On a Proposed Stock Assessment Method and its Application at the International Whaling Commission [1]

David Poole
University of Washington

Geof H. Givens
Colorado State University

Adrian E. Raftery
University of Washington

Technical Report No. 318
Department of Statistics
University of Washington

April 8, 1997

[1]David Poole is Graduate Research Assistant, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, USA. Tel: (206) 543-8484; Fax: (206) 685-7419; E-mail: poole@stat.washington.edu. Geof H. Givens is Assistant Professor, Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA. E-mail: geof@space.stat.colostate.edu; Web: http://www.stat.colostate.edu/~geof. Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, USA. E-mail: raftery@stat.washington.edu; Web: http://www.stat.washington.edu/raftery.

Abstract

The Scientific Committee of the International Whaling Commission (IWC) has engaged in considerable debate over various methods of assessment for the Bering-Chuckchi-Beaufort Seas stock of bowhead whales. At the 1996 meeting of the Committee, it was proposed that a method developed by Restrepo et al. (1991; 1992) for fish stocks be extended and applied to bowheads. The idea behind the method is to quantify the uncertainty in the results of a population analysis via Monte Carlo simulation from distributions on the inputs to the assessment model and to combine this with estimation based on observed data. This underlying strategy, while intuitively appealing and easily applied to many fisheries management problems, can unfortunately lead to inaccurate results. As such, it appears to be unsuited to the bowhead application or more general fisheries management problems.

KEY WORDS: stock assessment, Monte Carlo simulation, biological parameter uncertainty, bootstrap.

Contents

List of Tables

1 Introduction

Over the past 6 years, various assessment methods for the Bering-Chukchi-Beaufort stock of bowhead whales have been discussed by the Scientific Committee of the International Whaling Commission (IWC). A Bayesian approach (Raftery et al. 1995) was adopted and used as the basis for the IWC assessment of the stock in 1994. The method was developed after a 1991 Scientific Committee recommendation that methods for taking full account of uncertainty about inputs and outputs to population dynamics models be developed. An alternative maximum likelihood approach (Butterworth and Punt, 1995; Punt and Butterworth, 1996a) has also been proposed. In contrast to the adopted method, this approach does not allow for any uncertainty in the values of various biological parameters. Rather, it assumes that they are known exactly.

At the 1996 SC meeting, a modified maximum likelihood method that can take account of uncertainty in biological parameters was proposed (Punt and Butterworth, 1996b). This method is an extension of an approach developed by Restrepo et al. (1991; 1992). Punt and Butterworth (1996b) cite an example of the use of the Restrepo method by the International Commission for the Conservation of Atlantic Tunas, and Restrepo et al. have applied their method to swordfish and cod fishery assessments.

In this paper, we review the Restrepo method, and the Punt and Butterworth extension thereof, and evaluate its compliance with established statistical principles. These methods are shown to have properties which render them undesirable for the bowhead problem and for the wider class fisheries management problems for which they were suggested. In order to illustrate the results, some simulations are performed. The paper is presented solely as a scientific appraisal of the Restrepo approach, and the examples given are purely illustrative. We do not propose any alternative assessment methods here.

2 The Restrepo method

2.1 Description

Restrepo et al. (1992) describe an approach to quantifying the uncertainty in the results of sequential population analyses for various fish stocks. They motivate their approach by noting that

> Fisheries managers recognize the dangers of accepting parameter estimates without consideration of the variability inherent in the estimates of fish stock status and related parameters.... If all sources of error are not appropriately accounted for, then estimates of the uncertainty in the assessment results may be too small.

Their Monte Carlo approach proceeds as follows. Probability distributions ("priors"[1]) describe the uncertainty in the inputs to an assessment model. Model outputs such as a time trajectory of stock sizes are compared to observed data to formulate a likelihood (e.g. assuming lognormal deviations). Many parameter sets are drawn randomly from the specified "prior" distributions, and for each set, a **conditional** maximum likelihood estimate (MLE) is calculated for quantities of interest, given the fixed input parameters and the observed data. The simulation distribution of such conditional MLEs is used as an uncertainty estimate. The simulation is viewed as translating input uncertainties into output uncertainties.[2]

The Restrepo method is suggested for situations where (possibly many) nuisance parameters exist. The basic strategy is to estimate the quantities of interest (eg. current stock size and production rate) conditional on values of the nuisance parameters, and then to integrate over the "prior" for the nuisance parameters. This integration is achieved using the Monte Carlo simulation described above. The distribution of the conditional estimates of the parameters of interest is then examined for the purposes of inference. For example, if $\hat{\theta}_\gamma$ is an estimator of θ conditional on nuisance parameters γ, then

$$\hat{\theta}_{(1)} = \int \hat{\theta}_\gamma p(\gamma) d\gamma \tag{1}$$

is a Restrepo estimate of θ, where $p(\gamma)$ is the "prior" for γ.

The Restrepo approach is flawed. If one is going to express uncertainty in the value of a parameter by means of a probability distribution, then this distribution should be treated as a prior in a fully Bayesian setup. The method given in (1) can be described as an ad hoc partially Bayesian approach. As a result, estimators obtained using the method will not necessarily possess the desirable properties of either Bayesian or ML estimators.

In a Bayesian framework, the best estimator of θ (with respect to squared error loss) is the posterior mean, so it would be better to define the estimator as:

$$\hat{\theta}_{(2)} = E(\theta|\text{data}) = E[E(\theta|\gamma, \text{data})] = \int \hat{\theta}_\gamma \pi(\gamma|\text{data}) d\gamma \tag{2}$$

where $\hat{\theta}_\gamma$ is the posterior mean of θ conditional on γ, and $\pi(\gamma|\text{data})$ is the posterior distribution of γ. One might regard (2) as a general strategy and use it in cases where $\hat{\theta}_\gamma$ is not necessarily the Bayesian estimator. In this case, however, the properties of $\hat{\theta}_{(2)}$ are not clear.

[1]Since the framework of the method is not Bayesian, this "prior" distribution cannot be considered a true prior in a fully Bayesian context.

[2]Restrepo et al. also consider uncertainty in the assessment model itself, but this issue is not of primary interest here.

2.2 A Simple Example

Schweder and Hjort (1996) first identified potential weaknesses with the Restrepo method, and they described two situations where differences between the methods of (1) and (2) arose. The first of these is repeated here: let $x_1, ..., x_n$ be independent $N(\mu, \sigma^2)$, and assume there is a $N(\mu_0, \tau_0^2)$ prior for μ. Let σ^2 be the parameter for which inference is desired; μ is a nuisance parameter. σ^2 is regarded as fixed. The maximum likelihood estimate of σ^2 conditional on μ is

$$\hat{\sigma}_\mu^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

The estimators given by (1) and (2) are then

$$\hat{\sigma}_{(1)}^2 = E_{prior}(\hat{\sigma}_\mu^2) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2 + \tau_0^2$$

and

$$\hat{\sigma}_{(2)}^2 = E_{posterior}(\hat{\sigma}_\mu^2) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{post})^2 + \tau_{post}^2$$

respectively, where μ_{post} and τ_{post}^2 are the posterior mean and variance of the nuisance parameter μ. Note that in (2), $\hat{\sigma}_\mu^2$ (the conditional maximum likelihood estimator of σ^2) is being used as $\hat{\theta}_\gamma$. In addition we note that $\hat{\sigma}_{(2)}^2$ depends on σ^2 because both μ_{post} and τ_{post}^2 are functions of σ^2. In other words, the estimator $\hat{\sigma}_{(2)}^2$ depends on the quantity it is trying to estimate. This occurs when $\hat{\theta}_\gamma$ is not the Bayesian posterior mean of θ conditional on γ. In our examples, we simply plug in the ordinary MLE

$$\hat{\sigma}_{std}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

where needed to remove this dependency. Thus, to evaluate $\hat{\sigma}_{(2)}^2$ here, $\hat{\sigma}_{std}^2$ was used as a plug-in estimate of σ^2 in the expressions for μ_{post} and τ_{post}^2. $\hat{\sigma}_{std}^2$ is the usual MLE of σ^2 and is what would normally be used if conditioning on μ was not of interest.

Since $\tau_{post}^2 \to 0$ and $\mu_{post} \to \mu$, we have that $\hat{\sigma}_{(2)}^2 \to \sigma^2$ as $n \to \infty$. $\hat{\sigma}_{(1)}^2$, on the other hand, will converge to σ^2 only if $\mu_0 = \mu$ and $\tau_0^2 = 0$. It follows that $\hat{\sigma}_{(1)}^2$ will, in general, only yield accurate estimates if μ_0 is close to μ and τ_0^2 is small. We know from standard theory that $\hat{\sigma}_{std}^2 \to \sigma^2$ as $n \to \infty$.

To investigate the difference between the Restrepo method in (1) and the ad hoc "Bayes" approach in (2), we performed some simple simulations. For each of 9 combinations of μ_0 and τ_0^2, a random sample of size $n = 1,000$ was drawn from a $N(\mu = 50, \sigma^2 = 100)$ distribution. In each case, $\hat{\sigma}_{(1)}^2$, $\hat{\sigma}_{(2)}^2$, and $\hat{\sigma}_{std}^2$ were calculated. The results are shown in Table 1.

μ_0	τ_0^2	Estimators of σ^2 (true value is 100)		
		Restrepo: $\hat{\sigma}^2_{(1)}$	"Bayes": $\hat{\sigma}^2_{(2)}$	Standard: $\hat{\sigma}^2_{std}$
50	1	98.1	97.2	97.2
	9	112.9	103.8	103.8
	25	127.5	102.3	102.4
60	1	197.6	103.4	102.7
	9	199.2	94.6	94.6
	25	211.7	97.1	97.1
70	1	517.6	96.9	93.4
	9	523.2	101.7	101.7
	25	528.8	99.7	99.7

Table 1: Simulation results for the example of Schweder and Hjort (1996)

The results show very poor performance for $\hat{\sigma}^2_{(1)}$, and good, similar performances for $\hat{\sigma}^2_{(2)}$ and $\hat{\sigma}^2_{std}$. If conditioning on nuisance parameters is to be used here, the strategy presented in (2) appears to be preferable to the Restrepo approach in (1). As discussed earlier, however, method (2) has severe limitations of its own and we do not regard it as a viable alternative approach. Indeed, there is a considerable literature on the role of conditioning in inference. Reid (1995) presents a review of recent developments.

3 An Extension of the Restrepo Method

3.1 Description

Punt and Butterworth (1996b) present an extension of the original Restrepo method for use in bowhead whale assessment. In addition to the Monte Carlo simulation from a "prior" for the nuisance parameters, this extended version involves generation of pseudo-data using a parametric bootstrap of the observed data. The conditional estimator $\hat{\theta}_\gamma$ is then obtained using sampled values of the nuisance parameters and the pseudo-data from the bootstrap. In the form of an algorithm, the method proceeds as follows:

1. Obtain the MLE $\hat{\theta}_{\gamma_0}$ based on the observed data X and a likely value γ_0.

2. Sample γ^* from the "prior" $p(\gamma)$.

3. Use the parametric bootstrap to sample pseudo-data X^* from a distribution with density $f(x; \hat{\psi}(X))$, where $f(x; \psi)$ is a model for the data but not necessarily the

assessment model, and $\hat{\psi}(X)$ is an estimate of the parameters of this model. $\hat{\psi}(X)$ may depend on the results of step 1, namely $\hat{\theta}_{\gamma_0}$ and γ_0, or even on standard MLEs $\hat{\theta}$ and $\hat{\gamma}$. An example of f would be to assume $X^* \sim N(X, \hat{\psi})$ where $\hat{\psi}$ is an estimated dispersion matrix.

4. Find the conditional MLE $\hat{\theta}_{\gamma^*}$ given γ^* using the bootstrap data X^*. Store this value.

5. Repeat steps 2-4 many times to obtain a collection of $\hat{\theta}_{\gamma^*}$'s.

Inference about θ is then based on the distribution of this collection of $\hat{\theta}_{\gamma^*}$'s. Typically, the original $\hat{\theta}_{\gamma_0}$ or the mean or median of the Monte Carlo sample is used as a point estimate, and the 0.025 and 0.975 quantiles form the bounds of a 95% confidence interval.

3.2 Example : General confidence interval estimation

The example in Section 2.2 illustrates poor performance for the original Restrepo method with regards to point estimation. Similar problems occur with confidence interval estimation in the bootstrap extension of the method.

Consider the following example: $X_i \sim$ i.i.d. $\mathrm{U}(\gamma - \theta\gamma, \gamma + \theta\gamma)$, for $i = 1, \ldots, 100$, where γ is a nuisance parameter. The MLEs are $\hat{\gamma} = (\max X_i + \min X_i)/2$ and $\hat{\theta} = (\max X_i - \min X_i)/(\max X_i + \min X_i)$. The conditional MLE for θ given γ is $\hat{\theta}_\gamma = (\max X_i - \min X_i)/(2\gamma)$.

Suppose the nuisance parameter, γ, has a $\mathrm{U}(a, b)$ prior, $0 \leq a < b$. The extended Restrepo method would proceed as follows:

1. Sample $\gamma^* \sim \mathrm{U}(a, b)$.

2. Use the parametric bootstrap to sample $X_i^* \sim \mathrm{U}(\hat{\gamma} - \hat{\theta}\hat{\gamma}, \hat{\gamma} + \hat{\theta}\hat{\gamma})$, $i = 1, \ldots, 100$. In this example, the parametric bootstrap model is the same as the original model.

3. Find the conditional MLE, $\hat{\theta}_{\gamma^*}$, using the bootstrap data, and conditioning on the current γ^*.

4. Store $\hat{\theta}_{\gamma^*}$ and go to step 1. Use the collection of $\hat{\theta}_{\gamma^*}$'s to obtain a confidence interval using the quantile method.

The ad hoc "Bayes" method which relies on sampling from the posterior proceeds with the same steps, except that step 1 is replaced by

1.* Sample γ^* from its posterior distribution.

Here, if we think of the likelihood as a function of γ only, then the posterior for γ is proportional to:

$$\prod_i \left[\frac{1}{2\theta\gamma} I(\gamma - \theta\gamma \leq X_i \leq \gamma + \theta\gamma)\right] I(a \leq \gamma \leq b).$$

Since this posterior depends on θ, we plug in $\hat{\theta}$ in the same way as we did for the second method in Section 2.2.

In a simulation example, we let $\theta = 4$, $\gamma = 4$, $a = 0$, and $b = 8$ so the true value for γ is at the midpoint of its prior. The MLEs were $\hat{\theta} = 4.06$ and $\hat{\gamma} = 3.88$. The 95% Restrepo interval for $\hat{\theta}$ was (1.99, 87.55). The interval obtained with the second method was (3.78, 4.03). The Restrepo interval is 342 times wider than the second interval, and 2.7 times wider than the range of the observed data.

Note that the MLE $\hat{\theta}$ is not contained in the second interval. This is an undesirable result of the second method here. One possible remedy is as follows: instead of simply plugging $\hat{\theta}$ into the posterior for γ each time we sample from it, we attempt to "integrate over theta" by plugging in a different estimate of θ on each occasion. Each of these plug-in values for θ can be obtained by calculating the MLE of θ from a nonparametric bootstrap sample of the real data X. The confidence interval resulting from this strategy was (3.59, 4.08), so the MLE is now contained in the interval. We stress that this is again an ad hoc solution and we strongly favor standard maximum likelihood or Bayesian methods over either Restrepo or the ad hoc "Bayes" method.

Finally, the example can be twisted so that the Restrepo interval is too narrow, also. In fact, one may design all sorts of behavior through the choice of the function of γ specified by $\hat{\theta}_\gamma$. For example, suppose $X_i \sim$U$(\gamma - \theta\gamma^2, \gamma + \theta\gamma^2)$. This leads to $\hat{\theta}_\gamma = (\max X_i - \min X_i)/(2\gamma^2)$, a cusp-shaped function of γ over any interval that includes 0. Thus, considering γ priors of the form U$(-a, a)$ for $a > 0$, the Restrepo method leads to the surprising result that the width of a quantile-based confidence interval for θ approaches 0 as a increases, while holding the observed data fixed. In other words, the width of the confidence interval is entirely dependent on the prior, and wider priors lead to narrower Restrepo confidence intervals.

3.3 Example : Bowhead whale assessment

As noted in Section 3.1, Punt and Butterworth (1996b) propose the use of the extended Restrepo approach in the assessment of the Bering-Chukchi-Beaufort stock of bowhead whales. They generate pseudo-data using a parametric bootstrap. The data consist of abundance estimates and corresponding CV estimates for several years, and observed age class proportions. Specifically, each simulation consists of the following:

1. *Bootstrap the data.* A series of pseudo-abundance estimates is bootstrapped from the observed data (Punt and Butterworth's Table 1). Each estimate is assumed to be independent and from a lognormal distribution with mean and CV equal to the observed estimates from that year. Pseudo-data for fractions of calves and matures are generated from Table 4 of IWC (1995).

2. *Sample biological nuisance parameters.* Parameters such as age-at-maturity and natural mortality rates are generated from distributions from IWC (1995).

3. *Conditional estimation.* Conditional on the values of the nuisance parameters, maximum likelihood estimation is used with an age-structured density dependent population dynamics model to obtain estimates of the parameters of interest: carrying capacity (K) and a productivity parameter (MSYR). The likelihood contains contributions from both the abundance and proportion data.

4. *Uncertainty estimation.* The variation in conditional MLEs is used to represent uncertainty.

Note that K and MSYR are the parameters of interest θ in our previous notation, while the other biological parameters take the role of γ. The distributions from which they were simulated are the "prior" distributions. Punt and Butterworth (1996b) use the results of 1,000 replications of this procedure to form confidence intervals.

3.3.1 A Simple Population Dynamics Model

For the purposes of illustration, we apply this extended Restrepo method to the simple population dynamics model (PDM) used in Raftery, et al. (1996). This is a non-age-structured density dependent PDM given by:

$$P_{t+1} = P_t - C_t + 1.5(\text{MSYR})P_t(1 - (P_t/K)^2) \quad (3)$$

where P_t is the population in year t with $t = 0$ corresponding to the baseline year before commercial hunting started (here 1848), K is the initial population size or carrying capacity, MSYR is the maximum sustainable yield rate of production as a proportion of the population aged 1+, and C_t is the number of whales killed by hunting in year t (known exactly). This model is much simpler than the BALEEN II PDM (de la Mare and Cooke, 1993) used by the IWC for bowhead assessment, but it nevertheless captures many of the essential features of the bowhead population. Because it has only 2 inputs (K and MSYR) and one output (P_{1993}), it will be easier for us to evaluate the Restrepo method using this model. Specifically, we will treat MSYR as a nuisance parameter (γ in our previous notation) while K will be the parameter of interest (the θ from before).

3.3.2 Application of the Restrepo Method

The extended Restrepo method is applied to the simple PDM of (3) in a number of steps. First, "true" values of K and MSYR are selected, and the PDM is run to obtain the "true" value of P_{1993}. Since K (the parameter of interest here) has a "true" value, we will be able to assess the accuracy of the estimates produced by the simulations.

The "prior" for the nuisance parameter MSYR is Gamma(8.2, 372.7), and the likelihood for the observed total population in 1993 is $N(P_{1993}, 626^2)$. These choices are based on IWC consensus (IWC, 1995) and are the same as used in previous work (Raftery, et al. 1996).

We assume that we have a single observation from the likelihood for P_{1993}. In practice this is usually obtained via a census. The observation is typically a maximum likelihood estimate of P_{1993}, so we denote it by $\hat{P}_{1993}$.

An original conditional MLE is obtained by conditioning on a "likely" point estimate of MSYR, say MSYR_0. We choose $\text{MSYR}_0 = 0.02$, the mean of the "prior" for MSYR. The model is then run backwards (ie. with $\hat{P}_{1993}$ and MSYR_0 as inputs). The resulting output is the conditional maximum likelihood estimate $\hat{K}_{MSYR_0}$ of K because (given MSYR_0) it leads to the value of P_{1993} that maximizes the likelihood.

The bootstrapped Restrepo procedure then proceeds as follows:

1. Draw $\hat{P}^*_{1993}$ from $N(\hat{P}_{1993}, 626^2)$. This is the bootstrap from a distribution with mean given by the observed total population in 1993.

2. Draw MSYR^* from the "prior" for MSYR.

3. Obtain $\hat{K}_{MSYR^*}$ by running the model backwards with $\hat{P}^*_{1993}$ and MSYR^* as inputs.

4. Repeat steps 1-2 many times to form a collection of $\hat{K}_{MSYR^*}$ estimates. Like Punt and Butterworth (1996b), we use 1,000 replications.

5. Use $\hat{K}_{MSYR_0}$ and the distribution of the $\hat{K}_{MSYR^*}$ estimates to obtain inference about K. Specifically, the distribution of the $\hat{K}_{MSYR^*}$ estimates shows how the conditional MLE of K changes as MSYR is varied according to its prior.

3.3.3 Simulation results

Simulations were performed using three sets of "true" parameters as shown in Table 2. The values of MSYR in Sets 1, 2, and 3 correspond to the 0.5, 0.975, and 0.025 quantiles (respectively) of the Gamma (8.2, 372.7) "prior" for this parameter. In this way, we investigate the performance of the method when the true MSYR is at the center and the boundaries of

	MSYR	K	P_{1993}
Set 1	0.02	14,700	8,733
Set 2	0.04	11,300	9,896
Set 3	0.01	18,500	6,971

Table 2: True values of parameters used in the bowhead simulations

MSYR	K	P_{1993}	$\hat{P}_{1993}$	$\hat{K}_{MSYR_0}$	Quantiles of $\hat{K}_{MSYR^*}$			Coverage
					0.025	0.5	0.975	of 95% CI's
0.02	14,700	8,733	7,506	14,640	11,230	14,360	18,870	100%
			8,733	14,700	11,260	14,400	19,160	100%
			9,960	14,780	11,330	14,450	19,540	100%
0.04	11,300	9,896	8,669	14,700	11,250	14,400	19,120	68%
			9,896	14,780	11,290	14,470	19,450	55%
			11,123	14,880	11,580	14,560	19,840	0%
0.01	18,500	6,971	5,744	14,570	11,270	14,300	18,430	38%
			6,971	14,620	11,260	14,340	18,710	80%
			8,198	14,670	11,270	14,380	19,000	99%

Table 3: Simulation results for the extended Restrepo method applied to the simple bowhead whale PDM.

its "prior" likely interval. In each case, a value of K is chosen such that extinction does not occur and hence, that P_{1993} will be positive.

The results of the simulations are shown in Table 3. For each of the three sets of true parameters in Table 2, the simulation was run three times using the 0.025, 0.5, and 0.975 quantiles of the normal likelihood as the observed 1993 population, $\hat{P}_{1993}$. Then, this entire simulation design was replicated 500 times. The quantiles shown are medians across the 500 replicates, and the confidence interval coverage rates show the percentage of the 500 replicates for which the estimated 95% confidence interval covered the truth.

In the first set of simulations, where the true MSYR and the "prior" mean MSYR_0 are exactly equal, the conditional MLE $\hat{K}_{MSYR_0}$ is very good. This is to be expected in this optimistic (if somewhat unlikely) scenario. The confidence intervals provided by $\hat{K}_{MSYR^*}$ cover the true K in all cases. Also, the estimation is fairly insensitive to the accuracy of the estimate $\hat{P}_{1993}$. A difference of 1,227 whales in the estimate of P_{1993} (ie. two standard deviations) results in $\hat{K}_{MSYR_0}$ changing by less than 100 whales.

In the second set, the true value of MSYR is greater than the "prior" mean. This results in K being overestimated. Note that the overestimate is worse when $\hat{P}_{1993}$ is accurate than when $\hat{P}_{1993}$ is two standard deviations too small. This is not desirable - the performance of any inference technique should improve as the quality of the data improve. The 95% intervals provide very poor coverage, worse when $\hat{P}_{1993}$ is accurate than when it is too low.

In the final set, the true value of MSYR is at the low end of the "prior" interval, and we observe that K is underestimated by about 4,000 whales for all three values of $\hat{P}_{1993}$. The estimates of K are inaccurate regardless of the accuracy of $\hat{P}_{1993}$. Here also the 95% interval fails to cover the truth sufficiently often when $\hat{P}_{1993}$ is accurate or too small. Since a low true value of MSYR is of particular concern to the IWC, it is worrying that the method performs poorly in this specific case.

As a final point, these simulations and others we have run suggest that the estimates of K are heavily dependent on the "prior" distribution for MSYR, to the point of being undesirably insensitive to the true values of K, P_{1993}, and the data $\hat{P}_{1993}$. This behavior is more extreme than would be the case if the prior was updated to a posterior in a fully Bayesian framework. This concurs with the results in Section 2.2 where the Restrepo estimates were greatly influenced by the "prior" mean and variance.

3.4 Use of the Bootstrap

As in the original method proposed by Restrepo et al. (1992), the Punt and Butterworth (1996b) extension to bowhead whale assessment integrates conditional estimators over "priors" for nuisance parameters. We have shown this approach can lead to poor estimates. Now we show that Punt and Butterworth's suggested approach for weaving the bootstrap into the procedure is also worrisome.

The difficulties are illustrated by applying the Punt and Butterworth method to ordinary simple linear regression. By shifting from the complex whale assessment problem to this example, the essential points are brought to the forefront. For whales, there are periodic surveys that lie near some non-linear, parameterized population model trajectory, and the goal is to estimate key parameters that determine the trajectory. In our linear regression example, there are data spread out in X (analogous to time) which lie near some linear, parameterized model line, and the goal is to estimate the parameters that determine the line.

Suppose the regression data are $Y_i \mid X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$. A standard bootstrap would proceed by *resampling the residuals* (e.g. Efron and Gong, 1983, p. 43). Fit the model, and find the residuals $e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$. Resample the residuals with replacement, obtaining pseudo-residuals e_i^*. Add these scrambled residuals to the fitted values to obtain pseudo-data $Y_i^* = \hat{Y}_i + e_i^*$. Regress Y_i^* on X_i to obtain bootstrap estimates $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$. Repeat to form a large collection of bootstrap estimates, representing the bootstrap distribution. Modifications for unequal data precision and for the parametric bootstrap are straightforward, but they all involve the key idea of bootstrapping the residuals. The residuals should be approximately i.i.d. and hence suitable for resampling, but the Y_i are not.

The Punt and Butterworth approach allows a different model for the bootstrapping than is used to fit the data. In the whale case, their bootstrapping distributions are lognormal centered around the data, whereas the model is fit assuming lognormal distributions centered around the model trajectory. An analogy for the regression case would be if Punt and Butterworth were to bootstrap $Y_i^*|X_i, Y_1, \ldots, Y_n \sim N(Y_i, \tau^2)$, where τ^2 plays the role analogous to each survey CV in the whale case. Here we make the additional simplifying assumption of equal precision for all surveys[3]. The variance of their bootstrapped Y_i^* is $\sigma^2 + \tau^2$, under the model assumption that $Y_i \mid X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$. In contrast, the standard bootstrapping approach would result in $Y_i^*|X_i, Y_1, \ldots, Y_n \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, with variance σ^2.

Thus, when obtaining bootstrap parameter estimates from the bootstrap sample, the Punt and Butterworth approach introduces excess variability, effectively down-weighting the data. This excess variability is propagated to the standard errors or confidence intervals for the estimated model parameters. Since in many whale assessments τ^2 and σ^2 can be roughly the same magnitude, the Punt and Butterworth bootstrap approach could double the bootstrap sampling variance.

4 Conclusion

Instead of an ad hoc blend of Bayesian and maximum likelihood techniques like that of the Restrepo approach and Punt and Butterworth's extension, analysis should be based on fully Bayesian or classical frequentist approaches. As our examples illustrate, the use of ad hoc variants of either methodology can result in unreliable inference, whereas the optimal properties of more standard methods are well known. These conclusions hold specifically for the bowhead assessment and for more general stock assessment problems.

Our examples show important shortcomings in the Punt-Butterworth and Restrepo approaches, but they are not intended to denigrate the considerable contributions made by these scientists. The underlying problem is quite a challenging one, and ideas that ini-

[3]In the case of unequal variances, as the Punt and Butterworth whale analysis, the additional steps of scaling the residuals for constant variance before resampling, and rescaling by the standard deviation of the newly matched observation after resampling, are required. However, these implementation details are not relevant to the essential argument in the text.

tially seem promising may not hold up under closer scrutiny. Nevertheless, such ideas are a valuable part of the scientific process. Useful assessment approaches are built upon the foundation of what has been learned so far.

Acknowledgements

This research was supported by the North Slope Borough, Alaska, the State of Alaska (through the Alaska Department of Fish and Game), and the National Oceanic and Atmospheric Administration (through the National Marine Mammal Laboratory and the Alaska Eskimo Whaling Commission). The authors are grateful to Doug Butterworth, Carey Priebe, André Punt, Tore Schweder, Tim Smith, Paul Wade and Judy Zeh for helpful comments and input.

References

Butterworth, D.S. and Punt. A.E. 1995. On the Bayesian approach suggested for the assessment of the Bering-Chukchi-Beaufort Seas stock of bohwead whales. *Rep. int. Whal. Commn* 45:303-11.

de la Mare, W.K. and Cooke, J.G. 1993. BALEEN II: The population model used in the Hitter-Fitter Programs. Unpublished manuscript available from the IWC Secretariat.

Efron, B. and Gong, G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician* 37:36-48.

International Whaling Commission, 1995. Report of the Scientific Committee. *Rep. int. Whal. Commn* 45:53-222.

Punt, A.E. and Butterworth, D.S. 1996a. Further remarks on the Bayesian approach for assessing the Bering-Chukchi-Beaufort Seas stock of bowhead whales. *Rep. int. Whal. Commn* 46:481-91.

Punt, A.E. and Butterworth, D.S. 1996b. Assessments of the Bering-Chukchi-Beaufort Seas stock of bowhead whales (*Balaena mysticetus*) using maximum likelihood and Bayesian methods. Paper SC/48/AS1 presented to the IWC Scientific Committee, June 1996.

Restrepo, V.R., Powers, J.E. and Turner, S.C. 1991. Incorporating uncertainty in VPA results via simulation. *Collve. Vol. Scient. Pap. ICCAT* 35:355-61.

Restrepo, V.R., Hoenig, J.M., Powers, J.E., Baird, J.W. and Turner, S.C. 1992. A simple simulation approach to risk and cost analysis, with applications to swordfish and cod fisheries. *Fishery Bulletin* 90:736-748.

Raftery, A.E., Givens, G.H. and Zeh, J.E. 1995. Inference from a deterministic population dynamics model for bowhead whales (with Discussion and Rejoinder). *J. Amer. Statist. Ass.* 90:402–430.

Raftery, A.E., Poole, D. and Givens, G.H. 1996. The Bayesian Synthesis assessment method: resolving the Borel Paradox and comparing the backwards and forwards variants. Paper SC/48/AS16 presented to the IWC Scientific Committee, June 1996.

Reid, N. 1995. The roles of conditioning in inference. *Statistical Science* 10:138-157.

Schweder, T. and Hjort, N.L. 1996. Confidence Likelihood and Likelihood Synthesis. Presented at the Joint Statistical Meetings, August 1996.