Problematic Likelihood Functions from Sensible Population Dynamics Models: A Case Study

Geof H. Givens\textsuperscript{a,*} and David Poole\textsuperscript{b}

\textsuperscript{a}Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA
\textsuperscript{b}AT&T Labs-Research, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932, USA

September 1, 2000

Abstract

All International Whaling Commission assessments of the Bering-Chukchi-Beaufort Seas stock of bowhead whales (\textit{Balaena mysticetus}) rely on likelihood functions derived from nearly the same data and a particular family of population dynamics models. Eleven such past bowhead assessments are compared. We note that the type of dynamics model used in all of these assessments has a strong influence on the features of the likelihood surface. We examine how the likelihood surface created by such models exhibits a narrow, cusp-shaped, flat-topped, steep-edged ridge of strong non-linear dependency between key model parameters. We discuss how such features are generally troublesome for statistical inference and interpretation. Through examples we examine some of the implications for maximum likelihood estimation, Bayesian estimation, and bootstrapping. Although such a dynamics model is very useful for producing realistic population trajectories, it is a poor model from which to generate likelihood functions in this case because it and the data together establish a nearly chaotic dynamical nonlinear system.

\textit{Keywords:} Statistics; Simulation Model; Density Dependence; Likelihood Surface; Stock Assessment; Bowhead Whale; International Whaling Commission

1 Introduction

Recent International Whaling Commission (IWC) assessments of the Bering-Chukchi-Beaufort Seas stock of bowhead whales (\textit{Balaena mysticetus}) have generated substantial debate about population modelling and statistical estimation problems of interest to a wide variety of ecological modelers. The IWC effort provides a good example with which to examine the difficulties that can arise while modelling with diverse data in hand.

A broad summary of past IWC considerations of bowhead modelling and assessment might be categorized as follows:

1. Data Collection and Summary. Ice-based visual and acoustic survey abundance data and estimates (Raftery and Zeh, 1998); photogrammetric data on stock age structure (Angliss et al.,

\textsuperscript{*}Corresponding author. Tel.: +1-970-4916402; fax: +1-970-4917835. \textit{Email address:} geof@lamar.colostate.edu
1995); data on biological parameters, including whale aging via aspartic acid racemization (George et al., 1999) and survival rate analysis via photo-identification (Zeh et al., 2000); historical catch data (Bockstoce and Botkin, 1983) and estimation of uncertainty therein (Givens and Thompson, 1996).

2. **Population Model Structure and Selection.** Choice of density-dependent population dynamics model (de la Mare, 1989; de la Mare and Cooke, 1993; Punt, 1998); variations on handling density dependence (Punt, 1996); stochastic versions of the model (Givens, 1999; Punt, 1999).

3. **Estimation or Model Fitting Method.** Bayesian, frequentist, or other paradigms; choice of Bayesian priors to include and their form; choice of constraints among parameters. See IWC (1999a) and references therein for background material in this area.

The IWC has rightly focussed a great deal of attention on (1). New data and improved interpretations of old data are essential to advance any applied modelling effort.

The intense IWC scrutiny accorded to (3) is ironic considering that a data analysis problem is often considered to have reached maturity when a wide variety of estimation approaches provide the same fundamental scientific conclusions. At that point, scientific progress is replaced with debate among statisticians about the finer points of estimation: choosing a bias-variance tradeoff, reducing higher-order approximation error, philosophy.

It is remarkable that the IWC has focussed relatively less attention on (2). The choice of a suitable model can have much greater impact on estimates than the particular estimation approach employed. For example, consider estimating $\theta$ from data $X_1, \ldots, X_n \sim N(\theta, 1)$ using model $X_1, \ldots, X_n \sim N(g(\theta), 1)$. Using the correct model ($g(\theta) = \theta$) one can obtain an adequate estimate using maximum likelihood estimation (MLE), Bayesian estimation, method of moments, minimax theory, or a myriad of other estimation choices. However, with a poor model, one is in trouble regardless of estimation approach.

Virtually all recent IWC baleen whale assessment has relied on variants of one type of population dynamics model. In this paper, we investigate some features of this model type by examining the likelihood surface it generates for bowheads and the implications of that type of surface. Our exclusive focus on the likelihood surface is intended to address topic (2) above and to sidestep any arguments about estimation method.

1.1 **A Review of Recent Assessments**

Before abandoning topic (3), however, we provide a brief overview of past assessment results as further motivation for model scrutiny. Table 1 lists 11 recent bowhead assessments that used different methods. Each analysis is the most recent of its type. Assessments 'C' and 'F' are the ones upon which the IWC has most recently based its management advice (IWC, 1999a).

Of course, these 11 analyses are not strictly comparable for many reasons.\(^1\) Furthermore, the 11

---

\(^1\) Using the jargon developed in these assessments, some reasons are as follows. $Q_0$ is replaced by RY for 'H'-K'. The median replaces the mean for 'D'-F'. Estimates of depletion and $Q_0$ (or RY) pertain to various years between 1993 and 1998. Estimates of depletion for 'D'-F' pertain to the total population rather than the 1+ component. Estimates of MSYR\(_{1+}\) were translated from MSYR\(_{\text{mat}}\) for 'H'-K' using a factor of 2 (i.e. MSYR\(_{\text{mat}}\) = 2MSYR\(_{1+}\)). The choice of 2 was motivated by Punt and Butterworth (1996) who showed a variety of bowhead assessments for which the conversion factor ranged from 1.93 to 2.04.

2
analysis methods are not equally defensible. As a natural part of the scientific peer review process, most of them have received significant criticism since their introduction.

Nevertheless, it is interesting to examine results from the 11 assessments. These results are summarized in Figures 1 and 2. Figure 1 shows 90% confidence intervals (or posterior probability intervals in the Bayesian analyses) for six quantities: equilibrium carrying capacity, permissible annual catch in the sense of Givens et al. (1996) and Wade and Givens (1997), current depletion expressed as the ratio of stock size to carrying capacity in terms of both the whales aged 1+ and the mature females, current population size as a proportion of maximum sustainable yield level (MSYL), and net productivity rate at the MSYL (maximum sustainable yield rate, or MSYR). The point estimate or mean is shown in the interior of each interval with a symbol corresponding to the Table 1 labels. A few assessment results are not recorded in the available literature.

Figure 2 shows scatterplots of (a) the 5th percentile of permissible catch versus the point estimate of depletion, and (b) point estimates of MSYR versus population size as a proportion of MSYL. In panel (a), management implications for ‘G’ are very different than for all other assessments shown. ‘G’ favors a large catch on a stock that is well above MSYL, whereas the other assessments favor a smaller catch on a stock that is below MSYL. Similar results are shown in panel (b): ‘G’ estimates a high MSYR and a stock that is nearly at carrying capacity, whereas the other assessments estimate lower MSYR and notable stock depletion. Figure 1 also shows that ‘G’ is clearly atypical with respect to confidence interval widths and endpoints, and especially with respect to central estimates.

This comparison raises the question of why ‘G’ produced startlingly different results. The correctness of an assessment is not determined by how well it matches past assessments. However, until the reasons for its atypicality are well-understood, it is reasonable to maintain skepticism about assessment ‘G’. Some differences in results among these methods, and some criticisms of the methods themselves, are driven by how the methods react to features of the likelihood surface. Thus, we return now to a general scrutiny of the likelihood surface generated by typical bowhead assessment models.

2 Methods

2.1 The Full Model

Space permits only the briefest summary of some main features of the model used for IWC bowhead assessments, BALEEN II (de la Mare, 1989; de la Mare and Cooke, 1993; Punt, 1998). This discrete-time deterministic population dynamics model is age- and sex-stratified, and accounts for (i) density-dependent fecundity and (ii) natural and hunting mortality. Numbers of newborn calves, and whales in each age/sex stratum are updated annually to reflect these processes. The stock is initially assumed to be at carrying capacity before exploitation, with its corresponding stable age distribution.

Transition to sexual maturity occurs instantaneously when a female reaches age m. In year t, the sexually mature females have combined pregnancy-and-calf-survival rate

\[ f_t = \max\left(f_{aq}(1 + A(1 - (P^D_t/K^D_t)^2)), 0\right) \]

where \( P^D_t \) is the size in year t of the component of the population to which density dependence is related and \( K^D_t \) is the equilibrium carrying capacity of the same component. Newborns are split
equally into males and females. The parameters $f_{eq}$, $A$, and $z$, are related to the familiar concepts of MSYL and MSYR via complex nonlinear equations. MSYL is the value of $P_t^D/K_t^D$ that yields maximum net productivity, and MSYR is the net productivity rate at MSYL.

An age-specific annual natural mortality rate is applied to each age/sex class, including calves. Very old whales of each sex are pooled into two maximum-age classes, where they linger until death. Usually, modelers impose no maximum lifespan. Hunting mortality is distributed to all whales aged 1 or older, in proportion to the relative abundances in each age/sex class.

The likelihood for this model, and the data from which it is derived, are identical to the description given below in Section 2.3.

2.2 The Simple Model

For bowhead assessment, a simple population dynamics model captures most of the dynamics from far more complex models like BALEEN II. Denote by $\eta_t$ the modeled total whale abundance in year $t$. Simulation occurs from 1848 when the stock is assumed to be at equilibrium carrying capacity $K$ and commercial harvest began, through 1993 when the last available abundance survey was conducted. Figure 3 shows the historical catch record during this period (IWC 1999b, Appendix 3, Table 1); denote the catch in year $t$ by $C_t$. The simple model is then

$$\eta_{t+1} = \eta_t - C_t + 1.5(\text{MSYR})\eta_t (1 - (\eta_t/K)^2)$$

where MSYR is a productivity parameter. The IWC has previously used this model to help understand the main challenges faced during bowhead stock assessment (Raftery et al., 1996; Raftery and Poole, 1997).

2.3 The Likelihood

The likelihood components we used are nearly identical to those used in the 1998 IWC bowhead assessment (IWC, 1999b). Given values for the population model parameters (for either model), the stock is projected through time using the historical catch data. Projected population abundance is recorded in the set of years denoted $T = \{1978, 1980, 1981, 1982, 1983, 1985, 1986, 1987, 1988\}$ and in 1993. The likelihood of the observed abundance data ($N_t$) given the model projected abundances ($\eta_t$) is given by

$$N_{1993} \sim N_9 \left( \eta_{1993}, 626^2 \right) \}
\log (N_{\text{past}}) \sim N_9 \left( b\eta, \hat{\Sigma} \right)$$

independently

where $N_{\text{past}}$ and $\eta$ are column vectors of $N_t$ and $\eta_t$ respectively for $t \in T$, $\hat{\Sigma}$ is the estimated variance-covariance matrix of $\log(N_{\text{past}})$ given in Table 3, and $b = 8293/7778$. The likelihood is set to zero if any non-positive simulated abundance occurs. The rationale for this likelihood is given by IWC (1999b). The observed abundance data are given in Table 2.

The simple model in Section 2.2 has two unknown parameters: MSYR and $K$. The annual catches $C_t$ are treated as known, fixed constants. Since MSYR and $K$ determine the $\eta_t$ and $\hat{\Sigma}$ is treated as known, the likelihood (2) can be used to estimate the model parameters.

It turns out that the joint likelihood surface for $K$ and MSYR generated by the simple model is similar to the likelihood surface for these parameters produced by the full BALEEN II model.
In both cases the region of highest likelihood is a strip featuring a strong non-linear relationship between MSYR and K. Also in both cases, very few parameter sets reflecting high productivity are compatible with the observed data even though the trajectories produced by such parameter sets exhibit some of the highest likelihoods. It is easier to match the observed data with parameter sets reflecting lower productivity, though the match may be slightly poorer.

For simplicity, therefore, we focus hereafter on the likelihood generated by the two-parameter version of the model. We will show in this paper that the model and these data generate certain unattractive features in the likelihood surface. It is important to keep in mind that these features are not attributable to the distributional choices in equation (2) but rather to the linkages among data, simulated abundances, and catch enforced by the population dynamics model itself. These problems can be generalized not only to BALEEN II, but also to many other population modelling applications.

3 Results

3.1 Features of the Likelihood

Figure 4 shows contours of the log likelihood\(^2\). Because of the unusual nature of this function, it is impossible to highlight all the important features of the likelihood in a single fixed-scale graph. For example, it is difficult to discern in Figure 4 that the highest region is on a thin curving ridge near where the contour lines are most closely bunched. The widely spaced contour lines in the right portion of the figure are at levels far below the ridge top. We call this ridge top area the support region of the likelihood, and we list its features below.

1. Nonlinearity. The support region emphasizes a strong, nonlinear relationship between MSYR and K, as depicted in Figure 4. The support region is not at all elliptical.

2. Near Dependency. For a given MSYR value, only a very small range of K values are supported by the likelihood. For a given K, only a narrow MSYR range is supported. On the region of likelihood support, the two parameters are nearly functionally related.

3. Cusp Shape. The support region is increasingly narrow for high MSYR values; any likelihood contour looks like a ramphoid cusp with the pointed end vanishing in the limit as MSYR increases. The severity of this feature cannot be shown graphically. Indeed, Table 4 shows approximate ranges of K (in units of whales) spanned by a joint 95% confidence region for MSYR and K, at selected MSYR values\(^3\). Conditional 95% confidence intervals given these MSYR values would be even narrower.

\(^2\)Small fragmented segments along the surface edge for higher MSYR values are an artifact of the limited numerical precision of the computer. The function is smooth but incredibly narrow and steep along that region.

\(^3\)This confidence region is a very poor one because it is based on a likelihood contour at a level corresponding to an appropriate \(\chi^2\) quantile (Wilks, 1938; Wald, 1943). Since the conditions necessary for \(\chi^2\) approximation of likelihood ratios clearly do not hold here, such intervals should not be used for inference. When we refer to likelihood-based confidence regions hereafter, the same concerns hold. We use these regions only to provide a convenient reference point for comparisons of the shape of the likelihood surface.
4. **Flat Top.** The region of dominant likelihood support is relatively flat. Figure 5 shows the log profile likelihood for MSYR, namely

\[ \log L^*(\text{MSYR}) = \log \max_K L(\text{MSYR}, K) \]  

with horizontal lines superimposed corresponding to cutoffs for selected joint confidence regions based on $\chi^2$ quantiles. Thus, the height of the curve is the height of the joint likelihood surface along the very ridge top of the support region, and the range of MSYR values spanned by the curve above each confidence line is the range of MSYR values contained within the corresponding joint confidence region. The likelihood is especially flat for high MSYR values. Even a mere 5% confidence region contains MSYR values stretching from .037 to .055.

5. **Steep Edges.** Figure 6 shows the log conditional likelihood of K when MSYR is set to equal its maximum likelihood estimate. The range of K values spanned by the curve above this line (namely 10778.29–10778.35) represents a conditional 95% confidence interval for K. The location of the peak of this curve is the conditional MLE for K. Note that the conditional likelihood of a K value 3 whales less than its conditional MLE is about $e^{200}$ times smaller than at the MLE. This astounding drop-off in the likelihood surface is why the likelihood vanishes in bottom left portion of Figure 4.

4 Discussion

4.1 Inference and Interpretation via the Likelihood

Figure 7 shows a joint 95% confidence region for MSYR and K, shown with the vertical axis denoting K minus the conditional MLE of K given MSYR. This axis straightens out the curve and mitigates scaling difficulties compared to Figure 4. The inset in Figure 7 is the same region plotted on standard axes. The shape of the region is difficult to discern in the inset due to extreme scale differences, however the region does enclose more area for small MSYR values, as is seen in the main portion of Figure 7. The joint MLE is indicated by point A in the figure; a hypothetical Bayes estimate is indicated by B.

Consider the interpretation and summary of the confidence region shown in Figure 7. The maximum likelihood occurs at A, but the region is quite flat so likely MSYR values span a wide range, including where B lies. The region includes values of MSYR and K which, given the model, could have plausibly generated the observed data. The cusp shape of the confidence region shows that, for high MSYR values, there is only a tiny range of compatible, likely K values. Does A or B better summarize this region? We would argue that B is a better one-number summary, even though A has a higher likelihood.

Summarizing the likelihood surface with a single central number is a point estimation problem. The calculation of the 95% confidence region in Figure 7 is an example of summarizing the uncertainty expressed in the likelihood. In Section 4.3 we examine estimation of uncertainty via the bootstrap.

---

4 Numbers near and beyond the right edge of the graph cannot be calculated reliably due to the limits of computer numerical precision. Thus, the true curve is probably even flatter since we were unable to find the true maxima sufficiently precisely for high MSYR.
The shape of the confidence region has implications for our view of the model itself. Imagine that we knew that MSYR=0.06. In that case, no-one would seriously believe that one fewer whales in 1848 would have meant extinction. In that case, reasonable people would reject the model as being intolerably misspecified and would not use it for inference. Now suppose that we knew MSYR=.06 or MSYR=.01, but we weren’t sure which. We plan to use the model to choose. If our model-based decision ends up favoring 0.06, we are clearly in a quandary because our modeled answer still indicates that one fewer whale in 1848 would have meant extinction. If our model-based decision ends up favoring 0.01, no troubling historical implication arises because extinction would not have occurred for range of K values, but may we relax? No—the model is still a flawed tool for making this decision. The fact that this flaw didn’t manifest itself when the estimate turned out as 0.01 does not mean the decision was made on a sound basis. The actual situation for the bowheads is essentially the same as this last case, but with MSYR completely unknown. By using this model we are implicitly agreeing that we would accept a high MSYR estimate even though the implications of that estimate violate common sense within the model used to estimate it. A sounder strategy would be to base the MSYR estimate on a model that has no silly implications over a wide range of MSYR values. A more defensible MSYR estimate would result since, regardless of its value, there would be no unintended implications due to the model.

An alluring counterargument to these concerns is that they disregard the fact that the model is only a simplistic deterministic approximation to a complicated stochastic process. Use of a stochastic dynamics model, according to this counterargument, would alleviate the problem. In fact it does not. The likelihood for a stochastic dynamics model looks quite similar to the one shown here and some of the same problems arise even more severely for that case. For example, with the deterministic model and MSYR=0.05, K may vary by only a couple of whales while keeping the stock trajectory non-zero and somewhat consistent with the observed data. Such trajectories must follow a very narrow path to achieve these goals. With a stochastic model, such trajectories are even more difficult to achieve because stochastic births and deaths randomly knock trajectories off the narrow desirable path.

Mathematical modelers can also argue that the enormous compounding effect of a small change in stock size over a long time period makes perfect mathematical sense. Furthermore, the intermediate effects of such changes could have been modified by a different exploitation pattern arising from the altered dynamics. Of course these arguments are correct as far as they go. However, population assessment inference is generally conditional on a historical catch record, which in turn was presumably related the single true dynamics history. Therefore, these arguments express ideas not realizable in the modeling framework provided here. To reflect the preceding arguments would require—at a minimum—building into the framework stronger dependencies between each year’s catch and the previous abundance. Without this, the class of stock trajectories and historical dynamics available to assessors through the model-induced likelihood function does not preclude nonsensical hyper-sensitivity in some cases.

The troubling narrowness of the likelihood support region and the strong dependencies it causes between parameter estimates arise because of the nature of the population dynamics model itself. Bowhead abundances are modeled each year to depend on the abundance and catch in the previous year. This calculation is repeated over a long period during which there are no available data to provide a reality check. This allows tiny changes in the starting conditions or early dynamics to result in huge changes to the stock trajectory in recent years, which, in turn, suggests that analysis of the likelihood might be very sensitive to the historical catch record.
To investigate such sensitivity, we considered two revisions to the simple model. The first approach simulated random historical catch record variation as suggested by Givens and Thompson (1996). The second approach assumed that the historical catch in year $t$ was

$$
\begin{align*}
\alpha_1 C_t & \quad \text{if } t \leq 1855 \\
\alpha_2 C_t & \quad \text{if } 1856 \leq t \leq 1879 \\
\alpha_3 C_t & \quad \text{if } t \geq 1880
\end{align*}
$$

(4)

where the $\alpha_i$ are three additional parameters in the likelihood function. The dates for switching $\alpha_i$ were chosen to correspond to well-known change-points in the commercial hunting process.

For three models (the simple model and the two catch-related variants above) we applied two estimation methods: MLE with a parametric bootstrap as suggested by Punt and Butterworth (1999b), and a Bayesian approach (analogous to analyses by Punt and Butterworth (1999a) and Poole and Raftery (1998)) with uniform priors. To gauge sensitivity, we examined the range of the three point estimates of MSYR resulting from each method. The range of the bootstrapped median MSYR estimates for the MLE approach was 0.0186 with the random approach giving the lowest MSYR estimate and the three-bias approach giving the highest MSYR estimate. The range of the posterior median MSYR estimates for the Bayesian approach was 0.0080 with the reverse endpoints. These results are surprising; both MLE and Bayesian estimation rely on the likelihood function and one would expect both to be similarly affected by changes to how the historical catch record is modeled. The results show instead that the MLE approach is more sensitive to such changes.

4.2 Implications for Maximizing the Likelihood

Maximization of a likelihood like the one described here is not routine; when the model has more than the two parameters used here, the challenge can be even greater.

The two main difficulties are identification of a starting value and reliable estimation of the local gradient. We have shown above that MSYR and $K$ are very closely linked on the support region of the likelihood. This means that virtually any starting value chosen is likely to have essentially zero likelihood unless it lies on the narrow support region. With sheer computational brute force this problem can often be overcome by pre-testing a large array of starting values.

A more serious problem occurs once optimization is underway: estimates of the local gradient at the current candidate point may be gravely mistaken because the likelihood is so steep at the edges of the support region and so flat within. This means that an optimization routine may not adequately assess whether further improvement is possible, and if so how it can be achieved.

Table 5 shows the results of an experiment that illustrates such optimization difficulties. The bowhead likelihood conditional on three MSYR values (0.07, 0.08, and 0.09) was maximized from three starting values for $K$ (9,000, 10,000, and 20,000) using a sophisticated quasi-Newton optimizer with a double dogleg step and the BFGS secant update to the Hessian (Dennis and Mei, 1979; Dennis et al., 1981). This optimizer is widely and routinely used by mathematicians and statisticians. The true maxima are also tabulated. The results in Table 5 show that the optimization software failed to find the correct maxima, and the errors were sometimes substantial. For comparative purposes, the 50%, 75%, and 95% joint confidence levels correspond to likelihood heights of -9.851, -10.544, and -12.154, respectively. The errors made by the optimizer are substantial on this scale.

The experiment summarized in Table 5 essentially replicates the creation of the right side of Figure 5. Clearly if we had relied on routine use of optimization code, this figure would have been
in error (because the right side of the figure would be jagged and generally lower). We relied on
tedious, manual optimization to ensure that Figure 5 shows the true curve. However, even this
effort was thwarted for large MSYR values because machine precision was surpassed by the precision
with which small changes in K caused large changes in the likelihood. At this point, optimization
was impossible. Thus, our figures cannot fully capture the support given by the likelihood to high
MSYR values; we suspect that recent maximum likelihood bowhead assessments are also susceptible
to this shortcoming.

4.3 Implications for the Bootstrap

The premise of a well-designed bootstrap of \( \hat{\theta} \) as an estimator of \( \theta \) is often that \( \hat{\theta} - \theta \) has the
same distribution as \( \hat{\theta}^* - \hat{\theta} \) where \( \hat{\theta}^* \) is the estimate from a bootstrap sample, and ideally that
\( \hat{\theta} - \theta \) is pivotal. A pivotal quantity is a function of the data and the unknown parameter whose
distribution does not depend on the value of that parameter. A chain of argument supporting
bootstrap inference in this case might be as follows:

1. Since \( \hat{\theta} - \theta \) is pivotal its distribution, say \( f_0 \), doesn't depend on \( \theta \); hence inference about \( \theta \)
can be achieved by reference to \( f_0 \).

2. \( f_0 \) is not known, but if \( \hat{\theta} - \theta \) and \( \hat{\theta}^* - \hat{\theta} \) have the same distribution, then the bootstrap
distribution of \( \hat{\theta}^* - \hat{\theta} \) can serve as an estimate of \( f_0 \).

In our case, with \( \theta = (\text{MSYR}, K) \), it is clear that \( \hat{\theta} - \theta \) is not pivotal since the distribution of the parameter estimates has a shape that depends strongly on the true values. This dependency follows from the nonlinearity of the ridge of support (see Figure 4) and the varying width of the
ridge induced by its cusp shape (see Figure 7). Nonpivotality is easily confirmed through simulation.

In several respects the bootstrap does a better job estimating the distribution of a pivotal statistic than a nonpivotal one (Hall, 1992). Efron (1987) and Beran (1987) showed that failure to
prepivot a bootstrap generally leads to inferior intervals. Efron (1987; 1992) also showed that the
simple percentile method, where confidence intervals are based on quantiles of the distribution of a
simple bootstrap sample, is generally inferior to other bootstrap methods. Although the bootstrap
is powerful and very useful, it should be carefully applied, especially when the underlying likelihood
is unusual.

We were concerned about whether the bootstrap likelihoods resembled the original likelihood.
As a test, we carried out the parametric bootstrap used in some assessments presented to the IWC
(Punt and Butterworth, 1999b) and evaluated the relative likelihood of \( \hat{\theta}^* \) and \( \hat{\theta} \) with respect to
the original observed data. In other words, we calculated \( L(\hat{\theta}^*)/L(\hat{\theta}) \) where \( L() \) is the likelihood
function with respect to the observed data, \( \hat{\theta} \) is the MLE, \( \hat{\theta}^* \) is the MLE given the bootstrapped
data, and \( \theta = (\text{MSYR}, K) \). For comparative purposes, we calculated the same ratio for the ordinary
nonparametric bootstrap solution to the problem of estimating \( \theta \) from \( X_1, \ldots, X_{20} \sim N(\theta, 1) \).

Figure 8 shows the results. Bootstrapping the bowhead assessment problem produced many
bootstrap estimates which were highly unrepresentative of the original likelihood, i.e. the ratio
of the bootstrap likelihood to the original likelihood was not close to 1. Bootstrapping the normal
problem produced bootstrap estimates which were very typical of the original solution. This
suggests that the bootstrap bowhead likelihoods generated in the manner described may not be a
reliable basis for inference.
4.4 Concluding Remarks

We have examined a wide variety of problems with the likelihood resulting from the dynamics model discussed here; model variations and additional model complexity do little to mitigate these problems. The primary cause of these problems is that the model and historical catch data together establish a nearly chaotic dynamical nonlinear system in the sense of Rasband (1990). Bowhead abundances are projected with self-referencing dynamics over a long uninterrupted historical period without reference to external data. This allows tiny changes in the early portion of the trajectory to cause huge changes in recent years, and it is only the recent years for which we have available data.

The likelihood problems introduced occur despite the seemingly paradoxical presence of very reasonable-looking stock trajectories for well-chosen parameter sets. Pointing to such trajectories is not a defense of the likelihood function generated by the model.

Although the sort of dynamics model examined here is very useful for producing highly realistic whale population trajectories, it is a poor model from which to generate likelihood functions for assessments like that of the bowhead whale. It is a good illustration of the types of problems that may arise in diverse modelling efforts when data from varied sources over a great time span are compared to simulated results from iterated compositions of a nonlinear model. Inferential methods applied to such likelihood functions require substantially greater justification and scrutiny than in ordinary statistical problems.

Acknowledgements

This work was partially supported by the North Slope Borough (Alaska), the State of Alaska (through the Alaska Department of Community and Regional Affairs), and the National Oceanic and Atmospheric Administration (through the National Marine Mammal Laboratory to the Alaska Eskimo Whaling Commission). We thank Michael C.S. Kingsley for thought-provoking comments on an earlier version of this paper.
References


IWC documents are available from the IWC Secretariat, The Red House, 135 Station Road, Impington, Cambridge, UK CB4 9NP.
<table>
<thead>
<tr>
<th>Label</th>
<th>Reference</th>
<th>Inference Method</th>
<th>Variant</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>AEP (1999)</td>
<td>Bayesian</td>
<td>‘Forwards’</td>
<td>Stochastic model</td>
</tr>
<tr>
<td>C</td>
<td>AEP &amp; DSB (1999a)</td>
<td>Bayesian</td>
<td>‘Backwards’</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>DP (1999)</td>
<td>Bayesian</td>
<td>‘Full Pooling’</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>DP (1999)</td>
<td>Bayesian</td>
<td>‘Backwards’</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>AEP &amp; DSB (1999b)</td>
<td>Constrained MLE</td>
<td></td>
<td>See caption</td>
</tr>
<tr>
<td>H</td>
<td>AEP &amp; DSB (1997)</td>
<td>Conditional MLE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>AEP &amp; DSB (1997)</td>
<td>Similar to MLE</td>
<td>‘Restrepo’</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>AEP &amp; DSB (1997)</td>
<td>Bayesian</td>
<td>‘Sideways’</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>AEP &amp; DSB (1997)</td>
<td>Bayesian</td>
<td>‘Less Both’</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Eleven recent assessments. The labels correspond to those in Figures 1 and 2. Names are abbreviated for Punt (AEP), Butterworth (DSB), Poole (DP), and Raftery (AER). From the 16 choices for ‘Restrepo’ assessment (I) in Table 5 of Punt and Butterworth (1997), chosen here is their recommended analysis except employing their ‘IWC M prior’ option.
<table>
<thead>
<tr>
<th>$t$</th>
<th>$N_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1978</td>
<td>4,820</td>
</tr>
<tr>
<td>1980</td>
<td>3,900</td>
</tr>
<tr>
<td>1981</td>
<td>4,389</td>
</tr>
<tr>
<td>1982</td>
<td>6,572</td>
</tr>
<tr>
<td>1983</td>
<td>6,268</td>
</tr>
<tr>
<td>1985</td>
<td>5,132</td>
</tr>
<tr>
<td>1986</td>
<td>7,251</td>
</tr>
<tr>
<td>1987</td>
<td>5,151</td>
</tr>
<tr>
<td>1988</td>
<td>6,609</td>
</tr>
<tr>
<td>1993</td>
<td>7,778</td>
</tr>
</tbody>
</table>

Table 2: Abundance data used in the likelihood equation (2).
<table>
<thead>
<tr>
<th>Year</th>
<th>Correlation matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>1978</td>
<td>1.000</td>
</tr>
<tr>
<td>1980</td>
<td>0.166</td>
</tr>
<tr>
<td>1981</td>
<td>0.054</td>
</tr>
<tr>
<td>1982</td>
<td>0.168</td>
</tr>
<tr>
<td>1983</td>
<td>0.163</td>
</tr>
<tr>
<td>1985</td>
<td>0.126</td>
</tr>
<tr>
<td>1986</td>
<td>0.080</td>
</tr>
<tr>
<td>1987</td>
<td>0.175</td>
</tr>
<tr>
<td>1988</td>
<td>0.038</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Correlation matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>1978</td>
<td>1.000</td>
</tr>
<tr>
<td>1980</td>
<td>0.166 1.000</td>
</tr>
<tr>
<td>1981</td>
<td>0.054 0.047 1.000</td>
</tr>
<tr>
<td>1982</td>
<td>0.168 0.146 0.047 1.000</td>
</tr>
<tr>
<td>1983</td>
<td>0.163 0.141 0.046 0.143 1.000</td>
</tr>
<tr>
<td>1985</td>
<td>0.126 0.109 0.025 0.110 0.107 1.000</td>
</tr>
<tr>
<td>1986</td>
<td>0.080 0.070 0.012 0.070 0.068 0.108 1.000</td>
</tr>
<tr>
<td>1987</td>
<td>0.175 0.152 0.049 0.154 0.149 0.115 0.074 1.000</td>
</tr>
<tr>
<td>1988</td>
<td>0.038 0.033 0.012 0.033 0.032 0.018 0.009 0.035 1.000</td>
</tr>
</tbody>
</table>

Table 3: Correlation matrix for the logarithms of the bowhead abundance indices for \( t \in T \).
<table>
<thead>
<tr>
<th>MSYR</th>
<th>Range of K</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>289.54</td>
</tr>
<tr>
<td>0.02</td>
<td>86.65</td>
</tr>
<tr>
<td>0.03</td>
<td>18.98</td>
</tr>
<tr>
<td>0.04</td>
<td>4.03</td>
</tr>
<tr>
<td>0.05</td>
<td>0.90</td>
</tr>
<tr>
<td>0.06</td>
<td>0.22</td>
</tr>
<tr>
<td>0.07</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table 4: Approximate ranges of K (in units of whales) spanned by a joint 95% confidence region for MSYR and K, at selected MSYR values.
<table>
<thead>
<tr>
<th>Log Likelihood Maxima</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>From Starting Value:</td>
<td>9,000</td>
<td>-10.800</td>
<td>-9.621</td>
</tr>
<tr>
<td></td>
<td>10,000</td>
<td>-9.365</td>
<td>-9.945</td>
</tr>
</tbody>
</table>

Table 5: True and estimated maxima of the likelihood at three fixed MSYR values. The estimates are obtained using a sophisticated quasi-Newton optimizer with three different starting values for K.
Figure 1: 90% intervals (bars) and point estimates (letters) from 11 assessments as labeled in Table 1.

Figure 2: Scatterplots of results from assessments listed in Table 1. Point estimates are used except for permissible catch, where the 5th percentile is shown.

Figure 3: The historical bowhead catch, 1848-1993.

Figure 4: The log likelihood function for the simple dynamics model.

Figure 5: Log profile likelihood for MSYR. Horizontal lines are superimposed at heights corresponding to the indicated quantiles of a $\chi^2_2$ distribution so that the range of MSYR values spanned by the curve above each confidence line is the range of MSYR values contained within a corresponding joint confidence region.

Figure 6: Conditional log likelihood for K when MSYR is set to equal its maximum likelihood estimate. The range of K values spanned by the curve above the superimposed line (namely 10778.29–10778.35) represents a conditional 95% confidence interval for K.

Figure 7: Joint 95% confidence region for MSYR and K, shown with the vertical axis denoting K minus the conditional MLE of K given MSYR. The joint MLE is indicated by point A in the figure; a hypothetical Bayes estimate is indicated by B. Inset is the same confidence region plotted on standard axes.

Figure 8: Distributions of relative likelihoods of $\hat{\theta}^*$ and $\hat{\theta}$ with respect to the original observed data for the bowhead assessment problem and the normal problem described in the text.
Figure 1: 90% intervals (bars) and point estimates (letters) from 11 assessments as labeled in Table 1.
Figure 2: Scatterplots of results from assessments listed in Table 1. Point estimates are used except for permissible catch, where the 5th percentile is shown.
Figure 3: The historical bowhead catch, 1848-1993.
Figure 4: The log likelihood function for the simple dynamics model.
Figure 5: Log profile likelihood for MSYR. Horizontal lines are superimposed at heights corresponding to the indicated quantiles of a $\chi^2$ distribution so that the range of MSYR values spanned by the curve above each confidence line is the range of MSYR values contained within a corresponding joint confidence region.
Figure 6: Conditional log likelihood for K when MSYR is set to equal its maximum likelihood estimate. The range of K values spanned by the curve above the superimposed line (namely 10778.29–10778.35) represents a conditional 95% confidence interval for K.
Figure 7: Joint 95% confidence region for MSYR and $K$, shown with the vertical axis denoting $K$ minus the conditional MLE of $K$ given MSYR. The joint MLE is indicated by point A in the figure; a hypothetical Bayes estimate is indicated by B. Inset is the same confidence region plotted on standard axes.
Figure 8: Distributions of relative likelihoods of \( \hat{\theta}^a \) and \( \hat{\theta} \) with respect to the original observed data for the bowhead assessment problem and the normal problem described in the text.