ESTIMATING DENSITY DEPENDENCE, PROCESS NOISE, AND OBSERVATION ERROR

Coinvestigators:

José Ponciano, University of Idaho
Subhash Lele, University of Alberta
Mark Taper, Montana State University
David Staples, Montana State University
Model:

\[ N_t = g(N_{t-1}, \theta) \]

where \( N_t \) is population abundance at time \( t \)
\( \theta \) is a vector of parameters

Data:

American Redstart BBS # 02012 3336 08545
# 02014 3328 08636
American Redstart #02012 3336 08545

population abundance

time
1. OVERVIEW: CONNECTING ECOLOGICAL MODELS WITH TIME SERIES DATA

Approach: convert deterministic population models into stochastic time series models (explicitly model fluctuations)

A. Process noise

Types: demographic, environmental, catastrophic, etc.

ex. environmental

\[ N_t = g(N_{t-1}, \theta) \quad \text{(deterministic)} \]

\[ \ln N_t = \ln g(N_{t-1}, \theta) + E_t \quad \text{(stochastic)} \]

where \( E_t \) is a random noise process; typical assumption is that \( E_t \sim \text{normal}(0, \sigma^2) \) and \( E_1, E_2, E_3, \ldots \) are uncorrelated.

Notes:
- \( N_0, N_1, N_2, \ldots \) are dependent
- conditional distribution of \( \ln N_t \) given \( N_{t-1} = n_{t-1} \) is normal( \( \ln g(n_{t-1}, \theta), \sigma^2 \) )
Observations:  \( n_0, n_1, n_2, \ldots, n_q \)

Likelihood:

\[
L(\theta, \sigma^2) = f(n_1|n_0)f(n_2|n_1)\cdots f(n_q|n_{q-1})
\]

where

\[
f(n_t|n_{t-1}) = (\sigma^2 2\pi)^{-1/2} \exp \left[ - \frac{(\ln n_t - \ln g(n_{t-1}, \theta))^2}{2\sigma^2} \right]
\]

ML estimates of parameters in \( \theta \) minimize:

\[
\text{CSS} = \sum_{t=1}^q [\ln n_t - \ln g(n_{t-1}, \theta)]^2
\]

conditional sum of squares
B. Observation (sampling) error

\[ N_t = g(N_{t-1}, \theta) \]

\[ Y_t = N_t + F_t \]

where \( F_t \sim \text{normal}(0, \tau^2) \) and \( F_0, F_1, F_2, \ldots \) are uncorrelated.

Notes:

- \( N_t = h(t, \theta) \) is deterministic solution trajectory
- Initial pop. size \( N_0 = n_0 \) is an unknown model parameter
- \( Y_0, Y_1, \ldots, Y_q \) are independent
Observations: \( y_0, y_1, y_2, \ldots, y_q \)

Likelihood:
\[
L(\theta, \tau^2, n_0) = f(y_0) f(y_1) \cdots f(y_q)
\]
where
\[
f(y_t) = \left( \tau^2 2\pi \right)^{-1/2} \exp \left[ -\frac{(y_t - h(t, \theta))^2}{2\tau^2} \right]
\]

ML estimates of parameters in \( \theta \) minimize:
\[
TSS = \sum_{t=0}^{q} [y_t - h(t, \theta)]^2
\]
trajectory sum of squares
C. Combining process noise and observation error

\[ \ln N_t = \ln g(N_{t-1}, \theta) + E_t \]

\[ Y_t = N_t + F_t \]

“state space model”

Notes:

- Observations \( Y_0, Y_1, ..., Y_q \) are not just dependent, they are also not Markov.
- Usually, likelihood function (a repeated integral) cannot be written in simple form.
- Various approaches: likelihood via numerical simulation, Bayesian, etc (reviews by de Valpine 2002 *Bulletin of Marine Science*, and Clark and Bjørnstad 2004 *Ecology*).
- Parameter estimates (especially of \( \sigma^2 \) and \( \tau^2 \)) tend to be biased and confounded.
2. THE MODEL

A. The process model

Population abundance is assumed to change according to a discrete-time, stochastic Gompertz model. The Gompertz growth process takes the density dependence term to be proportional to \( \ln N_{t-1} \):

\[
N_t = N_{t-1} \exp(a + b \ln N_{t-1} + E_t).
\]

Here \( a \) and \( b \) are constants, and \( E_t \) has a normal distribution with mean 0 and variance \( \sigma^2 \), written \( E_t \sim \text{normal}(0, \sigma^2) \). Also, \( E_1, E_2, \ldots \) are assumed to be uncorrelated.

Let \( X_t = \ln N_t \). On the logarithmic scale,

\[
X_t = a + cX_{t-1} + E_t.
\]

Here \( c = (b + 1) \). Note: \( X_t \) is a first-order autoregressive process (AR(1) process).
**B. Properties of the process model**

- If $c < 1$, then the probability distribution for $X_t$ approaches a *stationary distribution* as $t$ becomes large:

\[
X_\infty \sim \text{normal} \left( \frac{a}{1 - c}, \frac{\sigma^2}{1 - c^2} \right).
\]

The stationary distribution for $N_\infty = \exp(X_\infty)$ is a *lognormal distribution*.

- If $c = 1$, then the model for $X_t$ is a discrete-time version of *Brownian motion with drift*. The corresponding model for $N_t$ is a discrete-time, stochastic model of exponential population growth (or decline). This is the density-independent population growth model studied by Dennis et al. (1991 *Ecological Monographs*).
C. Model with process noise and sampling error

Let $Y_t$ denote the estimated logarithmic population abundance (estimated value of $X_t$). The error is assumed to be normally distributed. Thus the full model is:

$$X_t = a + cX_{t-1} + E_t,$$

$$Y_t = X_t + F_t.$$  

Here $E_t \sim \text{normal}(0, \sigma^2)$, $F_t \sim \text{normal}(0, \tau^2)$, and the random errors/noises are assumed free of auto- or cross-correlations. The model implies that the sampling error inherent in estimating $N_t$ is lognormal. The model is a state-space model with an underlying, unobserved process $X_t$ and an observed process $Y_t$. The parameter $\tau^2$ is the variance of the log-scale estimation error.
D. Properties of the model with process noise plus sampling error

• If \( c < 1 \), then the probability distribution for \( Y_t \) approaches a stationary distribution as \( t \) becomes large:

\[
Y_{\infty} \sim \text{normal}\left(\frac{\alpha}{1-c}, \frac{\sigma^2}{1-c^2} + \tau^2\right).
\]

• If \( c = 1 \), then the model for \( Y_t \) is a discrete-time version of error-corrupted Brownian motion with drift. The model represents a discrete-time, stochastic model of exponential population growth (or decline) with lognormal sampling error. This is the model studied by Holmes (2001 *Proceedings of the National Academy of Sciences USA*) and Holmes and Fagan (2002 *Ecology*). They proposed a variance regression method for estimating parameters.
3. THE LIKELIHOOD FUNCTION: PROCESS NOISE PLUS SAMPLING ERROR

A. Multivariate normal likelihood

It can be shown that the observations $Y_0, Y_1, Y_2, ..., Y_q$ have a joint multivariate normal distribution, provided $Y_0$ arises from the stationary distribution, with:

$$E(Y_t) = \frac{a}{1 - c},$$

$$V(Y_t) = \frac{\sigma^2}{1 - c^2} + \tau^2,$$

$$\text{Cov}(Y_t, Y_{t+s}) = \frac{\sigma^2}{1 - c^2} c^{|s|}.$$  

The likelihood function is the multivariate normal pdf, evaluated at the data $y = [y_0, y_1, ..., y_q]'$:

$$L(a, c, \sigma^2, \tau^2) =$$

$$\left(\frac{1}{(2\pi)^{(q+1)/2} |V|^{1/2}}\right) \exp \left[-\frac{1}{2} (y - \mu)' V^{-1} (y - \mu) \right].$$
Interestingly, the likelihood function is identical to that of an AOV mixed effects model with repeated measures. SAS PROC MIXED can be “tricked” into calculating parameter estimates!

The AOV model: one subject (fixed intercept), with repeated measures on the subject (having AR(1) covariance structure), and random time effect. SAS program is appended with these notes.
B. The Kalman filter

Like the process-noise-only case, the likelihood $L(a, c, \sigma^2, \tau^2)$ for the model with process noise and sampling error can be decomposed into a product of univariate normal pdfs. However, the process $Y_t$ is \textit{not} a Markov process: that is, given $Y_{t-1} = y_{t-1}$, the distribution of $Y_t$ (or any future value of the process) \textit{does} depend on any and all values of the process prior to time $t - 1$. The pdf for $Y_t$, given $Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2}, \ldots, Y_0 = y_0$ is that of a normal distribution with mean $m_t$ and variance $v_t^2$ that are computed recursively using the history of the observations:

$$f(y_t \mid y_0, y_1, \ldots, y_{t-1})$$

$$= (v_t^2 2\pi)^{-1/2} \exp \left[ -\frac{(y_t - m_t)^2}{2v_t^2} \right].$$

The recursion relationships for $m_t$ and $v_t^2$ are

$$m_t = a + c \left[ m_{t-1} + \frac{v_{t-1}^2 - \tau^2}{v_{t-1}^2} (y_{t-1} - m_{t-1}) \right],$$

$$v_t^2 = c^2 \frac{v_{t-1}^2 - \tau^2}{v_{t-1}^2} \tau^2 + \sigma^2 + \tau^2.$$
If the initial population is assumed to arise from the stationary distribution, the recursions are initiated at the stationary mean and variance: $m_0 = a/(1 - c)$, $v_0^2 = [\sigma^2/(1 - c^2)] + \tau^2$. The pdf for $Y_0$ is that of the stationary normal distribution:

$$f(y_0) = (v_0^2 2\pi) \exp \left[ -\frac{(y_0 - m_0)^2}{2v_0^2} \right].$$

The recursion expressions for $m_t$ and $v_t^2$ are contained in a set of general equations known as the Kalman filter. Derivation of the expressions is straightforward; the derivation uses repeated applications of properties of the bivariate normal distribution.
With the conditional normal pdfs in hand, the likelihood function is thus

\[
L(a, c, \sigma^2, \tau^2) =
\]

\[
f(y_0) f(y_1 | y_0) f(y_2 | y_0, y_1) \cdots f(y_q | y_0, y_1, \ldots, y_{q-1})
\]

\[
= (2\pi)^{-(q+1)/2} \left( v_0^2 v_1^2 \cdots v_q^2 \right)^{-1/2} \times
\]

\[
\exp \left[ -\frac{1}{2} \sum_{t=0}^{q} \frac{(y_t - m_t)^2}{v_t^2} \right].
\]
4. REML ESTIMATION USING FIRST DIFFERENCES (*RESTRICTED MAXIMUM LIKELIHOOD*)

A. First differences

First differences are defined as:

\[ W_t = Y_t - Y_{t-1} \]

for \( t = 1, 2, ..., q \). Then \( W_1, W_2, ..., W_q \) have a joint multivariate normal distribution with

\[ \text{E}(W_t) = 0 \]

\[ \text{V}(W_t) = \frac{2\sigma^2}{1 - c^2} (1 - c) + 2\tau^2 \]

\[ \text{Cov}(W_t, W_{t+1}) = -\frac{\sigma^2}{1 - c^2} (1 - c)^2 - \tau^2 \]

\[ \text{Cov}(W_t, W_{t+s}) = -\frac{\sigma^2}{1 - c^2} (1 - c)^2 c^{|s|-1} \]
B. Likelihood function for REML

The data are \( w_1 = y_1 - y_0, \ w_2 = y_2 - y_1, \ldots, \ w_q = y_q - y_{q-1} \). The unknown parameters are \( c, \sigma^2, \tau^2 \) (\( a \) is eliminated in the distribution of the differences). The likelihood function is denoted \( L(c, \sigma^2, \tau^2) \):

\[
\mathbf{w} = [w_1, w_2, \ldots, w_q]'
\]

\[
L(c, \sigma^2, \tau^2) = \left( \frac{1}{(2\pi)^{q/2} |\Phi|^{1/2}} \right) \exp \left( -\frac{1}{2} \mathbf{w}' \Phi^{-1} \mathbf{w} \right)
\]
C. ML estimate of $a$, with the elements of $V$ known (i.e. fixed at REML values):

$$\hat{a} = (1 - c) \frac{j'V^{-1}y}{j'V^{-1}j}.$$
5. EXAMPLES

A. Data sets

   Breeding Bird Survey: American Redstart (2 locations)
   Simulation

B. Simulated properties of parameter estimates

   ML
   REML
estimated population abundances

population abundance

time

Estimated population abundances over time.
Figure 2.
Figure 3.
Figure 4.
Figure 5.
6. DISCUSSION POINTS

• There is information in population time-series data for jointly estimating density dependence, process noise, and observation error, and a variety of modeling approaches (of varying computational complexity). Estimation is tricky and needs hands-on attention. Linear Gaussian model can be adapted, via transforming to logarithmic scale, for more realistic ecological uses.

• ML estimation for the linear Gaussian model ( = Kalman filter) requires care but works reasonably well (simulations). Likelihood is routinely ridge-shaped & multimodal. The proper solution of the likelihood equation (giving statistically consistent estimates) frequently is not the global likelihood maximum. Published ML simulations for other models which did not accommodate multimodality are suspect.

• REML works reasonably well (preliminary simulations); seems to fix the some of the ML bias problems.
• Lognormal sampling model is a realistic model of ecological sampling under heterogeneous conditions: Poisson “mixture” models typically have constant coefficients of variation (as a function of size of population being sampled).

• Gompertz process model has held its own in comparative density dependence model-fitting studies (usually fits as well as logistic/Ricker/Bev-Holt).

• SAS program!
Supplement: SAS program for calculating parameter estimates for the Gompertz state space model.


 /*--------------------------------------------------------------------------*/
/* PARAMETER ESTIMATES FOR THE GOMPERTZ STATE SPACE MODEL */
/* SAS program to calculate parameter estimates for the Gompertz state- */
/* space model, using time series population abundance estimates. The */
/* GSS model is given by */
/* X(t) = a + cX(t-1) + E(t) */
/* Y(t) = X(t) + F(t) */
/* where X(t) is the natural logarithm of population abundance N(t) */
/* (assumed unknown), Y(t) is the observed value of X(t), E(t) has a */
/* normal distribution with mean 0 and variance sigmasquared, F(t) has */
/* a normal distribution with mean 0 and variance tausquared (with no */
/* auto- or cross-correlations in E(t) and F(t)), and t is time. Unknown */
/* model parameters are a, c, sigmasquared, tausquared. Data to be */
/* input into the program consist of observed or estimated population */
/* abundances O(0), O(1), O(2), ... O(q) (estimates of N(0), N(1), etc.), */
/* along with the values of t. The program currently does not accomodate */
/* missing observations. */
/* */
/* Program transforms data to logarithmic scale: Y(t) = ln[O(t)]. The */
/* program recasts the model as a linear mixed model with: (1) repeated */
/* measures on one subject having an AR(1) covariance structure, and (2) */
/* a random effect due to time (considered as a categorical variable). */
/* The random effect represents the extra variance component due to ob- */
/* servation error and produces a 'nugget' (augmented main diagonal) in */
/* the var-cov matrix for the observations. */
/* */
/* The example data are from the North American Breeding Bird Survey */
/* (record # 0214332008636, American Redstart), and correspond to Table 1 */
/* and Figure 1 of Dennis et al. (200X). */
 /*--------------------------------------------------------------------------*/

options nocenter;
data in;
input observed time;
y = log(observed);
cards;  
18 0  
10 1  
9 2  
14 3  
17 4  
14 5  
5 6  
10 7  
9 8  
5 9  
11 10 
11 11 
4 12 
5 13
```sas
proc mixed method=ml alpha=.05 noitprint noinfo data = in;
  class time;
  model y= ;
  random time;
  repeated / type=ar(1) subject=intercept;
  estimate 'intercept' intercept 1;
run;
quit;
```

The SAS System

The Mixed Procedure

Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>30</td>
<td>0 1 2 3 4 5 6 7 8 9 10 11 12</td>
</tr>
</tbody>
</table>

/* Restricted maximum likelihood (REML) is the default estimation method */
/* in PROC MIXED (SAS System for Windows Version 9.1). Delete "method= /*
/* ml" (or substitute "method=reml") in list of options in the above */
/* 'proc mixed' statement for REML estimation if desired. Also, the */
/* value of alpha, for asymptotic 100(1-alpha)% confidence intervals for */
/* parameters, can be changed in the option list. */

/* -----------------------------------------------*/
/* ANNOTATED OUTPUT OF THE GSS ESTIMATION PROGRAM */
/* */
/* The following output was generated using SAS/STAT software, Version */
/* 9.1 of the SAS System for Windows. Copyright (c) 2002-2003 SAS */
/* Institute Inc. SAS and all other SAS Institute Inc. product or */
/* service names are registered trademarks or trademarks of SAS Institute */
/* Inc., Cary, NC, USA. */
### Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td></td>
<td>0.2315</td>
<td>0.05</td>
<td>0.08439</td>
<td>1.7944</td>
</tr>
<tr>
<td>AR(1)</td>
<td>Intercept</td>
<td>0.7934</td>
<td>0.05</td>
<td>0.1859</td>
<td>1.4010</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>0.2625</td>
<td>0.05</td>
<td>0.08314</td>
<td>3.9119</td>
</tr>
</tbody>
</table>

/* In the 'Estimate' column, the value listed for 'time' is the estimate of $\tau^2$, for 'AR(1)' is $c$, and for 'residual' is $\sigma^2/(1 - c^2)$ (the stationary variance of $X(t)$). 'Lower' and 'Upper' columns give boundaries of asymptotic 95% confidence intervals for the parameters, based on inversion of the Hessian of the log-likelihood. The CIs have unknown coverage properties for small- and moderate-lengthed time series. The CI for $c$, along with the large value of the stationary variance upper bound, might suggest that the density independent model ($c=1$) is a viable model for the data. A SAS program to fit the density independent state space model was provided as a supplement to Staples et al. (2004). */

### Fit Statistics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 Log Likelihood</td>
<td>57.0</td>
</tr>
<tr>
<td>AIC (smaller is better)</td>
<td>65.0</td>
</tr>
<tr>
<td>AICC (smaller is better)</td>
<td>66.6</td>
</tr>
<tr>
<td>BIC (smaller is better)</td>
<td>70.6</td>
</tr>
</tbody>
</table>

/* These "fit statistics" can be used for model selection, in comparison to other models fitted to the data. */

### Estimates

| Label  | Estimate | Error | DF | t Value | Pr > |t| |
|--------|----------|-------|----|---------|------|--|
| intercept | 1.9021   | 0.2645 | 29 | 7.19    | <.0001 |

/* The estimate listed for "intercept" is the estimate of $a/(1-c)$, the stationary mean of $X(t)$. The "Standard Error" is an asymptotic estimate based on the information matrix. The t-test for the null hypothesis that $a/(1-c)=0$ is nonsensical in the context of the model. */

/* Thus for the example BBS data, the ML parameter estimates are: */

/* $\tau^2 = 0.2315$, */

/* $c = 0.7934$, */

/* $\sigma^2/(1 - c^2) = 0.09726$, */

/* $a = 1.9021*(1-c) = 0.3930$, */

/* Compare with ML estimates, Table 1, Dennis et al. (200X). Small numerical differences are due to roundoff error in SAS. */
REFERENCES


Brian Dennis
brian@uidaho.edu
http://www.cnrhome.uidaho.edu/fishwild/dennis

This presentation:
http://www.webpages.uidaho.edu/~brian/reprints/
kalmanESA2006.pdf

The paper: