Nested optimisation; Application to estimation of variation in annual mortality in fish populations*

Lennart Frimannslund† Hans Julius Skaug‡

March 31, 2006

Abstract

We study a population dynamics model incorporating natural mortality as well mortality due to human exploitation. The model is applicable to Norwegian spring spawning herring. Our goal is to make inference about annual variation in natural mortality. Using the Laplace approximation of the marginal likelihood, we derive a likelihood function for the unknown parameters in the model. The statistical properties of the estimators are investigated using simulated data sets. We do not find evidence for annual variation in mortality for Norwegian spring spawning herring, but our simulation experiments indicate that one would need much more data than currently available to be able to detect such an effect.

Keywords: Laplace Approximation, Monte Carlo-Simulation, Automatic Differentiation, Pattern Search.

1 Introduction

The annual mortality rate M is an important demographic parameter in wildlife populations. The probability that an individual survives from one year to the next is e^{-M}, and this serves to define M. In a fish population there can be large annual variations in M, due to

*This work was supported by the Norwegian research council (NFR).

†Department of Informatics, University of Bergen, Pb. 7800, 5020 Bergen, Norway.

‡Department of Mathematics, University of Bergen, Johannes Brunsgate 12, 5008 Bergen, Norway.

changes in environmental conditions and variation in predation pressure. Denote by $M + \epsilon_t$ the mortality rate in year t, where ϵ_t is a perturbation around the average mortality rate M. We assume that there is no direct measurement of ϵ_t available, and hence we shall view ϵ_t as a stochastic variable with expectation 0 and unknown variance τ^2. We formulate a stochastic population dynamic model, and derive an objective function that allows M and τ to be estimated from catch data and data from scientific surveys. For most marine fish populations such data are scarce, and estimation of M is a difficult problem. Needless to say, estimation of the level of annual variation (τ) is an even harder problem. Hence, with the current level of information we cannot hope to get reliable estimates of τ. In the present paper we use simulations to investigate how much data would need to be available in order for τ to be identifiable with a reasonable degree of precision.

The stochastic population dynamic model we use is an instance of a state-space model. There are two types of unknown quantities in such models: 1) the state variables, which here are the number of individuals being alive each year, and 2) the (structural) parameters: M, τ, along with some other parameters to be defined later. State-space models are often fit to data using Kalman-filter techniques [10]. When the model is non-linear, the equations must be linearised before one can apply the standard Kalman machinery. We use the Laplace approximation [7] to integrate out the state variables from the likelihood function. This leaves us with the marginal likelihood, which becomes our objective function for estimation of M and τ. The Laplace approximation is itself phrased as an optimisation problem, so our approach involves nested optimisation. The inner optimisation problem is solved using a quasi-Newton algorithm. We solve the outer problem using two approaches, quasi-Newton and a pattern search algorithm, the latter of which which allows for the inner problem to be solved inexactly, hence more cheaply.

The rest of the paper is organised as follows. In section 2 we outline the stochastic population dynamic model. In section 3 we outline the computational methods, and in section 4 we apply the method to data for Norwegian Spring Spawning Herring along with simulated datasets, which we discuss in section 5.

2 Population Dynamics of Exploited Fish Stocks

Most of our large fish populations are subject to human exploitation. We assume that the number of individuals C removed from the popu-

lation each year by fisheries is known. The mortality rate M referred to above is the "natural" mortality, and does not include the mortality caused by the fisheries.

We consider a period of n years, labeled $t = 1, \ldots, n$ for simplicity. Our population consists of A independent cohorts. In real life a cohort consists of all fish born in a particular year, but for simplicity we shall here treat the "cohorts" as being coexisting, but otherwise unrelated, developing populations. The basic equation governing the population dynamics of the jth, $j = 1, \ldots, A$ cohort is

$$N_{j,t} = (N_{j,t-1} - C_{j,t-1})\, e^{-(M+\epsilon_{t-1})}, \quad t = 1, \ldots, n, \tag{1}$$

where the quantities are:

$N_{j,t}$	Number of individuals in cohort j in year t,
$M + \epsilon_t$	Mortality in year t (applies to all cohorts),
$C_{j,t}$	Catches in numbers of individuals in cohort j in year t.

The model specification is completed with the requirement that ϵ_t has a Gaussian distribution with mean 0 and variance τ^2. Note that this assumption allows for $e^{-(M+\epsilon_t)} > 1$, which does not have an interpretation in terms of survival.

2.1 Available Data

In addition to the catch numbers C, data from acoustic scientific surveys are available. These surveys provide relative indices I of population size, in the sense that I is an estimate of $q \cdot N$, where q is a number satisfying $0 < q < 1$. We refer to q as the "catchability" parameter, and it may be given the interpretation that the survey covers only a proportion of the total population. By reading the age of individual fish in a random sample it is possible to calculate a survey index for each cohort. In the Norwegian Spring Spawning Herring data, and in our simulated datasets, there are four surveys each year, each with their own catchability parameter. The key quantities involved are:

$I_{j,s,t}$	Survey index for cohort j in survey s in year t,
q_s	"Catchability" in survey s.

The statistical assumption we make is that $\log(I_{j,s,t})$ has a normal distribution with expectation $\log(q_s \cdot N_{j,t})$ and variance σ^2.

2.2 Likelihood Function

In order to initialise the system (1) we need values for $(N_{1,0}, \ldots, N_{A,0})$, i.e. the state vector at time zero. These values will be estimated

along with the other parameters of the model. Hence, the parameters are: $\theta = (M, \tau, \sigma, q_1, \ldots, q_S, N_{1,0}, \ldots, N_{A,0})$. The other independent variables in the model are the parameters dealing with variation in mortality, $(\epsilon_1, \ldots, \epsilon_n)$. Let ϵ (without subscript) denote the vector $(\epsilon_1, \ldots, \epsilon_n)$, and similarly for the other variables. The log-likelihood function, from which we shall construct our objective function, has two parts:

$$\begin{aligned} l(\theta, \epsilon) &= \sum_{t=1}^{n} \sum_{s=1}^{S} \sum_{j=1}^{A} \left[-\log(\sigma) - \frac{(\log(I_{j,s,t}) - \log(q_s N_{j,t}))^2}{2\sigma^2} \right] \\ &\quad + \sum_{t=0}^{n-1} \left[-\log(\tau) - \frac{\epsilon_t^2}{2\tau^2} \right]. \end{aligned} \tag{2}$$

The first part arises from the distributional assumptions made about $I_{j,s,t}$, while the second part comes from the distributional assumptions made about the ϵ_t. Note that the two parts are coupled through (1), where $M + \epsilon_t$ occurs.

2.3 Laplace Approximation

Denote the function (2) by $l(\theta, \epsilon)$ where θ denotes all other independent variables than ϵ. It is well established in the statistical literature (e.g. [6], p. 466) that joint maximisation of l with respect to θ and ϵ does not give a good estimate of θ, and hence not of τ which is the parameter of primary interest to us. Instead, one can use the Laplace approximation [8]

$$l^*(\theta) = -\frac{1}{2} \log \det(-H(\theta)) + l(\theta, \bar{\epsilon}(\theta)), \tag{3}$$

of the marginal log-likelihood

$$l(\theta) = \log \left[\int \exp\{l(\theta, \epsilon)\} \, d\epsilon \right]. \tag{4}$$

In (3), $\bar{\epsilon}(\theta)$ is the maximiser of $l(\theta, \epsilon)$ with respect to ϵ for a fixed value of θ, and the symmetric matrix function H is defined as

$$H(\theta) = \frac{\partial^2}{\partial \epsilon^2} l(\theta, \epsilon)|_{\epsilon = \bar{\epsilon}(\theta)}. \tag{5}$$

Numerical evaluation of $l^*(\theta)$ may be done as follows:

- Maximise $l(\theta, \epsilon)$ with respect to ϵ to obtain $\bar{\epsilon}(\theta)$. This optimisation step is referred to as the "inner optimisation".

- Evaluate H at $\bar{\epsilon}$.
- Compute the determinant of H by means of a Cholesky factorisation and compute the expression (3).

The inner optimisation can be performed efficiently with a quasi-Newton method, with the gradient computed by Automatic Differentiation (AD), and H can also be computed by AD. AD (see, e.g. [3]) is a collection of techniques which can compute derivatives of a function defined through computer code, to machine precision. These techniques are attractive since they are usually transparent to the user, and can compute the gradient of a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ at between four and five times the cost of evaluating the function itself. AD may however require a large amount of storage space.

When $\bar{\epsilon}(\theta)$ does not maximise $l(\theta, \epsilon)$ exactly, in which case we write $\tilde{\epsilon}$, we must include a correction term in the Laplace approximation,

$$l^*(\theta) = -\frac{1}{2}\log\det(-H(\theta)) + l(\theta, \tilde{\epsilon}(\theta)) - \frac{1}{2}\nabla l^T H^{-1}\nabla l, \qquad (6)$$

where ∇l is the gradient of $l(\theta, \epsilon)$ with respect to ϵ evaluated at $\tilde{\epsilon}$, and H given by (5) now is evaluated at $\tilde{\epsilon}$. A proof of the result goes as follows. By a second order Taylor expansion, and skipping the argument θ from our notation, we get

$$l(\epsilon) \approx l(\tilde{\epsilon}) + \nabla l^T(\epsilon - \tilde{\epsilon}) + \frac{1}{2}(\epsilon - \tilde{\epsilon})^T H(\epsilon - \tilde{\epsilon}). \qquad (7)$$

By algebraic manipulation we find that

$$\begin{aligned}(\epsilon - \tilde{\epsilon} + H^{-1}\nabla l)^T H(\epsilon - \tilde{\epsilon} + H^{-1}\nabla l) \\ = (\epsilon - \tilde{\epsilon})^T H(\epsilon - \tilde{\epsilon}) + 2\nabla l^T(\epsilon - \tilde{\epsilon}) + \nabla l^T H^{-1}\nabla l,\end{aligned}$$

which can be used to rewrite the Taylor expansion (7) as

$$l(\epsilon) = l(\tilde{\epsilon}) - \frac{1}{2}\nabla l^T H^{-1}\nabla l + \frac{1}{2}(\epsilon - \tilde{\epsilon} + H^{-1}\nabla l)^T H(\epsilon - \tilde{\epsilon} + H^{-1}\nabla l).$$

Further, we have the multivariate normal integral

$$\int \exp\left[\frac{1}{2}(\epsilon - v)^T H(\epsilon - v)\right] d\epsilon = c \cdot \det(-H)^{-1/2}, \qquad (8)$$

where $c = (2\pi)^{n/2}$ is a constant that can be ignored in the present context. Since (8) holds for all values of v, and in particular

$$v = \tilde{\epsilon} - H^{-1}\nabla l,$$

the approximation (6) of the marginal likelihood (4) follows (after a few lines of thought).

3 Optimising the Likelihood Function

We consider two different methods for optimising the objective function (3). The first approach is to solve the problem by using the Quasi-Newton solver that is built into AD Model builder [1] (ADMB), a commercially available package for nonlinear statistical models. The second is to use a variant of the pattern search method of [2] for the outer problem, and the BFGS method to solve the inner problem, with gradients and H computed by the ADOL-C package [4].

A difference between the two approaches from a theoretical point of view is that the former requires the gradient of $l^*(\theta)$, which involves third order mixed derivatives of $l(\theta, \epsilon)$ [8], whereas the latter only requires the second derivative $H(\theta)$. When using AD, the computation of the gradient of a functional ($l^*(\theta)$ in our case) can be done with less than or equal to five times the amount of work required to compute the function value itself, and thus our problem appears well suited for a gradient-based method. The price to pay for the "cheap" gradient is that one has to store a computational graph (or an execution trace, sometimes called a *tape*) the size of which can be substantial in the sense that it can be larger than the available disk space. The size of the tape depends on the number and nature of the operations required to compute the function value.

In our case, the length of the vector ϵ, n, is important. In order to obtain the gradient of (3) we need to differentiate the Cholesky factorisation of the Hessian H, whose dimension is $n \times n$. The computational graph, and corresponding overhead, used in computing ∇l^* will therefore grow as n grows, and make the computation of ∇l^* more cumbersome. The principle that a gradient can be obtained at five times the cost in operation count of the function still applies, but the storage requirements may be substantial for large n, a problem which is not encountered on the same scale when calculating function values only.

In our context, however, since we only have one time step per year (that is, the formula (1) is only applied one time for each year considered) and fish have a limited life span, we do not expect ϵ to have significantly more elements than 20. Consequently, the computational graphs involved are of acceptable size on a personal computer, and the gradient-based method performs well. In addition, the ADMB package contains a differentiated version of the Cholesky algorithm (see e.g. [9]), which reduces storage requirements.

As for pattern search, its ability to cope with non-smoothness means that one may solve the inner optimisation problem inexactly initially and solve it more and more accurately as one approaches the

solution of the outer problem (6). This reduces the total time of the optimisation, since inexact function evaluations can be performed at a relatively low cost. The suggested pattern search method must be either be modified slightly to handle constraints, or one can handle the constraints by using Lagrange multipliers and returning infinity (or negative infinity) for points outside the domain of the function.

Both methods seem to benefit from the two-phase strategy outlined in [7]. The two-phase strategy should not be confused with the nested (inner-outer) optimisation scheme that is common to both the methods we discuss. In phase I the objective function is taken to be (2), but with $\epsilon = 0$ and τ fixed at some initial value. In particular, there is no Laplace approximation involved in the first phase. Note that the second term in (2) now can be ignored. Phase I hence provides estimates for all components of θ except τ. These estimates are used as initial values for phase II in which the objective function is taken to be the Laplace approximation (3).

Summing up, both methods are applicable to the problem; in our context the ADMB package is faster than our experimental codes, so we use the former in the next section.

4 Simulation Experiments

The question we ask in this section is: what type of, and how much, data do we need to be able to estimate τ? For this purpose we generate artificial data from the model (1) via Monte Carlo simulation. Hence we know what the true parameter values θ are. Then, we fit the model to the simulated data, as explained above and obtain an estimate $\widehat{\theta}$. This procedure is repeated many times, and we can measure the statistical properties (mean and standard deviation) of the estimator $\widehat{\theta}$. The variable of main interest to us is τ, so we created 1,000 data sets for each of the values

$$\tau_{\text{real}} = \{0.05,\ 0.1,\ 0.2\}\,,$$

where in addition

$$N_0 = \left[\ 200 \quad 200 \quad 200 \quad 200\ \right]^T,$$

(implying four cohorts)

$$q = \left[\ 0.5 \quad 0.5 \quad 0.5 \quad 0.5\ \right]^T,$$

implying four surveys, and finally

$$\sigma = 0.2,\ M = 0.15,\ n = 20.$$

a: $\tau_{\text{true}} = 0.05$

Data available	Mean($\widehat{\tau}$)	Std($\widehat{\tau}$)	# $\tau_{\min}$
50%	0.0181 (0.0559)	0.0274 (0.0183)	689
75%	0.0246 (0.0528)	0.0281 (0.0166)	544
100%	0.0297 (0.0488)	0.0265 (0.0160)	400

b: $\tau_{\text{true}} = 0.1$

Data available	Mean($\widehat{\tau}$)	Std($\widehat{\tau}$)	# $\tau_{\min}$
50%	0.0782 (0.0899)	0.0398 (0.0280)	132
75%	0.0881 (0.0927)	0.0319 (0.0254)	50
100%	0.0890 (0.0912)	0.0280 (0.0245)	25

c: $\tau_{\text{true}} = 0.2$

Data available	Mean($\widehat{\tau}$)	Std($\widehat{\tau}$)	# $\tau_{\min}$
50%	0.1870 (0.1908)	0.0424 (0.0373)	3
75%	0.1869	0.0386	0
100%	0.1905	0.0377	0

Table 1: Numerical results (1000 Monte Carlo replica) for different values of τ_{true} The numbers in parentheses show the results when only the instances where $\hat{\tau} \neq \tau_{\min}$ are included.

4.1 Results

The results are given in Table 1 a)–c). The columns of the tables signify, from left to right, the amount of survey data available, mean and standard deviation of $\widehat{\tau}$ and the number of times where $\widehat{\tau}$ was equal to the lower bound ($\tau_{\min} = 10^{-3}$) set by the optimisation algorithm. The table also shows results when the cases where $\hat{\tau}$ is equal to the lower bound are excluded. By available survey data, we mean the percentage of the $4 \cdot 4 \cdot 20 = 320$ acoustical observations available, where which observations are available is randomly selected.

5 Discussion

From Table 1 we draw the following conclusions:

- The more available data, the closer the mean of $\hat{\tau}$ is to τ_{true}.

- The more available data, the smaller the standard deviation of $\hat{\tau}$.
- Using only the cases where $\hat{\tau} \neq \tau_{\min}$ reduces the bias in the estimator.
- The larger the value of τ_{true}, the fewer cases of $\hat{\tau} = \tau_{\min}$.
- The larger the value of τ_{true}, the larger the standard deviation of $\hat{\tau}$.

We also applied the method to time series data from Norwegian spring spawning herring [5], which resulted in an estimate of $\tau = 0$. Apparently, there is no year-to-year variation in mortality, but an important point is that the uncertainty associated with the estimate is large. The objective function (3) is flat near its optimum, which is actually located at the boundary of the parameter space (τ must be non-negative). Hence, data provided little information about the true value of τ, which is what we expected. The simulation results presented in Table 1 support this conclusion, in that the probability that $\hat{\tau}$ ends up at zero is high, particularly when τ is small, like $\tau = 0.05$.

References

[1] D. Fournier. An introduction to AD MODEL BUILDER Version 6.0.2 for use in nonlinear modeling and statistics. Available from http://otter-rsch.com/admodel.htm, 2001.

[2] L. Frimannslund and T. Steihaug. A generating set search method using curvature information. To appear in Computational Optimization and Applications, 2006.

[3] A. Griewank. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Number 19 in Frontiers in Appl. Math. SIAM, Philadelphia, PA, 2000. ISBN 0–89871–451–6.

[4] A. Griewank, D. Juedes, and J. Utke. Algorithm 755: ADOL-C: a package for the automatic differentiation of algorithms written in C/C++. *ACM Transactions on Mathematical Software*, 22(2):131–167, June 1996.

[5] ICES. Report of the northern pelagic and blue whiting fisheries working group. ICES CM 2002/ACFM19, 2002.

[6] Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, 2001.

[7] H. Skaug and D. Fournier. Evaluating the Laplace approximation by automatic differentiation in nonlinear hierarchical models.

Technical report, Inst. of Marine Research, Box 1870 Nordnes, 5817 Bergen, Norway, 2005.

[8] H. Skaug and D. Fournier. Automatic approximation of the marginal likelihood in non-gaussian hierarchical models. Technical report, 2006. To appear in Computational Statistics and Data Analysis.

[9] S. P. Smith. Differentiation of the Cholesky algorithm. *Journal of Computational and Graphical Statistics*, 4(2):134–147, 1995.

[10] M. West and P. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York, 1997.