_3			$R_3(1-S)r$	$R_3 S(1-S)r$	$R_3 SS(1-S)r$
R_4				$R_4(1-S)r$	$R_4 S(1-S)r$
R_5					$R_5(1-S)r$

Model $\{S, r, r_3\}$ This notation denotes that S and r are constant, except for r in year 3.

Number Tagged   **Expected Number of Tags Reported (dead)**

	$j=1$	2	3	4	5
R_1	$R_1(1-S)r$	$R_1S(1-S)r$	$R_1SS(1-S)r_3$	$\mathbf{R_1SSS(1-S)r}$	$R_1SSSS(1-S)r$
R_2		$R_2(1-S)r$	$R_2S(1-S)r_3$	$R_2SS(1-S)r$	$R_2SSS(1-S)r$
R_3			$R_3(1-S)r_3$	$R_3S(1-S)r$	$R_3SS(1-S)r$
R_4				$R_4(1-S)r$	$R_4S(1-S)r$
R_5					$R_5(1-S)r$

WHY? Reporting probability in year 3 hypothesized to be very different than in normal years.

The striped bass could have been shown with the number never recovered shown:

i	R_i	m_{ij}					
		$j=1$	2	3	4	5	never
1	2000	30	70	114	**43**	15	1728
2	2000		80	97	55	19	**1749**
3	2000			169	46	10	1775
4	2000				72	24	1904
5	2000					34	1966

The number never recovered from tagging in year 2 is 1749 and its expectation under Model $\{S_t, r_t\}$ is:

$$R_2 - \left(R_2(1-S_2)r_2 + R_2S_2(1-S_3)r_3 + R_2S_2S_3(1-S_4)r_4 + R_2S_2S_3S_4(1-S_5)r_5 \right)$$

or

$$2000 - (80 + 97 + 55 + 19) = 1749.$$

The expected number never recovered varies by model; e.g., for Model $\{S, r\}$ is

$$R_2 - \left(R_2(1-S)r + R_2S(1-S)r + R_2SS(1-S)r + R_2SSS(1-S)r \right)$$

and this could be computed given the parameters S and r.

Program *MARK* computes these values, however, one must be aware of this additional cell in the multinomial distribution.

For modeling in a likelihood framework, we will need probabilities. These can be obtained by merely dividing the expected cell values by the number banded/tagged.

The **Saturated Model**

Number Tagged		Matrix of Cell Probabilities for m_{ij}		$E(m_{ij}/R_i)$	
	$j=1$	2	3	4	5
R_1	θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}
R_2		θ_{22}	θ_{23}	θ_{24}	θ_{25}
R_3			θ_{33}	θ_{34}	θ_{35}
R_4				θ_{44}	θ_{45}
R_5					θ_{55}

None of these K parameters are of any biological interest. The saturated model serves as a basis for judging fit of other models. The saturated model fits the data perfectly; no other model will fit as well. The **deviance** of the saturated model is zero.

Thus, $K = 15$ as there are 15 cells.

Model $\{S_t, r_t\}$

Number Tagged	Matrix of Cell Probabilities for m_{ij} $E(m_{ij}/R_i)$				
R_1	$(1-S_1)r_1$	$S_1(1-S_2)r_2$	$S_1S_2(1-S_3)r_3$	$S_1S_2S_3(1-S_4)r_4$	$S_1S_2S_3S_4(1-S_5)r_5$
R_2		$(1-S_2)r_2$	$S_2(1-S_3)r_3$	$S_2S_3(1-S_4)r_4$	$S_2S_3S_4(1-S_5)r_5$
R_3			$(1-S_3)r_3$	$S_3(1-S_4)r_4$	$S_3S_4(1-S_5)r_5$
R_4				$(1-S_4)r_4$	$S_4(1-S_5)r_5$
R_5					$(1-S_5)r_5$

Model $\{S, r\}$

Number Tagged	Matrix of Cell Probabilities for m_{ij}					$E(m_{ij}/R_i)$
R_1	$(1\text{-}S)r$	$S(1\text{-}S)r$	$SS(1\text{-}S)r$	$SSS(1\text{-}S)r$	$SSSS(1\text{-}S)r$	$(1\text{-}(1\text{-}S^5)r)$
R_2		$(1\text{-}S)r$	$S(1\text{-}S)r$	$SS(1\text{-}S)r$	$SSS(1\text{-}S)r$	$(1\text{-}(1\text{-}S^4)r)$
R_3			$(1\text{-}S)r$	$S(1\text{-}S)r$	$SS(1\text{-}S)r$	$(1\text{-}(1\text{-}S^3)r)$
R_4				$(1\text{-}S)r$	$S(1\text{-}S)r$	$(1\text{-}(1\text{-}S^2)r)$
R_5					$(1\text{-}S)r$	$(1\text{-}(1\text{-}S)r)$

We will switch back and forth between expected values and probabilities, so people need to get familiar with both expressions.

Consider the $\log_e \mathcal{L}(S, r \mid R_1, m_{1j})$ for the first year of release of tagged bass. This is the log of the likelihood of the parameters (the constant survival probabilities S and r), given the data from tagging in year 1 (the R_1 and m_{1j})

Consider the original data set on (simulated) striped bass and the first row of the m_{ij} matrix;

i	R_i	$j=1$	2	3	4	5
1	2000	30	70	114	43	15
2	2000		80	97	55	19
3	2000			169	46	10
4	2000				72	24
5	2000					34

The cell probabilities under Model $\{S, r\}$ are:

Number Tagged	Matrix of Cell Probabilities for m_{ij}					$E(m_{ij}/R_i)$
R_1	$(1\text{-}S)r$	$S(1\text{-}S)r$	$SS(1\text{-}S)r$	$SSS(1\text{-}S)r$	$SSSS(1\text{-}S)r$	$(1\text{-}(1\text{-}S^5)r)$
R_2		$(1\text{-}S)r$	$S(1\text{-}S)r$	$SS(1\text{-}S)r$	$SSS(1\text{-}S)r$	$(1\text{-}(1\text{-}S^4)r)$
R_3			$(1\text{-}S)r$	$S(1\text{-}S)r$	$SS(1\text{-}S)r$	$(1\text{-}(1\text{-}S^3)r)$
R_4				$(1\text{-}S)r$	$S(1\text{-}S)r$	$(1\text{-}(1\text{-}S^2)r)$
R_5					$(1\text{-}S)r$	$(1\text{-}(1\text{-}S)r)$

The the log-likelihood of the parameters S and r, *given* the data (the number released in year 1 and the number of tag recoveries reported in year j **from those released in year 1).**

$$\log_e \mathcal{L}(S, r \mid R_1, m_{1j}) = m_{11}\log\left((1\text{-}S)r\right) + m_{12}\log\left(S(1\text{-}S)r\right) + m_{13}\log\left(SS(1\text{-}S)r\right) +$$

$$m_{14}\log\left(SSS(1\text{-}S)r\right) + m_{15}\log\left(SSSS(1\text{-}S)r\right)$$

$$+ \left(\text{a term for those never reported, } MARK \text{ handles this}\right)$$

The log-likelihood function for the tag recoveries from releases in year 2 is similar,

$$\log_e \mathcal{L}(S, r \mid R_2, m_{2j}) = m_{22}\log\Big((1\text{-}S)r\Big) + m_{23}\log\Big(S(1\text{-}S)r\Big) + m_{24}\log\Big(SS(1\text{-}S)r\Big) +$$

$$m_{25}\log\Big(SSS(1\text{-}S)r\Big) + \text{"never"}$$

The log-likelihood functions for the tag recoveries from releases in years 3, 4 and 5 are

$$\log_e \mathcal{L}(S, r \mid R_3, m_{3j}) = m_{33}\log\Big((1\text{-}S)r\Big) + m_{34}\log\Big(S(1\text{-}S)r\Big) + m_{35}\log\Big(SS(1\text{-}S)r\Big) + \text{"never"}$$

$$\log_e \mathcal{L}(S, r \mid R_4, m_{4j}) = m_{44}\log\Big((1\text{-}S)r\Big) + m_{45}\log\Big(S(1\text{-}S)r\Big) + \text{"never"}$$

$$\log_e \mathcal{L}(S, r \mid R_5, m_{5j}) = m_{55}\log\Big((1\text{-}S)r\Big) + \text{"never"}$$

Example, the log-likelihood for releases in year 3:

$$\log_e \mathcal{L}(S, r \mid 2000, 169, 46, 10) = 169\log\Big((1\text{-}S)r\Big) + 46\log\Big(S(1\text{-}S)r\Big) + 10\log\Big(SS(1\text{-}S)r\Big)$$

$$+ \textbf{(2000-169-46-10)log}\left(\textbf{1--}\left[\Big(\textbf{(1-S)r}\Big) + \Big(\textbf{S(1-S)r}\Big) + \Big(\textbf{SS(1-S)r}\Big)\right]\right)$$

This function assumes the "data" are given; indeed, the data here include the 2000 fish tagged, and the number of recoveries in years 3, 4, and 5. Only the parameters are unknown and the objects of interest. The log-likelihood function here is a function only of S and r.

Note the final term for those fish <u>never</u> <u>reported</u> from the 2,000 released is shown explicitly in the final expression (above, in bold blue).

The total log-likelihood for all the recovery data for the 5 years of release is merely the sum of the individual log-likelihoods (assuming independence of tagged cohorts),

$$\log_e \mathcal{L}(S, r \mid R_i, m_{ij}) = \log_e(\mathcal{L}_1) + \log_e(\mathcal{L}_2) + \ldots + \log_e(\mathcal{L}_5).$$

Thus,

$$\log_e \mathcal{L}(S, r \mid R_i, m_{ij}) = m_{11}\log\Big((1\text{-}S)r\Big) + m_{12}\log\Big(S(1\text{-}S)r\Big) + m_{13}\log\Big(SS(1\text{-}S)r\Big) +$$

$$m_{14}\log\Big(SSS(1\text{-}S)r\Big) + m_{15}\log\Big(SSSS(1\text{-}S)r\Big)$$

$$+$$

$$m_{22}\log\Big((1\text{-}S)r\Big) + m_{23}\log\Big(S(1\text{-}S)r\Big) + m_{24}\log\Big(SS(1\text{-}S)r\Big) + m_{25}\log\Big(SSS(1\text{-}S)r\Big)$$

$$+$$

$$m_{33}\log\Big((1\text{-}S)r\Big) + m_{34}\log\Big(S(1\text{-}S)r\Big) + m_{35}\log\Big(SS(1\text{-}S)r\Big)$$

$$+$$

$$m_{44}\log\Big((1\text{-}S)r\Big) + m_{45}\log\Big(S(1\text{-}S)r\Big)$$

$$+$$

$$m_{55}\log\Big((1\text{-}S)r\Big) \qquad \text{(plus a complex "never recovered" term).}$$

Note, each term involving parameters in the log-likelihood function is of the form

---

DATA * LOG(PROBABILITY).

The "DATA" are the R_i (number tagged in year i) and the m_{ij} (number of tags recovered in year j from releases in year i). The "PROBABILITY" of observing the data is some expression of the unknown, underlying *parameters*, given a particular *model* (the S_j and the r_j). These are often called "cell probabilities."

Little complications that are important (trust *MARK*):

1. A multinomial cell must be included for the fish in each released cohort that are "never" recovered.

2. The multinomial coefficient $\binom{n}{y_i}$ is shorthand for

$$n! / \Big((y_1)! \, (y_2)! \, \cdots \, (y_k)! \Big),$$

This term does not involve any of the unknown parameters and is ignored for many estimation issues. *MARK* handles this issue.

In the striped bass tagging we have 5 multinomial coefficients in the overall likelihood; each of the form

$$\begin{pmatrix} R_i \\ m_{ij} \end{pmatrix} , \text{ where } i = 1, ..., 5.$$

For $i = 1$,

$$R_1! \Big/ \Big((m_{11})! \, (m_{12})! \, \cdot \, \cdot \, \cdot \, (m_{15})! \, (R_1 - \sum m_{1j})! \Big)$$

or

$$2000! \Big/ \Big(30! \; 70! \; 114! \; 43! \; 15! \; (2000-272)! \Big) .$$

You can see why it is convenient to ignore this term! Program *DERIVE* does such things easily; we will get to it soon.

Some Questions:

1. What if some tags are taken from animals harvested in year j, but not reported to the agency until year $j+\kappa$? Thus, the analyst thinks the animal was harvest in a latter year? What is the likely effect of the estimators S_i?

2. What if the sample size R_i was actually made up of m groups of brood mates. What might one worry about (in terms of the multinomial model)?

3. What if you banded animals for 19 years and then stopped; however, sport harvest continued for many more years. What might parameters might you estimate? What might be expected about the precision of the estimators?

4. The total log-likelihood is a sum of the log-likelihoods for each of the 5 cohorts. But this is based on the notion that the cohorts are independent. Is this a good assumption? Why? Why not? When might it fail?

Why is Brownie et al. (1985) hard to follow? Why so much notation that seems not to be covered in lectures? Why is the log-likelihood function not more prominent?

When the first edition of the "Banding Analysis Handbook" was published (1978) the emphasis was on estimators that existed in "closed form." That is, the calculus was used to take first partial derivatives of the log-likelihood function, set these to zero, and solve the resulting set of equations to get computable estimators "formulae." The parameter being estimated was on the RHS and the data could be substituted for the notation on the LHS and a numerical value produced. For example, on the cover we see the estimator (in closed form),

$$\hat{f}_i = \frac{R_i C_i}{N_i T_i} \quad \text{for} \ \ i = 1, ..., k.$$

The 4 quantities on the LHS are *statistics* that come from the data and are known (although R_i, C_i, and T_i are observed random variables). A numerical value of the estimator (i.e., an estimate) is obtained by plugging in values for the 4 statistics and computing an estimate of f for each year i Easy enough, that is what one might expect. This was a useful approach and served the research and management community well during the 1970s and 1980s. There are problems with this approach:

1. Lots of notation is needed (see Chapters 2, 3, and 4). This is a bother to students trying to cope with the mass of symbols, subscripts, Greek characters, etc.

2. The notation has changed over the years as close relationships were found with the open capture-recapture models.

3. Many (most) models of real biological data do not have closed form estimators (they do not exist).

4. In some ways, the focus on the closed form estimators hinds the important concept of a log-likelihood and its parameterization, the notion of parameter values that maximize this function, etc.

5. While some estimators are "computable" and exist in "closed form" they are often tedious and error-prone if done by hand (e.g., Brownie et al. 1985:75-78). All such calculations are done by computer anyway, thus we are teaching FW663 with relatively little emphasis on closed form estimators and all the associated notation required. Our approach puts a premium on the numerical maximization of the log-likelihood and the shape of this function.