Model identification, parameter redundancy and exhaustive summaries

Byron Morgan

ncse
National Centre for Statistical Ecology

Collaborators

- Ted Catchpole (Canberra and Kent)
- Diana Cole (Kent)

Outline

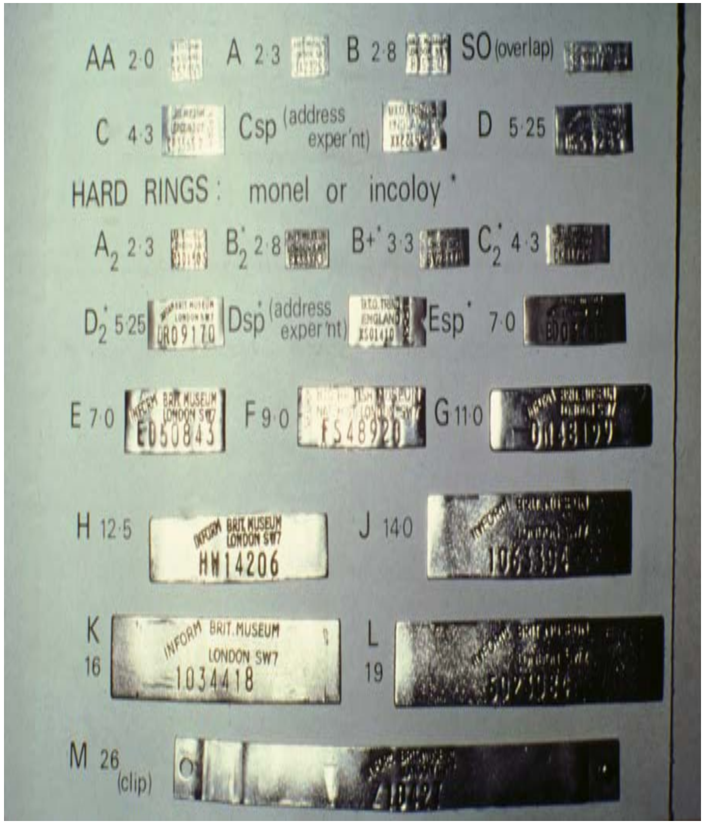
- Motivation
- CJS model (1965)
- Methods
- Covariates
- Fisheries example (2007)
- General rules
- Bayesian perspective
- New work and other areas
- References

Motivation

- Estimation of the annual survival probabilities of wild animals.
- Collect data on previously marked animals.
- These are either found dead or alive.
- Form probability models.
- Fit to data using maximum likelihood.

Models for survival: Marking

- We obtain information on survival from studying previously marked animals
- These may be observed again alive or dead.
- It is assumed that marking does not affect behaviour

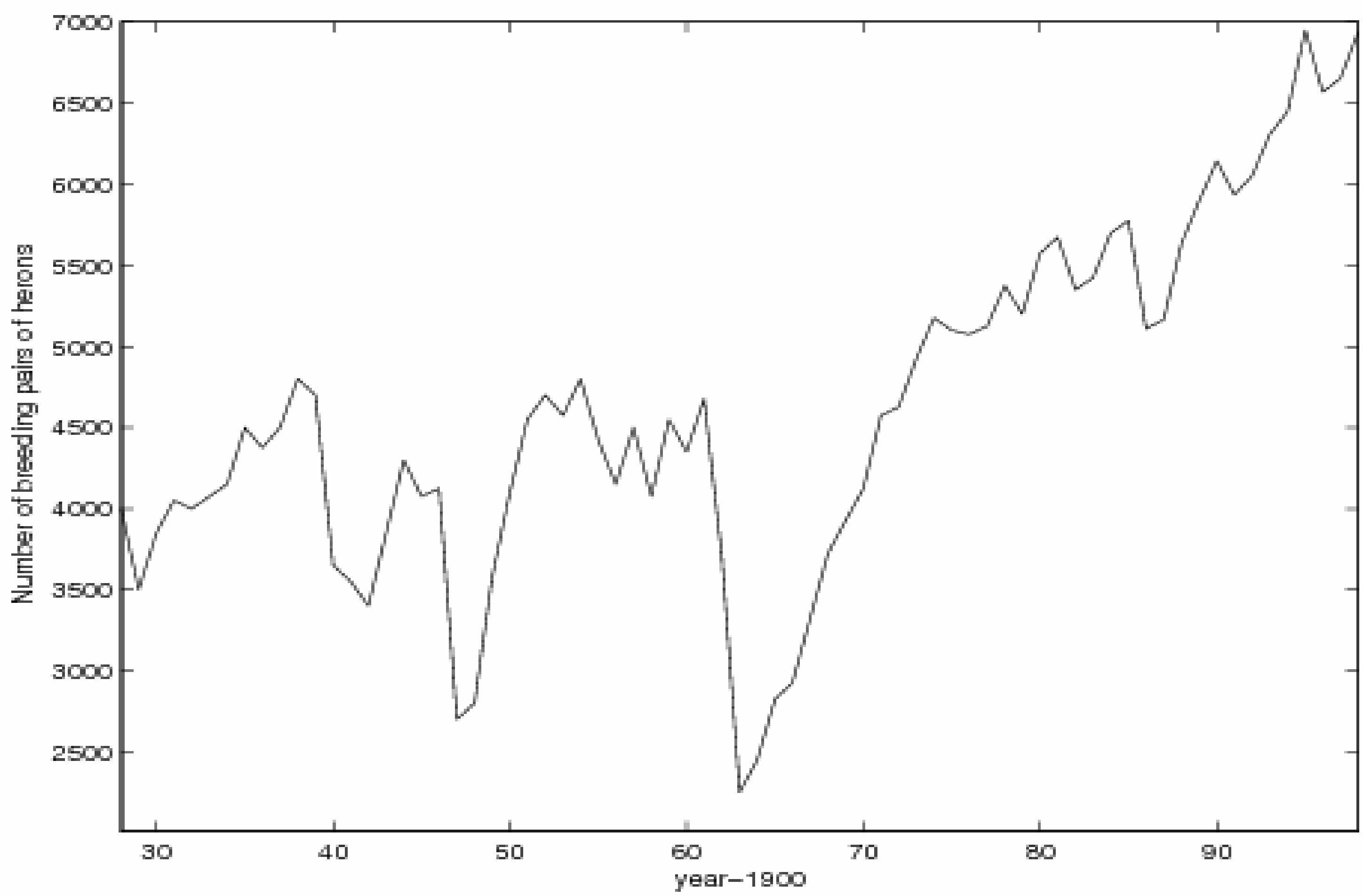


ncse
National Centre for Statistical Ecology

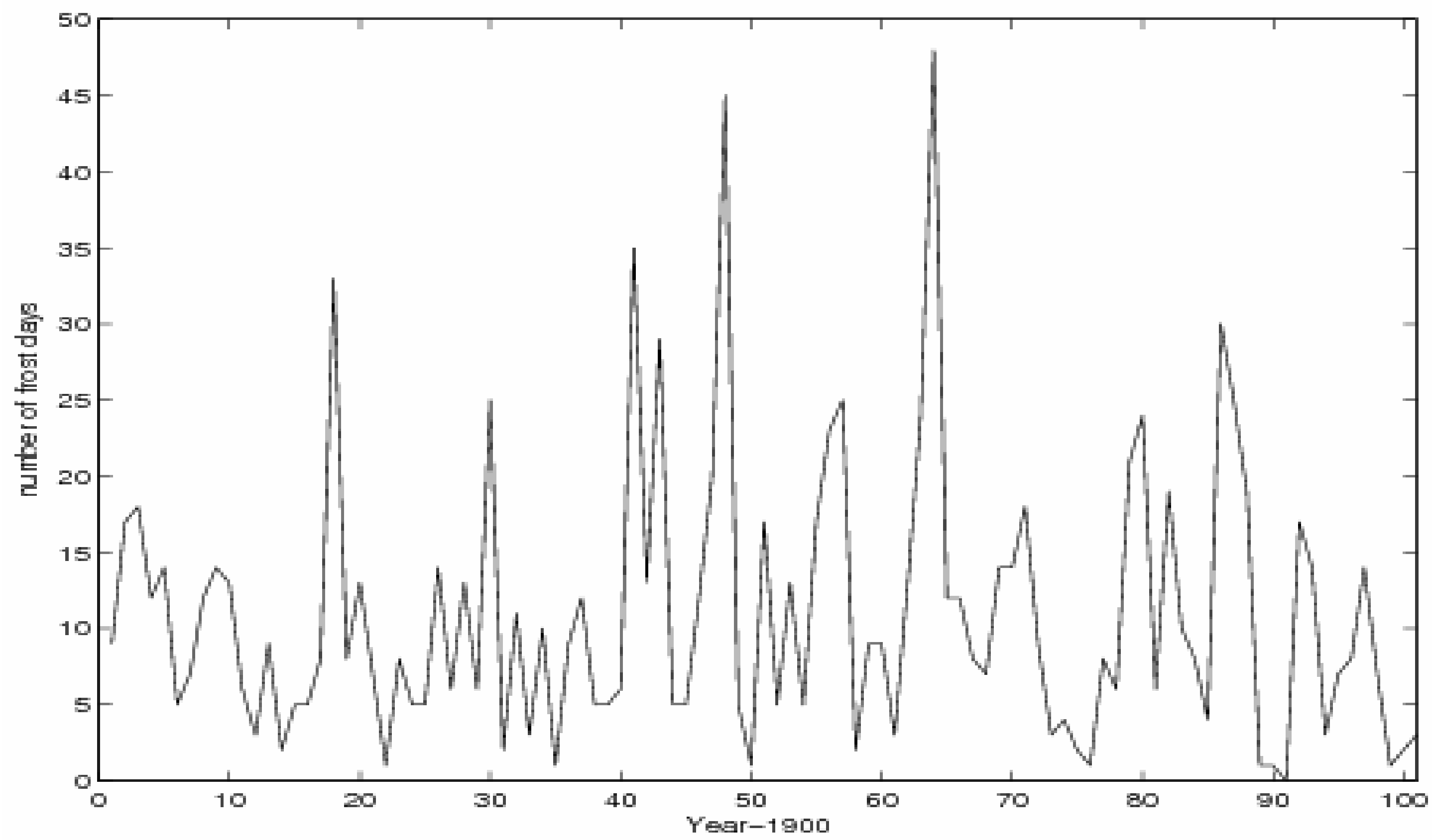
Complexity

- Models may be complicated, incorporating age, cohort and time components.
- Models may be simplified by the use of covariates.
- Modern focus on multi-site data can produce models with many parameters.
- It is often unclear how many parameters can be estimated.

The British heron census, *Ardea cinerea*

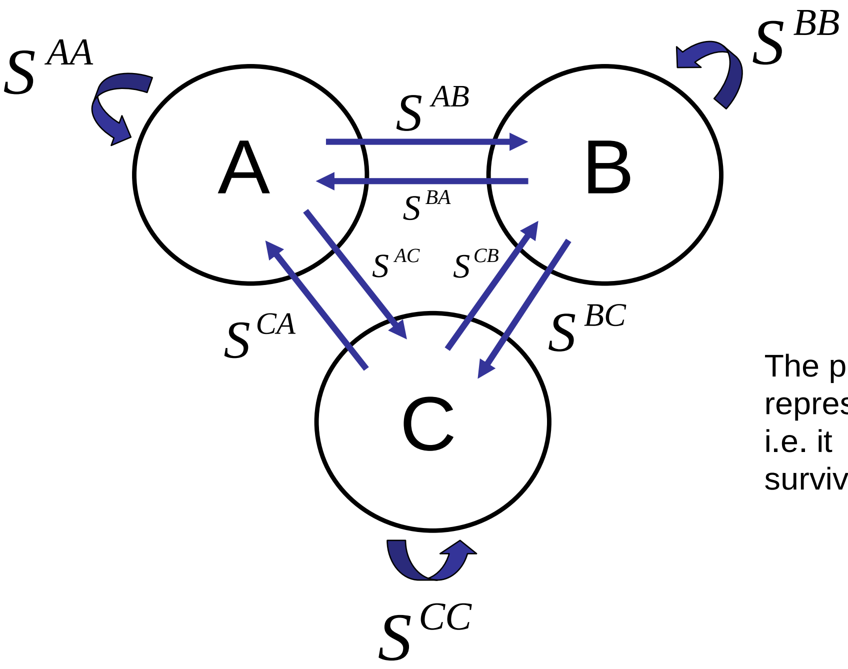


Climatic covariates: number of frost-days in Central England.



An example of a multi-site system

Multisite Systems



The parameter S represents the "transition", i.e. it represents both survival and movement

The Cormack-Jolly-Seber (CJS) model (1965)

Consider a simple case in which all animals are adults, sharing a common probability of annual survival, ϕ. If *p* denotes the probability of recapture then the multinomial probabilities corresponding to any cohort, of known size, of marked birds have the form:

ϕ p, ϕ^2 p(1-p), ϕ^3 p (1-p)^2 , ...

ncse
National Centre for Statistical Ecology

CJS model continued

If we allow each parameter to be time-varying, then there is a pair of parameters, $\phi_{t-1}p_t$, which only occur together. They are confounded, and so can only be estimated as a product when the likelihood, in this case a product of multinomials, one from each cohort, is maximised. All of the other parameters in the model can be estimated.

Illustration of CJS recapture probabilities: a 3-year study

$\phi_1\ p_2$	$\phi_1\ \phi_2\ (1-p_2)p_3$	$\phi_1\ \phi_2\ \phi_3(1-p_2)(1-p_3)p_4$
	$\phi_2\ p_3$	$\phi_2\ \phi_3(1-p_3)p_4$
		$\phi_3\ p_4$

ncse
National Centre for Statistical Ecology

CJS recapture probabilities: what we can estimate

$\phi_1\ p_2$	$\phi_1\ \phi_2\ (1-p_2)p_3$	$\phi_1\ \phi_2\ \phi_3(1-p_2)(1-p_3)p_4$
	$\phi_2\ p_3$	$\phi_2\ \phi_3(1-p_3)p_4$
		$\phi_3\ p_4$

ncse
National Centre for Statistical Ecology

Parameter redundancy

- This model has parameter redundancy of one: we can only estimate the product, $\phi_3 p_4$. All the other parameters can be estimated.
- What if we only have two years of ringing?

National Centre for Statistical Ecology

Illustration of CJS recapture probabilities: a 3-year study + 2 cohorts

$\phi_1\ p_2$	$\phi_1\ \phi_2\ (1-p_2)p_3$	$\phi_1\ \phi_2\ \phi_3(1-p_2)(1-p_3)p_4$
	$\phi_2\ p_3$	$\phi_2\ \phi_3(1-p_3)p_4$

ncse
National Centre for Statistical Ecology

Parameter redundancy and identifiability

- A model is identifiable if no two values of the parameters give the same probability distribution for the data.
- A model is locally identifiable if there is a distance $\delta > 0$, such that any two parameter values that give the same distribution must be separated by at least δ.
- A parameter redundant model is not locally identifiable.
- An essentially full rank model is locally identifiable.
- Are essentially full rank models identifiable?

ncse
National Centre for Statistical Ecology

How to test for parameter redundancy

- Form an appropriate derivative matrix, D.
- Use Maple to determine the symbolic row rank of D.
- We can also determine which parameter combinations can be estimated.
- We use expansion theorems to demonstrate that results hold for model structures of different sizes.

The method

The approach is for exponential family models. It is performed using a symbolic algebra package such as Maple.

1. Calculate **D** = $\left[\dfrac{\partial \mu_j}{\partial \theta_i}\right]$ (μ is the mean, θ are parameters).

2. The number of estimable parameters = rank(**D**).

3. Solve α^T**D** = 0. The location of the zeros in α indicates which are the estimable parameters.

4. Solve $\sum_{i=1}^{p} \alpha_{ij} \dfrac{\partial f}{\partial \theta_i} = 0$ to find the full set of estimable parameters; (j is the index for >1 solution to α^T**D** = 0).

5. Perform a modified PLUR decomposition of **D**.

Example 1: Cormack-Jolly-Seber Model

Little Penguins, *Eudyptula minor*, capture recapture data (1994 to 1997)

$$\mathbf{N} = \begin{bmatrix} 30 & 58 & 37 \\ 0 & 20 & 37 \\ 0 & 0 & 18 \end{bmatrix}$$

ϕ_i – probability a penguin survives from occasion i to $i+1$

p_i – probability a penguin is recaptured on occasion i

The set of parameters is: $\boldsymbol{\theta} = [\phi_1, \phi_2, \phi_3, p_2, p_3, p_4]$

$$\mathbf{P} = \begin{bmatrix} \phi_1 p_2 & \phi_1 \overline{p}_2 \phi_2 p_3 & \phi_1 \overline{p}_2 \phi_2 \overline{p}_3 \phi_3 p_4 \\ 0 & \phi_2 p_3 & \phi_2 \overline{p}_3 \phi_3 p_4 \\ 0 & 0 & \phi_3 p_4 \end{bmatrix}$$

$\overline{p}_2 = 1 - p_2$ etc

Forming the derivative matrix

$$\mathbf{D} = \frac{\partial \ln(\mathbf{P})}{\partial \theta} = \begin{bmatrix} \phi_1^{-1} & \phi_1^{-1} & \phi_1^{-1} & 0 & 0 & 0 \\ 0 & \phi_2^{-1} & \phi_2^{-1} & \phi_2^{-1} & \phi_2^{-1} & 0 \\ 0 & 0 & \phi_3^{-1} & 0 & \phi_3^{-1} & \phi_3^{-1} \\ p_2^{-1} & -\bar{p}_2^{-1} & -\bar{p}_2^{-1} & 0 & 0 & 0 \\ 0 & p_3^{-1} & -\bar{p}_3^{-1} & p_3^{-1} & -\bar{p}_3^{-1} & 0 \\ 0 & 0 & p_4^{-1} & 0 & p_4^{-1} & p_4^{-1} \end{bmatrix}$$

rank(**D**) = 5 < 6, so the model is parameter redundant.

In order to see which of the original parameters we can estimate:

Set $\alpha^T \mathbf{D} = 0 \Rightarrow \alpha^T = [\, 0, 0, -\phi_3 / p_4, 0, 0, 1]$

Solving PDE, we find that the estimable parameters are: ϕ_1, ϕ_2, p_2, p_3, $\phi_3 p_4$

National Centre for Statistical Ecology

Use of the PLUR decomposition

If parameter redundant:

Solve $\alpha^T \mathbf{D} = 0$. Zeros in α indicate estimable parameters.

Solve $\sum_{i=1}^{p} \alpha_{ij} \frac{\partial f}{\partial \theta_i} = 0$ to find full set of estimable parameters.

If full rank:

Determine whether essential ($\forall \theta$) or conditionally ($\exists \theta$) full rank using the PLUR decomposition.

$\mathbf{D} = \mathbf{PLUR}$. If det($\mathbf{U}$) = 0, model is parameter redundant. If det($\mathbf{U}$) is close to 0 model is near parameter redundant.

Example – Cormack-Jolly-Seber Model with covariates

We now set

$\phi_i = 1/\{1+\exp(\ a + bx_i\ )\}$

For example, x_i could be the mean annual banding weight, or the SOI.

$\theta = [a, b, p_2, p_3, p_4]$,

and we find that the model is now full rank.

Use of the PLUR decomposition

We have **D=PLUR.**

We find that

$$\mathrm{Det}(\mathbf{U}) = \frac{-(x_1 - x_2)(1 - p_2)\,p_3 p_4 \exp(a + bx_1)\exp(a + bx_2)}{\{1 + \exp(a + bx_1)\}^4 \{1 + \exp(a + bx_2)\}^4 \{1 + \exp(a + bx_3)\}^2}$$

Hence the model is full rank only if $x_1 \neq x_2$

Example 2: Near-singular model

Consider the model with parameter set,

$$\theta = [\phi_{1,1}, \phi_{1,2}, \phi_{1,3}, \phi_a, \lambda_1, \lambda_a]$$

This model is full rank, but from the **PLUR** decomposition, we find

$$\text{Det(U)} = \frac{(\phi_{1,2} - \phi_{1,1})}{(1-\phi_{1,1})\lambda_1\phi_{1,1}\phi_a(1-\phi_{1,2})\lambda_a\phi_{1,2}(1-\phi_{1,3})}.$$

Example 3 – Tag Returns Fisheries Model Jiang et al (2007): Striped Bass, *Morone saxatilis*.

θ = [F, M_1, M_2, M_3, C_1, C_2, C_3, λ],

F – instantaneous fishing mortality rate

M_a – instantaneous natural mortality rate, at age a

C_a – selectivity coefficient for age a ($a > 3$ $C_a = 1$)

λ – reporting probability

P_{ijk} – probability fish tagged at age k, released year i harvested and returned year j

$$P_{ijk} = \left[\prod_{v=i}^{j-1} \exp\left\{-\left(F\,C_{k+v-i} + M_{k+v-i}\right)\right\}\right]\left[1 - \exp\left\{-\left(F\,C_{k+j-i} + M_{k+j-i}\right)\right\}\right]\frac{FC_{k+j-i}\lambda}{FC_{k+j-i} + M_{k+j-i}}$$

Exhaustive summaries

- In this example Maple lacks memory.
- Exhaustive summaries are particular reparameterisations.
- We seek exhaustive summaries that give cell probabilities that are structurally simpler.
- This can result in greater parameter redundancy.
- In this example we move from 16 to 24 parameters.
- We find a deficiency of 8 in the new parameter space.
- Note that Jiang et al., used numerical analysis and found a deficiency of 9.

ncse
National Centre for Statistical Ecology

General rules

- In some cases it is possible to establish general rules for models of particular stuctures.

- This avoids having to use Maple.

- A particular illustration of this occurs with age-dependent recovery models

Model notation for recovery models

Ring-recovery models are described as, for example:

C/A/C, T/A/C, T/A/T, C/C/T.

In this notation, each model is specified by 3 letters, which designate, in order,

1. The way we model first-year survival: C or T;
2. The way we model adult survival: C, A or T;
3. The way we model the recovery probability: C, A or T.

ncse National Centre for Statistical Ecology

Steps: age-dependence also in λ.

- Consider, for example, the model denoted by C/A(2,2,3)/A(2,1,1,4). What can we estimate here?

- Here we have the parameters:

$$\phi_1, \phi_2, \phi_2, \phi_3, \phi_3, \phi_4, \phi_4, \phi_4$$
$$\lambda_1, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_4, \lambda_4, \lambda_4$$

Steps: age-dependence also in λ.

- Consider, for example, the model denoted by C/A(2,2,3)/A(2,1,1,4). What can we estimate here?

- Here we have a single step, as shown:

$$\phi_1, \phi_2, \phi_2 \mid \phi_3, \phi_3, \phi_4, \phi_4, \phi_4$$
$$\lambda_1, \lambda_1, \lambda_2 \mid \lambda_3, \lambda_4, \lambda_4, \lambda_4, \lambda_4$$

Theorem 1

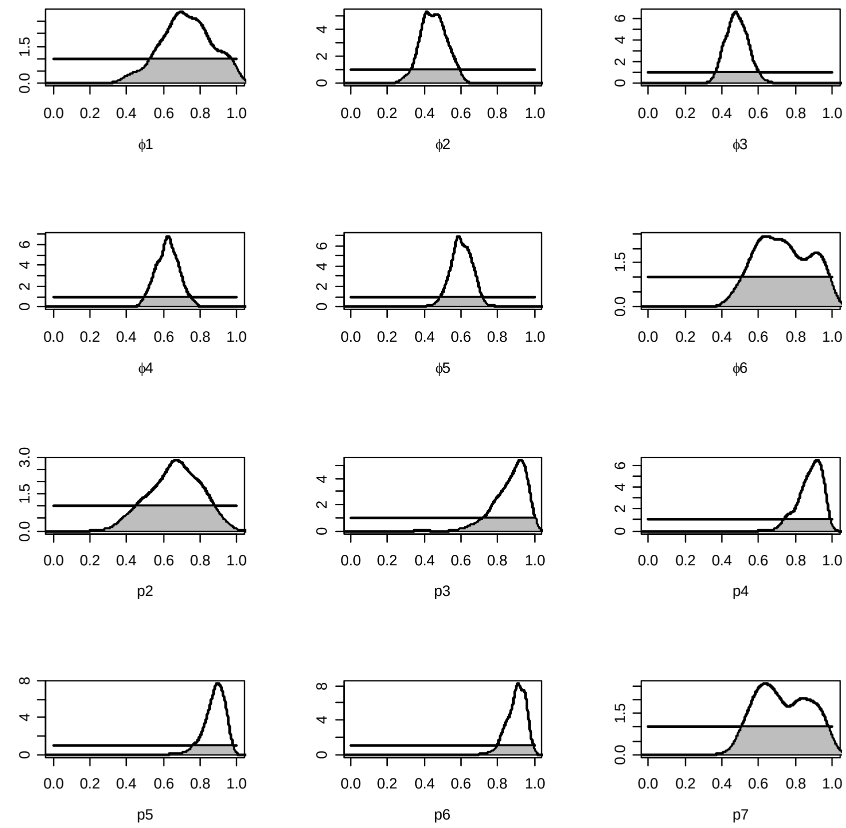
- Suppose the first step occurs at age n, and let m be the number of parameters used in the first n years.
- If $m = n+1$, the model is parameter redundant.
- If $1 < m < n+1$, then the step does not cause parameter redundancy. Furthermore, to test for parameter redundancy, the parameters occurring in the first n years can be discarded, and the count started anew in year $n+1$.

Theorem 2

- In the age-dependent model T/A/A
- The step at age 1 year does not cause parameter-redundancy
- To determine any possible redundancy caused by a subsequent step, the age and parameter counts begin again after age 1 year, as in Theorem 1.

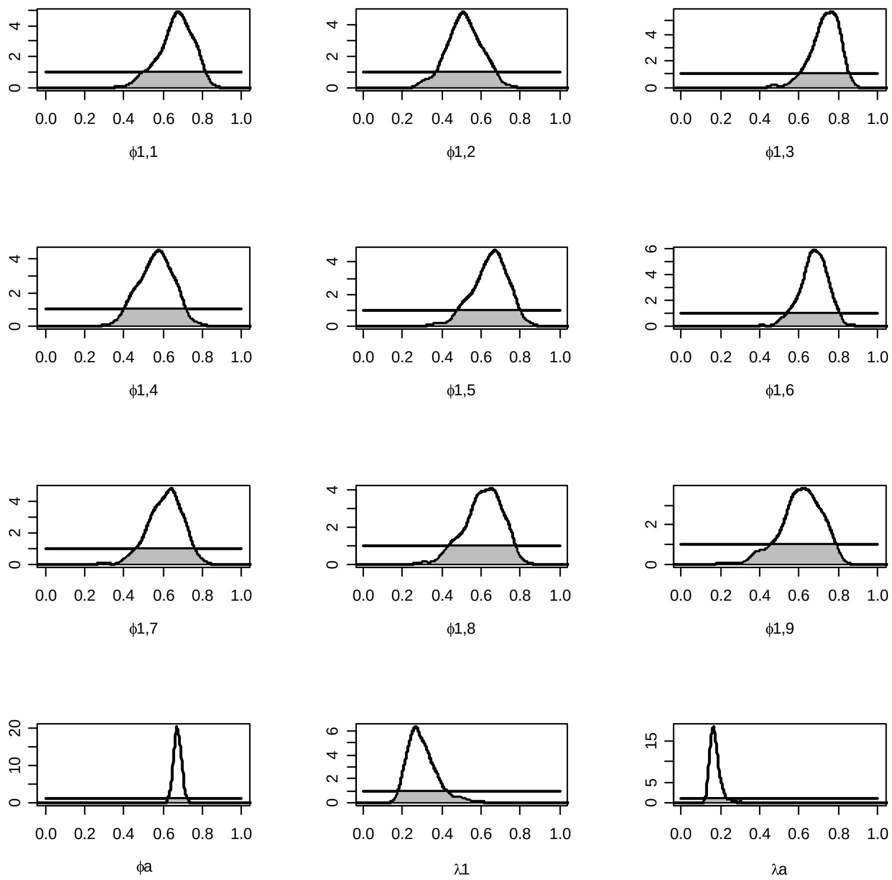
A Bayesian perspective: the CJS model

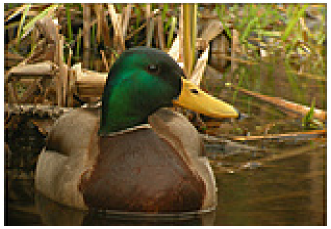
- In population ecology we may devise models with parameters that cannot be estimated from the data.
- Symbolic algebra can be used to examine whether a model is parameter-redundant.
- In a Bayesian context, it is interesting to consider the overlap between priors, $p(\theta)$ and posteriors $\pi(\theta|x)$.



ncse
National Centre for Statistical Ecology

Male mallard, *Anas platyrhyncos*





Model: $\phi_{1,i}$, ϕ_a, λ_1, λ_a here only two parameters, ϕ_a and λ_a are strongly identified.

National Centre for Statistical Ecology

New work

- Use of PLUR decomposition
- Use of covariates
- Use of exhaustive summaries
- Overlap of priors and posteriors

Other areas

- Econometrics (Rothenburg)
- Compartment modelling (Walter)
- Contingency tables (Goodman)

Acknowledgement

The work of Diana Cole is supported by the EPSRC grant for the NCSE.

ncse
National Centre for Statistical Ecology

References

- Bekker et al (1994) *Identification, equivalent models and computer algebra*.
- Bellman and Aström, 1970, *Mathematical Biosciences.*
- Catchpole, Freeman and Morgan, 1996, *JRSS B*.
- Catchpole and Morgan, 1997, *Biometrika*
- Catchpole, Morgan and Freeman, 1998, *Biometrika*.
- Catchpole and Morgan, 2001, *Biometrika*.
- Garrett and Zeger, 2000, *Biometrics*.
- Gimenez et al., 2003, *Biometrical J*.
- Gimenez, Morgan and Brooks, 2007. *J. Env. and Ecol. Stats*.
- Goodman, 1974, *Biometrika*.
- Jiang et al., 2007, *JABES*.
- Rothenberg, 1971, *Econometrica*.
- Walter, 1982, *Identifiability of state space models.*

ncse
National Centre for Statistical Ecology