State-space likelihoods for nonlinear fisheries time-series

Perry de Valpine and Ray Hilborn

Abstract: State-space models are commonly used to incorporate process and observation errors in analysis of fisheries time series. A gap in analysis methods has been the lack of classical likelihood methods for nonlinear state-space models. We evaluate a method that uses weighted kernel density estimates of Bayesian posterior samples to estimate likelihoods (Monte Carlo Kernel Likelihoods, MCKL). Classical likelihoods require integration over the state-space, and we compare MCKL to the widely used errors-in-variables (EV) method, which estimates states jointly with parameters by maximizing a nonintegrated likelihood. For a simulated, linear, autoregressive model and a Schaefer model fit to cape hake (*Merluccius capensis* × *M. paradoxus*) data, classical likelihoods outperform EV likelihoods, which give asymptotically biased parameter estimates and inaccurate confidence regions. Our results on the importance of integrated state-space likelihoods also support the value of Bayesian analysis with Monte Carlo posterior integration. Both approaches provide valuable insights and can be used complementarily. Previously, Bayesian analysis was the only option for incorporating process and observation errors with complex nonlinear models. The MCKL method provides a classical approach for such models, so that choice of analysis approach need not depend on model complexity.

Résumé : Les modèles d'espace d'états servent couramment à incorporer les erreurs de processus et d'observation dans l'analyse des séries chronologiques en pêcheries. L'absence de méthodologie classique de vraisemblance pour les modèles d'espace d'états non linéaires est une carence dans ces méthodes analytiques. Nous évaluons une méthode qui utilise des estimations de la densité pondérée des noyaux des échantillons bayésiens a posteriori pour estimer les vraisemblances (MCKL, vraisemblances kernel de Monte Carlo). Les vraisemblances classiques requièrent une intégration dans l'espace des états ; nous comparons donc la méthode MCKL à la méthode très utilisée des erreurs sur les variables (EV) qui estime les états conjointement avec les paramètres en maximisant une vraisemblance non intégrée. Dans l'ajustement d'un modèle d'autorégression linéaire simulé et d'un modèle de Schaefer à des données sur des merlus de cap (*Merluccius capensis* × *M. paradoxus*), les vraisemblances classiques fonctionnent mieux que les vraisemblances EV qui fournissent à l'asymptote des estimations distordues des paramètres et des régions de confiance erronées. Nos résultats sur l'importance des vraisemblances intégrées d'espace d'états appuient aussi la valeur de l'analyse bayésienne avec une intégration Monte Carlo a posteriori. Les deux méthodologies fournissent des perspectives précieuses et peuvent être utilisées de façon complémentaire. Antérieurement. l'analyse bayésienne était la seule manière d'incorporer les erreurs de processus et d'observation dans des modèles non linéaires complexes. La méthode MCKL fournit une approche classique pour ces modèles et le choix de la méthode d'analyse ne dépend plus nécessairement de la complexité du modèle.

[Traduit par la Rédaction]

Introduction

A major challenge of fitting fisheries models to time-series data has been to incorporate the combined effects of process and observation error in the framework of state-space models, which combine a stochastic model for population (state) dynamics with a stochastic model for observations or catches (Schnute 1994). The primary methods for fitting state-space models have been the Kalman Filter (KF) for classical analysis with linear models (Sullivan 1992; Pella 1993; Gudmundsson 1994), the "errors-in-variables" (EV) method, in which the joint likelihood of states and population dynamics parameters is maximized (Ludwig and Walters 1981; Ludwig et al. 1988; Richards et al. 1997), and the Bayesian approach, usually implemented by a Markov Chain Monte Carlo (MCMC) sample of the posterior distribution of states and parameters (McAllister et al. 1994; Meyer and Millar 1999; Schnute and Kronlund 2002).

From the point of view of asymptotic likelihood theory, which motivates both Bayesian and classical analysis with the assurance that more data provide better estimates on average, it is desirable to have methods that calculate the likelihood of population dynamics parameters by integrating over the state-space of population trajectories. The KF for classical analysis and MCMC posterior sampling for Bayesian analysis both accomplish this integration, but the KF is limited to linear models. Thus, a gap in available methodology has been the lack of methods to conduct classical analysis with nonlinear state-

Received 9 March 2004. Accepted 4 March 2005. Published on the NRC Research Press Web site at http://cjfas.nrc.ca/ on 30 August 2005.
J18020

Perry de Valpine.[1,2] National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, 735 State Street, Suite 300, Santa Barbara, CA 93101, USA.
Ray Hilborn. School of Aquatic and Fishery Sciences, University of Washington, Box 355020, Seattle, WA 98195, USA.

[1] Present Address: University of California, Berkeley, Department of Environmental Science, Policy, and Management, 137 Mulford Hall, MC #3114, Berkeley, CA 94720.
[2] Corresponding author (e-mail: devalpin@nceas.ucsb.edu).

 doi: 10.1139/F05-116

space models using likelihoods that integrate over the state-space (but see Kitagawa (1987) and de Valpine and Hastings (2002) for methods that do this for low-dimensional models). The EV method can use nonlinear models and has sometimes been treated as a classical maximum likelihood method, but it does not use an integrated likelihood and thus can have asymptotically biased parameter estimates and unknown confidence intervals (CIs).

In this paper, we evaluate and test the Monte Carlo Kernel Likelihood method (MCKL; de Valpine 2003, 2004) for classical likelihood analysis with nonlinear state-space models. MCKL estimates likelihoods with a weighted kernel density estimate of a Bayesian posterior sample, so it is as flexible as Bayesian MCMC methods and can be used as a complementary tool. Our purpose is to fill a gap in available methodology, not to weigh the merits of Bayesian and classical approaches for various applications. In many management contexts, Bayesian analysis may remain the method of choice, particularly as a way to provide state (stock) estimates that incorporate uncertainty in the population dynamics parameters (Punt and Hilborn 1997). MCKL can provide state distributions given particular population parameters (analogous to the KF; Schnute 1994), but the classical interpretation of parameters makes it harder to incorporate parameter uncertainty in as simple a way as the Bayesian approach, and state estimation is not a focus of this paper.

We motivate the need for the MCKL method by comparing integrated with nonintegrated likelihoods in a classical framework. The EV method maximizes the joint likelihood of population states and parameters without integration over the state-space. It has long been recognized that EV does not use classical likelihoods (e.g., Mendelssohn 1988; Schnute 1994; Millar and Meyer 2000) and can give biased estimates (Polacheck et al. 1993; Schnute and Richards 1995). However, EV has sometimes still been treated as if asymptotic likelihood theory applies, such as by estimating confidence intervals using the second derivatives of the joint likelihood to estimate a Fisher information matrix (Ludwig and Walters 1981; Richards et al. 1997; Otter Research 2000) and using information criteria for model selection. To compare EV with classical maximum likelihood in a simple case, we use a linear autoregressive model. Even for this model, EV gives asymptotically biased parameter estimates and inaccurate CIs. Although Bayesian methods are not a focus of this paper, our results comparing integrated with nonintegrated likelihoods also support Bayesian analysis, since a posterior MCMC sample effectively integrates over the state-space.

Our comparison of EV and likelihood approaches uses both theory and simulations. State-space (integrated) likelihood methods have generally been evaluated by applying them to one or a few real data sets (e.g., Freeman and Kirkwood 1994; Kimura et al. 1996; Millar and Meyer 2000). However, this approach has the shortcoming that for any single data set, it is impossible to know which method happens to provide a good inference. An alternative approach is to compare methods using many simulated time series. This reveals the distribution of estimation behavior for each method, which allows comparisons of bias, variance, and accuracy of CIs (or Bayesian confidence regions). This approach has been used by Ludwig et al. (1988), Ludwig and Walters (1989), and Polacheck et al. (1993) in comparison of EV and non-state-space methods and by de Valpine and Hastings (2002) for a state-space likelihood method. Here we use replicated simulations to compare state-space likelihood and EV methods for linear and nonlinear models, including a realistic example based on the cape hake (*Merluccius capensis* × *M. paradoxus*) data analyzed by Hilborn and Mangel (1997), although our simulations are limited by using only the single recorded catch data.

This paper is organized into four sections. First we introduce the state-space framework and the difference between classical (integrated) likelihoods and EV (nonintegrated) likelihoods. Then we present simulation and theoretical results for the linear, autoregressive model. Next we introduce the MCKL method, and finally we compare it with EV using the Namibian hake example. To allow comparison against a known benchmark, we first fit a Schaefer model to the hake data using MCKL, simulate many data sets using the estimated parameters and actual catch records, and then fit the simulated series with both MCKL and EV.

State-space framework

This introduction to the general state-space model framework draws on Kitagawa (1987), Harvey (1989), and Schnute (1994) and uses notation similar to that of Schnute (1994).

Denote the true state of the population at time t as a scalar or vector $\boldsymbol{X}_t$, which might be population size or a stage or age vector. Let $\boldsymbol{X}^t = (\boldsymbol{X}_1, ..., \boldsymbol{X}_t)$, and let T be the final observation time, so the entire series is $\boldsymbol{X}^T$. Let the stochastic model for the true population dynamics be

$$\boldsymbol{X}_t = F(\boldsymbol{X}_{t-1}, \nu_{t-1}) \quad (1)$$

where ν_{t-1} is a random variable that provides the process error of the model. F might be a biomass dynamics model, a catch-at-age model, or some other fisheries model.

Denote the observation at time t as scalar or vector $\boldsymbol{Y}_t$, which might include total biomass, stage-structure or other information, and might aggregate or omit components of $\boldsymbol{X}_t$. Let $\boldsymbol{Y}^t = (\boldsymbol{Y}_1, ..., \boldsymbol{Y}_t)$. Denote the stochastic model for the observations as

$$\boldsymbol{Y}_t = G(\boldsymbol{X}_t, \varepsilon_t) \quad (2)$$

where ε_t is a random variable that provides observation error. The model could also include control variables such as catch (Schnute 1994), but for simplicity we initially omit these from the notation.

The parameters of F and G will be denoted as a vector, $\boldsymbol{\Theta}$, which we want to estimate. In this paper we reserve the term parameter for quantities like fecundity, survival, growth rate, and the means and variances of the distributions of process and observation errors. For convenience, we do not use "parameter" to refer to unknown states (or process errors), as it is sometimes used in Bayesian analysis and EV analysis, so that we can discuss these two categories unambiguously.

Likelihoods vs. EV likelihoods

For a state-space model, the states and observations are jointly distributed random variables, and the likelihood is the marginal

probability of the observations, integrated over all possible state trajectories. This is given by

$$(3)\qquad L(\boldsymbol{\Theta}|\boldsymbol{Y}^T) = P(\boldsymbol{Y}^T|\boldsymbol{\Theta}) = \int P(\boldsymbol{Y}^T, \boldsymbol{X}^T|\boldsymbol{\Theta})\,\mathrm{d}\boldsymbol{X}^T = \int P(\boldsymbol{Y}^T|\boldsymbol{X}^T, \boldsymbol{\Theta})P(\boldsymbol{X}^T|\boldsymbol{\Theta})\,\mathrm{d}\boldsymbol{X}^T$$

where $P(\boldsymbol{Y}^T, \boldsymbol{X}^T|\boldsymbol{\Theta})$ is the probability density of states and observations given parameters, calculated using the models F and G. Equation 3 corresponds to the classical definition of likelihood to which the usual asymptotic (large-sample) theory applies. The high-dimensional integral over $\boldsymbol{X}^T$ cannot be solved analytically except in the simplest cases, such as by the KF if F and G are linear with additive Gaussian noises. Equation 3 here plays the role of eq. 2.9 of Schnute (1994). Classical likelihood analysis calls for maximizing $L(\boldsymbol{\Theta}|\boldsymbol{Y}^T)$ over $\boldsymbol{\Theta}$ and estimating a confidence region, often approximated using asymptotic likelihood theory.

The EV likelihood handles the unknown states differently than the classical likelihood (eq. 3) in that it considers a single, specific trajectory of states. The EV likelihood is defined as the probability of a specific state trajectory, and the observations:

$$(4)\qquad L^{\mathrm{EV}}(\boldsymbol{\Theta}, \boldsymbol{X}^T|\boldsymbol{Y}^T) = P(\boldsymbol{Y}^T, \boldsymbol{X}^T|\boldsymbol{\Theta}) = P(\boldsymbol{Y}^T|\boldsymbol{X}^T, \boldsymbol{\Theta})P(\boldsymbol{X}^T|\boldsymbol{\Theta})$$

Note that it equals the integrand of the likelihood (eq. 3) for the particular states $\boldsymbol{X}^T$. Equation 4 here plays the role of eq. 2.6 in Schnute (1994). EV model-fitting calls for maximizing $L^{\mathrm{EV}}(\boldsymbol{\Theta}, \boldsymbol{X}^T|\boldsymbol{Y}^T)$ over both $\boldsymbol{\Theta}$ and $\boldsymbol{X}^T$. It has become a standard part of EV analysis to assume a known constraint on the process and observation error variances, usually in the form of a known ratio of the two variances.

In Bayesian analysis for $\boldsymbol{\Theta}$, one can consider the posterior for states and parameters together

$$(5)\qquad \pi(\boldsymbol{\Theta}, \boldsymbol{X}^T|\boldsymbol{Y}^T) = \frac{P(\boldsymbol{Y}^T|\boldsymbol{X}^T, \boldsymbol{\Theta})P(\boldsymbol{X}^T|\boldsymbol{\Theta})P(\boldsymbol{\Theta})}{P(\boldsymbol{Y}^T)}$$

as well as the posterior for parameters

$$(6)\qquad \pi(\boldsymbol{\Theta}|\boldsymbol{Y}^T) = \frac{P(\boldsymbol{Y}^T|\boldsymbol{\Theta})P(\boldsymbol{\Theta})}{P(\boldsymbol{Y}^T)}$$

which can be obtained by integrating (eq. 5) over the state dimensions:

$$(7)\qquad \pi(\boldsymbol{\Theta}|\boldsymbol{Y}^T) = \int \pi(\boldsymbol{\Theta}, \boldsymbol{X}^T|\boldsymbol{Y}^T)\,\mathrm{d}\boldsymbol{X}^T$$

In a Monte Carlo implementation, one obtains a sample of $(\boldsymbol{\Theta}, \boldsymbol{X}^T)$ points from $\pi(\boldsymbol{\Theta}, \boldsymbol{X}^T|\boldsymbol{Y}^T)$. Dropping the $\boldsymbol{X}^T$ dimensions represents a Monte Carlo approximation of the integral (eq. 7), so the remaining $\boldsymbol{\Theta}$ values are a sample from $\pi(\boldsymbol{\Theta}|\boldsymbol{Y}^T)$.

The EV method has a complicated history of justification and application. EV originated as an improvement over simply ignoring either process or observation errors (Ludwig and Walters 1981; Collie and Sissenwine 1983; Ludwig et al. 1988). In that context, the requirement of assuming a known ratio of the two error variances (or some constraint on them) was better than just omitting one type of error. Ludwig and Walters (1981) referred to the approach simply as "maximum likelihood" and suggested that "conventional large-sample" results applied to it, including the use of a Fisher information matrix "describing the distribution of the maximum likelihood estimates." (They also proposed a Bayesian interpretation for EV and pointed out that the EV variance estimate is badly biased, which is an indication that conventional large-sample theory for classical likelihoods — including asymptotically unbiased estimates — does not apply.) In a major synthesis of the state-space framework, Schnute (1994) compared the EV and the KF approaches. He placed EV in a Bayesian framework as a highest-posterior-density estimate of (eq. 5), but he also raised serious cautions and pointed toward other Bayesian estimation possibilities (which have now become feasible with MCMC sampling; Schnute and Kronlund 2002).

While it was recognized that EV likelihoods are not classical likelihoods (e.g., Mendelssohn 1988; Schnute 1994; Millar and Meyer 2000), the practice of applying asymptotic likelihood theory to EV estimates became somewhat common. For example, Schnute and Richards (1995) used $-2 \log L^{\mathrm{EV}}(\boldsymbol{\Theta}, \boldsymbol{X}^T|\boldsymbol{Y}^T)$ as an "inference function", a quantity that corresponds to approximations from classical likelihood theory. Similarly, Richards et al. (1997) called the Hessian of the EV estimate for the entire state–parameter vector an "observed Fisher information matrix" and stated that "highest posterior density confidence regions correspond exactly to regions obtained by likelihood ratio tests." This practice is also reflected in the manual for AD Model Builder (Otter Research 2000), a software package widely used for fisheries analysis. The manual gives an example presenting the inverse Hessian as a "covariance matrix of model parameters", including states (Otter Research 2000, pp. 1–22). All of these papers and the AD Model Builder package have made valuable contributions to fisheries science, which we do not discuss here; our scope is limited to evaluating the application of likelihood theory to EV likelihoods.

Given the practice of applying classical likelihood theory to EV likelihoods, it is worth emphasizing that some of the most useful properties of classical and Bayesian analysis apply only to eqs. 3 and 6, respectively, not to eqs. 4 or 5. For example, in the limit as $T \to \infty$, $\pi(\boldsymbol{\Theta}|\boldsymbol{Y}^T)$ is asymptotically Gaussian (with covariance that may be estimated using the Fisher information matrix from the Hessian matrix of second derivatives) and puts more and more weight on the true $\boldsymbol{\Theta}$, but the same does not hold for $\pi(\boldsymbol{\Theta}, \boldsymbol{X}^T|\boldsymbol{Y}^T)$ with respect to the true values of $\boldsymbol{\Theta}$ or $\boldsymbol{X}^T$. Our investigation of EV estimation represents a reevaluation of its justification starting from Schnute's (1994) cautionary note, but where he referred to "KF", we simply refer to "likelihood" methods to generalize the discussion to nonlinear and (or) non-Gaussian models.

Example 1: autoregressive (AR) model

How much do these theoretical issues matter in practice? In particular, how do EV estimates with likelihood-style approximate CIs compare to true maximum likelihood estimates and approximate CIs? We start with a first-order autoregressive

process with additive Gaussian errors, which allows us to combine theoretical results and simulations. Although this model is simpler than realistic fisheries models, it allows us to test the theoretical comparison of EV and likelihood methods in the simplest case. Since likelihood theory addresses parameter estimation rather than state (or latent variable) estimation, we focus on parameter estimation. State estimates are often of great interest in fisheries analysis but are beyond the scope of this paper.

For the constraint on error variances that is required for the EV method, we use the correct value. This is highly unrealistic, but it emphasizes that even under the correct variance constraint, EV methods behave differently than true likelihood methods.

The states of the model, X_t, will have dynamics:

$$X_t = \mu + \rho(X_{t-1} - \mu) + v_{t-1} \quad (8)$$

Observations will be made from states with error:

$$Y_t = X_t + \varepsilon_t \quad (9)$$

Here μ is the mean population size, ρ is the correlation between X_t and X_{t-1}, and v_t and ε_t are independently (from each other and through time) and identically normally distributed, with means of 0 and variances of σ_v^2 and σ_ε^2, respectively. We denote parameters by $\boldsymbol{\Theta} = (\mu, \rho, \sigma_v^2, \sigma_\varepsilon^2)$, states by $\boldsymbol{X}^T = (X_1, ..., X_T)$, and observations by $\boldsymbol{Y}^T = (Y_1, ..., Y_T)$.

Fitting methods

Denote the EV likelihood as

$$L_\lambda^{\mathrm{EV}}(\boldsymbol{\Theta}_C, \boldsymbol{X}^T|\boldsymbol{Y}^T) = P_\lambda(\boldsymbol{Y}^T|\boldsymbol{X}^T, \boldsymbol{\Theta}_C) P_\lambda(\boldsymbol{X}^T|\boldsymbol{\Theta}_C) \quad (10)$$

where the parameters to be estimated are constrained to $\boldsymbol{\Theta}_C = (\mu, \rho, \sigma_\varepsilon^2)$, and the variance ratio

$$\lambda = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\varepsilon^2} \quad (11)$$

is assumed known. Note that $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_C, \frac{\lambda}{1-\lambda}\sigma_\varepsilon^2)$. The assumption of known λ is awkward but has become a standard assumption for EV analysis.

For purposes of comparison to the EV method, we consider two versions of the classical likelihood. The first version assumes a known λ and will be called the constrained Kalman Filter (CKF). The second version assumes nothing about the error variances and is the KF. The CKF method shares the same parameter space as the EV method, so comparison of results from EV and CKF will reveal the difference between EV (non-integrated) likelihoods and true (integrated) likelihoods. Comparison between CKF and KF will reveal the difference between assuming λ known vs. unknown, respectively.

The CKF likelihood is given by

$$L_\lambda^{\mathrm{CKF}}(\boldsymbol{\Theta}_C|\boldsymbol{Y}^T) = \int P_\lambda(\boldsymbol{Y}^T|\boldsymbol{X}^T, \boldsymbol{\Theta}_C) P_\lambda(\boldsymbol{X}^T|\boldsymbol{\Theta}_C) \mathrm{d}\boldsymbol{X}^T \quad (12)$$

and the KF likelihood is given by

$$L^{\mathrm{KF}}(\boldsymbol{\Theta}|\boldsymbol{Y}^T) = \int P(\boldsymbol{Y}^T|\boldsymbol{X}^T, \boldsymbol{\Theta}) P(\boldsymbol{X}^T|\boldsymbol{\Theta}) \, \mathrm{d}\boldsymbol{X}^T \quad (13)$$

The steps for KF calculations are straightforward and are given in many sources, including Harvey (1989, 1993) and Schnute (1994). For the simple example here, equivalent calculations are even easier using the full multivariate normal distribution of $\boldsymbol{Y}^T$ (Appendix A).

Theoretical predictions for the AR model

Several analytical results on L^{EV} for the AR model are presented in Appendix A, with log likelihoods denoted $l = \log(L)$, and can be summarized by the following: (*i*) As the amount of data increases, l^{EV} parameter estimates do not converge to the true parameters (i.e., they are asymptotically biased), even if λ is correct. l^{EV} differs from l^{KF} and l^{CKF}, which do converge to the true parameters (if λ is correct for l^{CKF}). (*ii*) l^{KF} and l^{EV} converge as observation error variance becomes small, so the asymptotic bias of l^{EV} goes to zero as observation error variance goes to zero. (*iii*) EV systematically underestimates the error variances. (*iv*) For EV, approximate CIs based on likelihood-type approximations are not accurate and will become less accurate as the amount of data increases, even if λ is correct. As the amount of data increases, the Hessian will be more sharply peaked around asymptotically incorrect parameters, so EV will offer increasingly specious certainty about parameters increasingly certain to be incorrect.

Simulations for the AR model

We simulated 200 data sets under every combination of $\rho = (-0.25, 0.25, 0.75)$, $\lambda = (0.25, 0.5, 0.75)$, and $n = (40, 200)$. In every case, the variances were chosen to keep constant the stationary variance in observations:

$$\sigma_y^2 = \frac{\sigma_v^2}{1 - \rho^2} + \sigma_\varepsilon^2 \quad (14)$$

We used $\sigma_y^2 = 20$ for all simulations. The shorter series lengths ($n = 40$) are optimistic for well-studied fisheries, while the longer lengths allow comparison with asymptotic results. We used $\mu = 100$ for all simulations.

For every simulated data set, we estimated parameters using EV, CKF, and KF. For EV and CKF, we used the correct value of λ. We also calculated the likelihood-based, approximate p value of the true parameters for each method and reported the percentage of rejections of the true parameters compared with the expected number of rejections under asymptotic likelihood theory for Type I error rates of $\alpha = 0.05, 0.10, 0.25$, and 0.50.

Under asymptotic likelihood theory, –2 log ratio of the true parameter likelihood to the estimated parameter likelihood is distributed as χ_d^2, where d is the number of estimated parameters. For KF, we used this result to estimate p values for the true parameters. For EV and CKF, we used the related result that $(\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}})' \hat{\mathbf{I}}(\hat{\boldsymbol{\Theta}})(\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}) \sim \chi_d^2$, where $\hat{\mathbf{I}}(\hat{\boldsymbol{\Theta}})$ is the estimated inverse covariance matrix of $\hat{\boldsymbol{\Theta}}$, calculated as the Hessian of the log-likelihood surface (i.e., the matrix of second derivatives) (e.g., Richards et al. 1997; Severini 2000). For EV we used the estimated inverse covariance matrix of only the parameter dimensions (i.e., we inverted the state–parameter Hessian to obtain the estimated state–parameter covariance matrix, extracted the estimated covariance of the parameter dimensions, and inverted that covariance matrix to estimate p values for the true parameters). For EV in a linear regression problem,

Table 1. Mean, standard deviation (Std. Dev.), and root mean squared error (RMSE) of $\hat{\rho}$ in the autoregressive (AR) linear model example.

	$\rho = -0.25$			$\rho = 0.25$			$\rho = 0.75$		
	$\lambda = 0.25$	$\lambda = 0.50$	$\lambda = 0.75$	$\lambda = 0.25$	$\lambda = 0.50$	$\lambda = 0.75$	$\lambda = 0.25$	$\lambda = 0.50$	$\lambda = 0.75$
EV, $n = 40$									
Mean	−0.577	−0.423	−0.309	0.475	0.303	0.239	0.865	0.779	0.734
Std. Dev.	0.548	0.466	0.234	0.623	0.477	0.245	0.136	0.189	0.140
RMSE	0.638	0.497	0.241	0.663	0.480	0.245	0.179	0.192	0.141
CKF, $n = 40$									
Mean	−0.202	−0.243	−0.234	0.0527	0.158	0.176	0.536	0.589	0.641
Std. Dev.	0.363	0.275	0.186	0.401	0.262	0.183	0.332	0.203	0.155
RMSE	0.367	0.275	0.187	0.447	0.277	0.197	0.395	0.259	0.189
KF, $n = 40$									
Mean	−0.323	−0.358	−0.307	0.244	0.289	0.279	0.528	0.600	0.637
Std. Dev.	0.261	0.286	0.257	0.319	0.283	0.262	0.344	0.199	0.177
RMSE	0.271	0.306	0.263	0.319	0.286	0.263	0.409	0.249	0.210
EV, $n = 200$									
$\lim_{n\to\infty} \hat{\rho}$	−0.830	−0.595	−0.352	0.830	0.595	0.352	0.924	0.882	0.828
Mean	−0.820	−0.548	−0.345	0.815	0.553	0.320	0.921	0.865	0.808
Std. Dev.	0.122	0.201	0.103	0.124	0.157	0.116	0.0189	0.0329	0.0383
RMSE	0.583	0.359	0.141	0.579	0.341	0.135	0.172	0.120	0.0692
CKF, $n = 200$									
Mean	−0.238	−0.245	−0.251	0.201	0.233	0.231	0.720	0.719	0.728
Std. Dev.	0.221	0.133	0.0836	0.216	0.116	0.0908	0.0764	0.0635	0.0495
RMSE	0.222	0.133	0.0836	0.221	0.117	0.0928	0.0821	0.0705	0.0541
KF, $n = 200$									
Mean	−0.305	−0.362	−0.328	0.319	0.359	0.302	0.698	0.712	0.724
Std. Dev.	0.266	0.260	0.190	0.275	0.250	0.185	0.137	0.101	0.0657
RMSE	0.271	0.283	0.206	0.283	0.273	0.192	0.146	0.108	0.0705

Note: Column headers indicate parameters used for simulation. Row sections give results for different fitting methods (EV, errors-in-variables; KF, Kalman Filter; CKF, constrained Kalman Filter) and different series lengths. Calculated asymptotic value is shown as $\lim_{n\to\infty} \hat{\rho}$.

Schnute et al. (1990) showed that one-half the usual likelihood ratio calculation (or related Hessian calculation) gave the usual χ^2_d result, and Schnute (1994) raised the question of whether a similar result would apply to EV for time series, so we also calculated p values based on this quantity ("EV adjusted").

Results for the AR simulations

Simulations confirm the theoretical predictions about EV, CKF, and KF (Tables 1, 2; Figs. 1, 2). In nearly every case, CKF has lower root mean squared error ($\text{RMSE} = \sqrt{\text{bias}^2 + \text{variance}}$) than EV. For EV, more data often lead to increased bias, as the estimates converge to an asymptotically biased estimate, while for CKF more data lead to decreased bias. The arrows in Figs. 1 and 2 and the line "$\lim_{n\to\infty}(\cdot)$" in Tables 1 and 2 give the theoretical asymptotic parameter estimates for EV, obtained computationally as described in Appendix A. The means of the $n = 200$ results are close to these values. The only cases where EV has smaller RMSE than CKF are the ρ estimates with $n = 40$ and $\lambda = 0.75$, but even in these cases, EV estimates converge away from the true values as n increases, while CKF estimates converge toward the true values. For both $n = 40$ and $n = 200$, EV systematically underestimates the error variances, consistent with Ludwig and Walters (1981), Schnute and Kronlund (2002), and results therein.

Likelihood-based p values for EV are not even approximately accurate (Table 3). For example, using the χ^2_d cutoff that asymptotically rejects the true parameters with frequency 5%, EV rejected the true parameters in 99% of simulations, while CKF rejected them in 11% of simulations. The EV-adjusted results calculated from twice the usual test statistic are also very inaccurate. For CKF, p value percentiles differ from the asymptotic expectation but may be considered to be a useful approximation.

In terms of both parameter estimates and p values, KF performs substantially worse than CKF owing to its difficulty estimating error variances. Nevertheless, its confidence regions (as reflected by p value results) offer a useful approximation of uncertainty. For example, although KF does not estimate error variances well, it does include the true parameters in a putative 95% confidence region roughly 99% of the time (across the range of simulation parameters). In summary, classical likelihoods outperform EV likelihoods (CKF vs. EV), while it remains true that separating process and observation error is an inherently difficult problem (KF vs. CKF).

Having compared classical and EV likelihoods with theory and simulations, we reevaluate Schnute's (1994) proposed justification of the EV method from a Bayesian point of view, which may be summarized as follows: (*i*) Bayesian analysis treats both states and parameters as random variables. (In a fre-

Table 2. Mean, standard deviation (Std. Dev.), and root mean squared error (RMSE) for $\hat{\sigma}_\varepsilon$ in the autoregressive (AR) linear model example.

	$\rho = -0.25$			$\rho = 0.25$			$\rho = 0.75$		
	$\lambda = 0.25$ $\sigma_\varepsilon = 3.84$	$\lambda = 0.50$ $\sigma_\varepsilon = 3.11$	$\lambda = 0.75$ $\sigma_\varepsilon = 2.18$	$\lambda = 0.25$ $\sigma_\varepsilon = 3.84$	$\lambda = 0.50$ $\sigma_\varepsilon = 3.11$	$\lambda = 0.75$ $\sigma_\varepsilon = 2.18$	$\lambda = 0.25$ $\sigma_\varepsilon = 3.37$	$\lambda = 0.50$ $\sigma_\varepsilon = 2.47$	$\lambda = 0.75$ $\sigma_\varepsilon = 1.60$
EV, $n = 40$									
Mean	2.54	2.08	1.48	2.56	2.08	1.49	2.28	1.68	1.08
Std. Dev.	0.303	0.262	0.163	0.301	0.235	0.165	0.266	0.217	0.129
RMSE	1.33	1.07	0.723	1.32	1.05	0.715	1.12	0.819	0.534
CKF, $n = 40$									
Mean	3.69	2.98	2.10	3.70	2.98	2.11	3.32	2.42	1.53
Std. Dev.	0.429	0.372	0.230	0.429	0.333	0.234	0.390	0.314	0.184
RMSE	0.454	0.395	0.245	0.453	0.356	0.245	0.392	0.318	0.194
KF, $n = 40$									
Mean	2.48	2.82	3.41	2.17	2.64	3.12	2.39	2.59	2.75
Std. Dev.	1.89	1.78	1.38	1.95	1.83	1.54	1.21	0.987	0.624
RMSE	2.33	1.80	1.84	2.57	1.89	1.80	1.56	0.994	1.31
EV, $n = 200$ + $\lim_{n\to\infty} \hat{\sigma}_\varepsilon$	2.62	2.17	1.54	2.62	2.17	1.54	2.32	1.72	1.12
Mean	2.62	2.15	1.53	2.62	2.16	1.53	2.31	1.71	1.12
Std. Dev.	0.142	0.116	0.0816	0.126	0.113	0.0744	0.112	0.0875	0.0555
RMSE	1.23	0.963	0.660	1.23	0.958	0.654	1.07	0.761	0.481
CKF, $n = 200$									
Mean	3.82	3.09	2.16	3.81	3.09	2.17	3.34	2.46	1.59
Std. Dev.	0.201	0.164	0.115	0.182	0.160	0.105	0.165	0.127	0.0791
RMSE	0.202	0.165	0.116	0.184	0.161	0.106	0.166	0.127	0.0793
KF, $n = 200$									
Mean	2.65	2.89	3.51	2.40	2.88	3.55	2.09	2.50	2.79
Std. Dev.	1.79	1.48	1.09	1.80	1.45	1.06	0.664	0.464	0.351
RMSE	2.14	1.50	1.72	2.30	1.47	1.73	1.44	0.465	1.24

Note: Column headers indicate parameters used for simulation. Row sections give results for different fitting methods (EV, errors-in-variables; KF, Kalman Filter; CKF, constrained Kalman Filter) and different series lengths. Calculated asymptotic value is shown as $\lim_{n\to\infty} \hat{\rho}$.

quentist view, parameters are not random variables, but states are.) (*ii*) Bayesian analysis sometimes estimates parameters by finding the maximum posterior probability. (*iii*) If states and parameters are both random variables, they can be estimated the same way (i.e., by maximizing their joint posterior). For an uninformative prior, this is the same as maximizing the EV likelihood, which is the EV method. The reasoning is summarized by Schnute's (1994, p. 1680) statement that "the Bayes framework easily allows the random states $\mathbf{X}_t$ to be considered parameters, so that the posterior (2.11) [here eq. 5] is meaningful without integration."

As shown for even a simple linear model, a weakness in this reasoning is that likelihood theory does not apply to joint estimates of states and parameters. The breakdown in the proposed justification is that the usual likelihood properties of Bayesian and classical inference apply only to the parameters, not the states, so that using the joint state–parameter Bayesian posterior as a justification for the EV likelihood subtly drops the applicability of standard likelihood theory.

Monte Carlo state-space likelihoods

Next we present a method for classical maximum likelihood estimation with nonlinear state-space models, based on Monte Carlo approximations of the likelihood integral (de Valpine 2003, 2004). Several algorithms have been developed for this problem in non-fisheries contexts, including basic Monte Carlo integration (Durbin and Koopman 1997, 2000), Monte Carlo expectation maximization (Wei and Tanner 1990; Chan and Ledolter 1995; Booth and Hobert 1999), maximization based on local Monte Carlo likelihood ratio estimation (Geyer and Thompson 1992; McCulloch 1997), and "particle filter" methods (Gordon et al. 1993; Pitt and Shephard 1999; Hürzeler and Kunsch 2001). For the case of stage-structured population models, de Valpine (2004) found that a new method (MCKL) was more efficient than these others. In addition, MCKL has the advantage that it estimates classical likelihoods from a Monte Carlo Bayesian posterior, so that the same computational work can be used for both Bayesian and classical analyses. We present the basic rationale behind the MCKL method; theoretical convergence and accuracy are considered by de Valpine (2004).

Consider the general problem of estimating the likelihood, $L(\boldsymbol{\Theta}) \equiv P(\boldsymbol{Y}^T|\boldsymbol{\Theta})$, from a Monte Carlo sample of the posterior $\pi(\boldsymbol{\Theta}|\boldsymbol{Y}^T)$, denoted by $[\boldsymbol{\Theta}^{(j)}]_{j=1}^m$ for a sample with m points. The difficulty is that $\pi(\boldsymbol{\Theta}|\boldsymbol{Y}^T) \propto L(\boldsymbol{\Theta})P(\boldsymbol{\Theta})$, where $P(\boldsymbol{\Theta})$ is the prior, so we need to divide by $P(\boldsymbol{\Theta})$ after Monte Carlo sampling. (We cannot assume that $P(\boldsymbol{\Theta})$ can be chosen as flat everywhere because then $\pi(\boldsymbol{\Theta}|\boldsymbol{Y}^T)$ may not have finite integral, leading MCMC algorithms to fail.)

To solve this problem, MCKL uses two tools of Monte Carlo

Fig. 1. Estimates of ρ for the autoregressive (AR) model. Box plots of 200 point estimates of ρ from simulated time series of length 40 for each of nine simulation parameter values ($3\lambda \times 3\rho$ values: (*a*) $\rho = -0.25$, (*b*) $\rho = 0.25$, (*c*) $\rho = 0.75$). EV, errors-in-variables; CKF, constrained Kalman Filter; KF, Kalman Filter. Boxes show 25th, 50th, and 75th percentiles; dashed whiskers extend to the farthest value within 1.5 times the interquartile range; dots show all remaining values; *, mean; dotted horizontal line, true value. Arrows for EV shows asymptotic value (i.e., the bias given infinite data). (Asymptotic value for CKF and KF is the true value.)

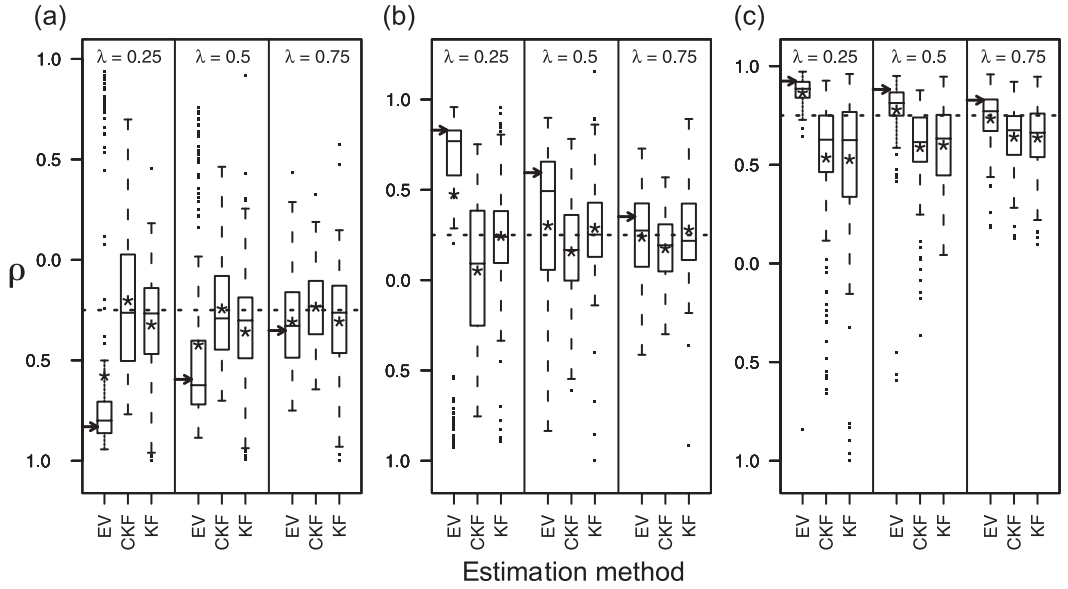
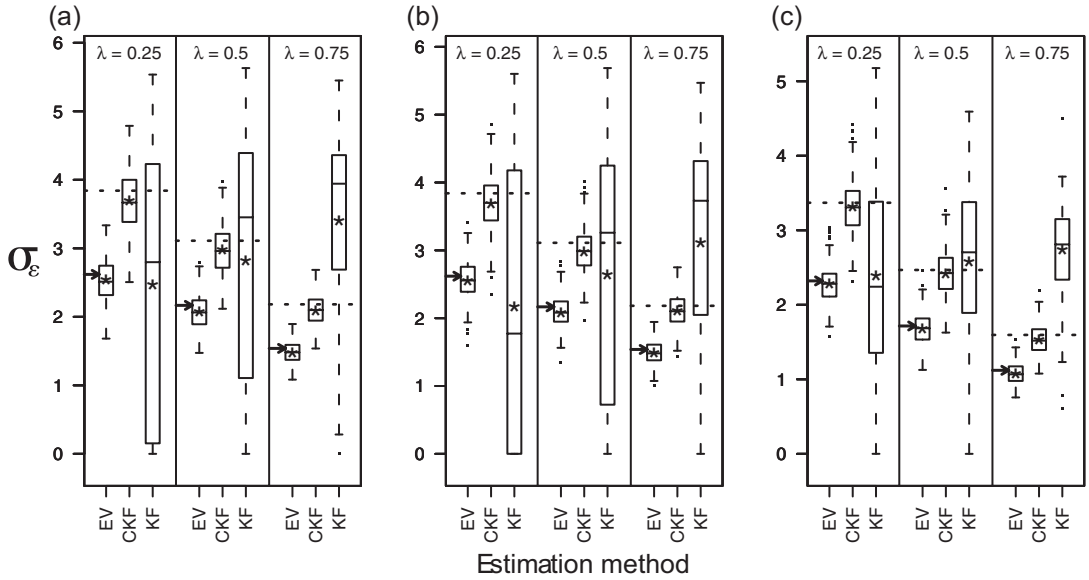


Fig. 2. Estimates of σ_ε for the autoregressive (AR) model. Box plots of 200 point estimates of ρ from simulated time series of length 40 for each of nine simulation parameter values ($3\lambda \times 3\rho$ values: (*a*) $\rho = -0.25$, (*b*) $\rho = 0.25$, (*c*) $\rho = 0.75$). EV, errors-in-variables; CKF, constrained Kalman Filter; KF, Kalman Filter. Boxes show 25th, 50th, and 75th percentiles; dashed whiskers extend to the farthest value within 1.5 times the interquartile range; dots show all remaining values; *, mean; dotted horizontal line, true value. Arrows for EV shows asymptotic value (i.e., the bias given infinite data). (Asymptotic value for CKF and KF is the true value.)



methodology: kernel density estimation and importance sampling. If the goal was to estimate values of $\pi(\boldsymbol{\Theta}|\boldsymbol{Y}^T)$, this would be a standard problem of density estimation from a sample, and we could use kernel density estimation alone. We define $K_{\mathbf{h}}(\boldsymbol{\Theta})$ to be a multivariate normal density function (the kernel) with mean zero and a diagonal covariance matrix with values $(h_1^2, h_2^2, \ldots, h_d^2)$. (The choice of diagonal covariance matrix is for simplicity of notation only.) A kernel density estimate of $\pi(\boldsymbol{\Theta}|\boldsymbol{Y}^T)$ is

$$\hat{\pi}(\boldsymbol{\Theta}|\boldsymbol{Y}^T) \equiv \frac{1}{m}\sum_{j=1}^{m} K_{\mathbf{h}}[\boldsymbol{\Theta} - \boldsymbol{\Theta}^{(j)}] \quad (15)$$

Intuitively, $\hat{\pi}(\boldsymbol{\Theta}|\boldsymbol{Y}^T)$ is a weighted average of the number of points in the Monte Carlo sample that are near $\boldsymbol{\Theta}$, with weights determined by the kernel. Choice of $\mathbf{h}$ is determined by a balance

Table 3. Confidence interval (CI) performance of fitting methods for the autoregressive (AR) linear model example.

	$\rho = -0.25$			$\rho = 0.25$			$\rho = 0.75$		
	$\lambda = 0.25$	$\lambda = 0.50$	$\lambda = 0.75$	$\lambda = 0.25$	$\lambda = 0.50$	$\lambda = 0.75$	$\lambda = 0.25$	$\lambda = 0.50$	$\lambda = 0.75$
Expected percentile	$\sigma_\varepsilon = 3.84$	$\sigma_\varepsilon = 3.11$	$\sigma_\varepsilon = 2.18$	$\sigma_\varepsilon = 3.84$	$\sigma_\varepsilon = 3.11$	$\sigma_\varepsilon = 2.18$	$\sigma_\varepsilon = 3.37$	$\sigma_\varepsilon = 2.47$	$\sigma_\varepsilon = 1.60$
EV, $n = 40$									
0.05	0.990	0.970	0.985	0.990	0.980	0.975	0.955	0.950	0.975
0.10	0.990	0.975	0.995	0.995	0.985	0.985	0.975	0.960	0.985
0.25	0.995	0.980	0.995	0.995	0.995	0.995	0.985	0.975	0.985
0.50	1	1	1	0.995	0.995	1	0.990	0.990	0.990
EV adjusted, $n = 40$									
0.05	0.950	0.920	0.885	0.940	0.915	0.835	0.840	0.825	0.820
0.10	0.975	0.930	0.920	0.970	0.955	0.885	0.890	0.865	0.880
0.25	0.990	0.970	0.955	0.990	0.975	0.955	0.930	0.940	0.970
0.50	0.990	0.975	0.995	0.995	0.985	0.990	0.985	0.970	0.985
CKF, $n = 40$									
0.05	0.110	0.090	0.065	0.135	0.090	0.055	0.100	0.115	0.145
0.10	0.150	0.135	0.105	0.185	0.125	0.100	0.150	0.165	0.205
0.25	0.270	0.245	0.195	0.280	0.225	0.170	0.235	0.280	0.340
0.50	0.410	0.455	0.360	0.460	0.425	0.370	0.395	0.480	0.500
KF, $n = 40$									
0.05	0.015	0.015	0.005	0.025	0.035	0.010	0.020	0.035	0.025
0.10	0.045	0.040	0.015	0.025	0.050	0.020	0.045	0.070	0.055
0.25	0.095	0.120	0.105	0.075	0.110	0.070	0.140	0.175	0.165
0.50	0.220	0.305	0.260	0.205	0.245	0.220	0.340	0.415	0.430
EV, $n = 200$									
0.05	1	1	1	1	1	1	1	1	1
0.10	1	1	1	1	1	1	1	1	1
0.25	1	1	1	1	1	1	1	1	1
0.50	1	1	1	1	1	1	1	1	1
EV adjusted, $n = 200$									
0.05	1	1	1	1	1	1	1	1	1
0.10	1	1	1	1	1	1	1	1	1
0.25	1	1	1	1	1	1	1	1	1
0.50	1	1	1	1	1	1	1	1	1
CKF, $n = 200$									
0.05	0.060	0.045	0.050	0.065	0.015	0.035	0.015	0.045	0.015
0.10	0.110	0.085	0.070	0.075	0.035	0.050	0.035	0.120	0.075
0.25	0.255	0.205	0.165	0.175	0.135	0.125	0.120	0.185	0.190
0.50	0.410	0.395	0.355	0.355	0.320	0.330	0.320	0.355	0.365
KF, $n = 200$									
0.05	0.010	0.025	0.025	0.005	0.020	0.015	0.005	0.030	0.015
0.10	0.040	0.075	0.040	0.015	0.030	0.035	0.035	0.060	0.050
0.25	0.105	0.155	0.125	0.085	0.070	0.090	0.105	0.205	0.125
0.50	0.285	0.355	0.270	0.235	0.265	0.265	0.315	0.380	0.380

Note: Column headers indicate parameters used for simulation. Row sections give results for different fitting methods (EV, errors-in-variables; KF, Kalman Filter; CKF, constrained Kalman Filter) and different series lengths. Rows are labeled by the p value cutoff that was used (0.05, 0.10, 0.25, 0.50). Table values give percent of simulations in which the true parameters were rejected using the given cutoff. For example, EV rejected the true parameters over 95% of the time when a nominal rejection percentage of only 5% was chosen.

between bias and variance. Small $\mathbf{h}$ gives low bias but high variance, and vice-versa.

A more general view of eq. 15 is that it is a Monte Carlo approximation to the convolution (or smoothing) integral of $K_{\mathbf{h}}$ and π, which is defined by

$$\tilde{\pi}(\boldsymbol{\Theta}|\boldsymbol{Y}^T) \equiv \int K_{\mathbf{h}}(\boldsymbol{\Theta} - \boldsymbol{\phi})\pi(\boldsymbol{\phi}|\boldsymbol{Y}^T)\, \mathrm{d}\boldsymbol{\phi} \quad (16)$$

This integral can be interpreted as the average of $K_{\mathbf{h}}(\boldsymbol{\Theta} - \boldsymbol{\phi})$ over the distribution $\pi(\boldsymbol{\phi}|\boldsymbol{Y}^T)$. A natural way to estimate an average over a distribution is to use a sample of points from the distribution and calculate the average directly, which explains why eq. 15 is an estimate of eq. 16.

The next step is to adjust the kernel density estimate to recover $L(\boldsymbol{\Theta})$ instead of $\pi(\boldsymbol{\Theta}|\boldsymbol{Y}^T)$ (i.e., to divide out the prior after the Monte Carlo sample has been obtained). There must still be some smoothing, since we are estimating a density from a Monte Carlo sample, so a goal is to estimate the convolution

Table 4. Mean, standard deviation (Std. Dev.), and root mean squared error (RMSE) of $\log(\hat{r})$ for the hake (*Merluccius capensis* × *M. paradoxus*) example using errors-in-variables (EV; λ known) and Monte Carlo Kernel Likelihood (MCKL; λ known vs. unknown).

	Parameter used for simulation		
	$\boldsymbol{\Theta}_{.01}$[a]	$\boldsymbol{\Theta}_{.05}$[b]	$\boldsymbol{\Theta}_{.10}$[c]
EV (λ known)			
Mean	−1.78	−1.51	−1.41
Std. Dev.	0.723	0.612	0.336
RMSE	0.918	0.726	0.525
MCKL (λ known)			
Mean	−1.17	−1.12	−1.12
Std. Dev.	0.523	0.461	0.302
RMSE	0.524	0.461	0.320
MCKL (λ unknown)			
Mean	−1.25	−1.18	−1.13
Std. Dev.	0.583	0.490	0.331
RMSE	0.584	0.495	0.352

[a] $\log(r) = -1.21$.
[b] $\log(r) = -1.18$.
[c] $\log(r) = -1.01$.

Table 5. Mean, standard deviation (Std. Dev.), and root mean squared error (RMSE) for $\log(\hat{q})$ for the hake (*Merluccius capensis* × *M. paradoxus*) example using errors-in-variables (EV; λ known) and Monte Carlo Kernel Likelihood (MCKL; λ known vs. unknown).

	Parameter used for simulation		
	$\boldsymbol{\Theta}_{.01}$[a]	$\boldsymbol{\Theta}_{.05}$[b]	$\boldsymbol{\Theta}_{.10}$[c]
EV (λ known)			
Mean	−7.49	−7.24	−7.22
Std. Dev.	0.379	0.555	0.399
RMSE	0.601	0.844	0.684
MCKL (λ known)			
Mean	−8.07	−7.98	−7.91
Std. Dev.	0.351	0.346	0.260
RMSE	0.367	0.361	0.291
MCKL (λ unknown)			
Mean	−8.04	−7.98	−7.90
Std. Dev.	0.359	0.361	0.291
RMSE	0.367	0.377	0.316

[a] $\log(q) = -7.96$.
[b] $\log(q) = -7.87$.
[c] $\log(q) = -7.78$.

of the likelihood (rather than the posterior), defined by

$$\tilde{L}(\boldsymbol{\Theta}) \equiv \int K_{\mathbf{h}}(\boldsymbol{\Theta} - \boldsymbol{\phi}) L(\boldsymbol{\phi})\, \mathrm{d}\boldsymbol{\phi} \quad (17)$$

This is equivalent to

$$\tilde{L}(\boldsymbol{\Theta}) = \int K_{\mathbf{h}}(\boldsymbol{\Theta} - \boldsymbol{\phi}) \frac{L(\boldsymbol{\phi})}{\pi(\boldsymbol{\phi}|\boldsymbol{Y}^T)} \pi(\boldsymbol{\phi}|\boldsymbol{Y}^T)\, \mathrm{d}\boldsymbol{\phi} \quad (18)$$

which is the average of $K_{\mathbf{h}}(\boldsymbol{\Theta} - \boldsymbol{\phi}) \frac{L(\boldsymbol{\phi})}{\pi(\boldsymbol{\phi}|\boldsymbol{Y}^T)}$ over the distribution $\pi(\boldsymbol{\phi}|\boldsymbol{Y}^T)$. Substituting $\frac{L(\boldsymbol{\phi})}{\pi(\boldsymbol{\phi}|\boldsymbol{Y}^T)} = \frac{P(\boldsymbol{Y}^T)}{P(\boldsymbol{\phi})}$, the natural Monte Carlo estimate of eq. 18 is

$$\hat{L}(\boldsymbol{\Theta}) \equiv \frac{P(\boldsymbol{Y}^T)}{m} \sum_{j=1}^{m} \frac{K_{\mathbf{h}}[\boldsymbol{\Theta} - \boldsymbol{\Theta}^{(j)}]}{P[\boldsymbol{\Theta}^{(j)}]} \quad (19)$$

The step of multiplying and dividing by $\pi(\boldsymbol{\phi}|\boldsymbol{Y}^T)$ in eq. 18, so that each sample point $\boldsymbol{\Theta}^{(j)} \sim \pi(\boldsymbol{\Theta}|\boldsymbol{Y}^T)$ is weighted by $1/P[\boldsymbol{\Theta}^{(j)}]$ in eq. 19, is known as importance sampling. Finally, the unknown constant $P(\boldsymbol{Y}^T)$ can be ignored for maximization of eq. 19 over $\boldsymbol{\Theta}$, giving the MCKL estimate:

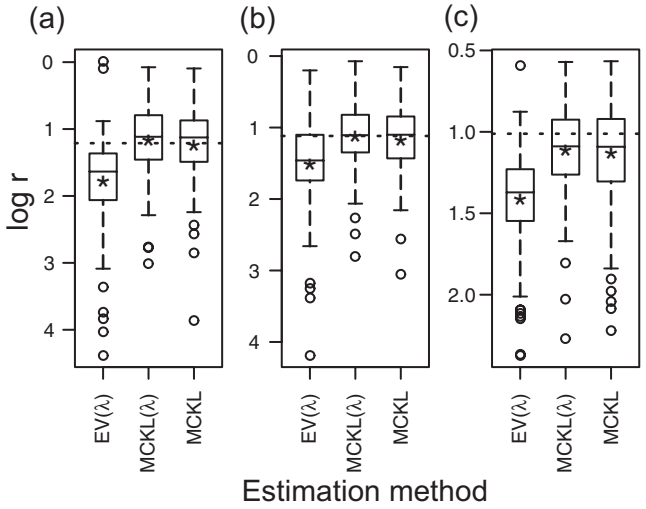
$$\hat{L}_{\mathrm{U}}(\boldsymbol{\Theta}) \equiv \frac{1}{m} \sum_{j=1}^{m} \frac{K_{\mathbf{h}}[\boldsymbol{\Theta} - \boldsymbol{\Theta}^{(j)}]}{P[\boldsymbol{\Theta}^{(j)}]} \quad (20)$$

where subscript U indicates unnormalized. In summary, eq. 20 is a weighted, importance-sampled, kernel density estimate of the likelihood up to an unknown constant, where the Bayesian posterior serves as the importance sampling density. For some analyses (such as model selection using information criteria), a normalized likelihood is necessary. $P(\boldsymbol{Y}^T)$ may be estimated by methods such as importance sampling or particle filtering that would be too slow for maximization but are feasible for a single calculation, such as at the maximum of $\hat{L}_{\mathrm{U}}(\boldsymbol{\Theta})$.

Since eq. 20 is an approximation of a convolved likelihood, it raises several issues of convergence, accuracy, and practical implementation. In terms of theoretical convergence, standard results for kernel density estimation give that $\lim_{m\to\infty} \hat{L}_{\mathrm{U}}(\boldsymbol{\Theta}) = \tilde{L}(\boldsymbol{\Theta})/P(\boldsymbol{Y}^T)$, and $\lim_{h\to 0} \tilde{L}(\boldsymbol{\Theta}) = L(\boldsymbol{\Theta})$ (Silverman 1986; Scott 1992). For these results to hold here, the following conditions are required (de Valpine 2004). We define $\psi_j(\boldsymbol{\Theta}) \equiv K_{\mathbf{h}}[\boldsymbol{\Theta} - \boldsymbol{\Theta}^{(j)}]/P[\boldsymbol{\Theta}^{(j)}]$, and we define $\nabla\boldsymbol{\psi}_j(\boldsymbol{\Theta})$ and $\nabla^2\boldsymbol{\psi}_j(\boldsymbol{\Theta})$ to be the vector of first derivatives and matrix of second derivatives with respect to $\boldsymbol{\Theta}$, respectively. Then it is required that $E[\nabla^2\boldsymbol{\psi}_j(\boldsymbol{\Theta}_{\mathbf{h}})]$ be invertible and $E\{[\nabla\boldsymbol{\psi}_j(\boldsymbol{\Theta}_{\mathbf{h}})][\nabla\boldsymbol{\psi}_j(\boldsymbol{\Theta}_{\mathbf{h}})]^T\}$ exist, where $\boldsymbol{\Theta}_{\mathbf{h}}$ is the maximum of $\tilde{L}$, and the expectations are with respect to the sampling distribution $\pi[\boldsymbol{\Theta}^{(j)}|\boldsymbol{Y}^T]$.

Implementation of MCKL requires two steps: (*i*) obtain an MCMC sample of the Bayesian posterior, and (*ii*) maximize eq. 20. MCMC sampling methodology is advancing quickly. Equation 20 is straightforward to program, but one must choose sample size (m) and multivariate smoothing bandwidth ($\mathbf{h}$). Optimal choice of bandwidth in standard kernel density estimation is a highly developed topic, but the current problem is different because of the inverse-prior weighting and the possibility of simulating additional points for the smoothing sample. de Valpine (2004) gives an approach to choosing $\mathbf{h}$ based on the approximately Gaussian shape of the log likelihood surface near its maximum. One can even choose $\mathbf{h}$ manually, based on inspection of estimated likelihood surfaces, until better-developed, automated approaches become available. In addition, it is possible to improve accuracy after obtaining an initial maximum likelihood estimate by using a new prior, with heavier weight near the maximum, and obtaining a second MCMC sample that

Fig. 3. Estimates of log(r) for the hake (*Merluccius capensis* × *M. paradoxus*) model. Box plots of 100 point estimates of log(r) for each of three simulation parameters: (*a*) $\sigma_\varepsilon = 0.01$, (*b*) $\sigma_\varepsilon = 0.05$, (*c*) $\sigma_\varepsilon = 0.10$ (see text). EV, errors-in-variables; MCKL, Monte Carlo Kernel Likelihood. (λ) indicates that λ is assumed correctly known.



gives greater MCKL accuracy near the maximum ("zooming in"; de Valpine 2004).

Note that the density estimation must be conducted in the same coordinates as the sampling; transformation to other coordinates before density estimation would introduce a correction based on the determinant of the Jacobian of the transformation (as in standard probability theory), which we do not give here. It is safe, however, to use a linear transformation, such as principle components, since the Jacobian is simply a constant.

In fisheries applications, the distribution of states, particularly of the last state, given parameters may be of great interest. This distribution may be obtained in one of two ways. First, one may include the final state dimension in the kernel-smoothing integral after estimating the maximum likelihood parameters $\hat{\boldsymbol{\Theta}}$. For example, one would first estimate $\hat{\boldsymbol{\Theta}}$ and then use eq. 20 with the X_T dimension included to estimate $P(X_T|\hat{\boldsymbol{\Theta}})$. Second, one could run the MCMC algorithm for states only with parameters fixed at $\hat{\boldsymbol{\Theta}}$ to obtain a sample from $P(X_T|\hat{\boldsymbol{\Theta}})$ (obtaining this sample will be much faster than obtaining the joint state–parameter sample). As noted in the Introduction, management-oriented incorporation of parameter uncertainty into estimates of the state distribution still provides strong motivation for a Bayesian approach. This paper fills in the gap of classical likelihood methods for state-space models but does not conflict with Bayesian approaches. Having indicated how one would estimate state distributions using MCKL, we defer further comparison of state estimates to further research and focus here on comparison of parameter estimates using different methods.

Example 2: hake

To illustrate MCKL and compare it with EV in a real but relatively simple fisheries problem, we use the hake data (annual catch and abundance index from 1965 to 1987) and Schaefer

Fig. 4. Estimates of log(q) for the hake (*Merluccius capensis* × *M. paradoxus*) model. Box plots of 100 point estimates of log(q) for each of three simulation parameters: (*a*) $\sigma_\varepsilon = 0.01$, (*b*) $\sigma_\varepsilon = 0.05$, (*c*) $\sigma_\varepsilon = 0.10$ (see text). EV, errors-in-variables; MCKL, Monte Carlo Kernel Likelihood. (λ) indicates that λ is assumed correctly known.

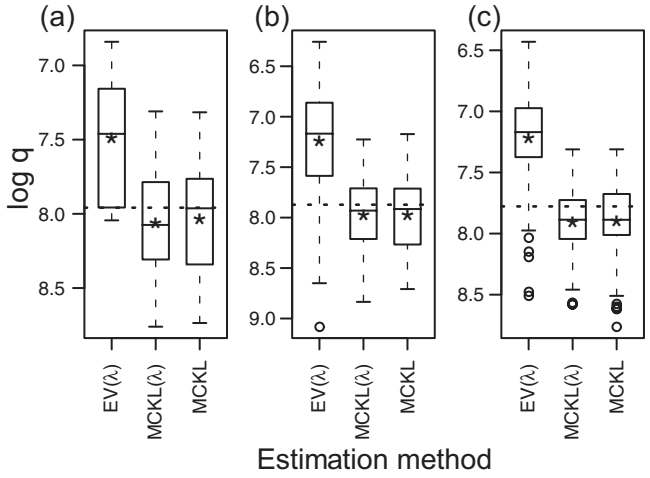
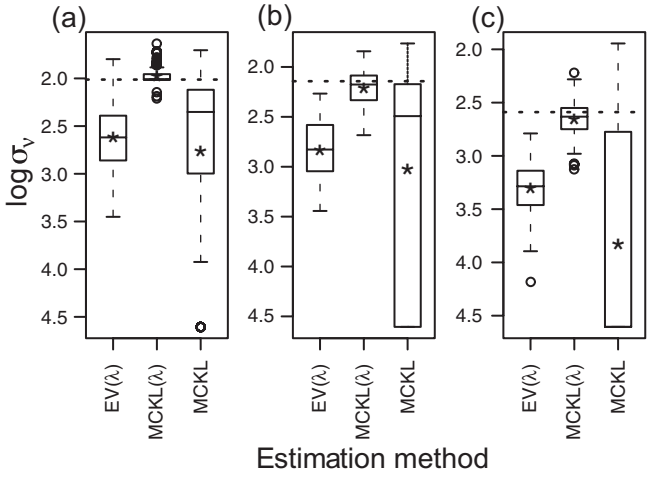


Fig. 5. Estimates of log(σ_ν) for the hake (*Merluccius capensis* × *M. paradoxus*) model. Box plots of 100 point estimates of log(q) for each of three simulation parameters: (*a*) $\sigma_\varepsilon = 0.01$, (*b*) $\sigma_\varepsilon = 0.05$, (*c*) $\sigma_\varepsilon = 0.10$ (see text). EV, errors-in-variables; MCKL, Monte Carlo Kernel Likelihood. (λ) indicates that λ is assumed correctly known.



model investigated by Hilborn and Mangel (1997).

The model is

$$B(t+1) = \left[B(t) + rB(t)(1 - \frac{B(t)}{K}) - C(t)\right]\exp(\nu_t) \quad (21)$$

$$I(t) = qB(t)\exp(\varepsilon_t) \quad (22)$$

where $B(t)$, $C(t)$ and $I(t)$ are biomass, catch, and abundance index in year t, respectively; r is intrinsic growth rate; K is carrying capacity; q is fraction of biomass estimated by the survey; and ν_t and ε_t are process and observation errors at time t, respectively, with means 0 and variances σ_ν^2 and σ_ε^2, respectively. The parameters to be estimated are $\boldsymbol{\Theta} \equiv (r, K, q, \sigma_\nu, \sigma_\varepsilon)$.

Table 6. Mean, standard deviation (Std. Dev.), and root mean squared error (RMSE) for $\log(\hat{\sigma}_v)$ for the hake (*Merluccius capensis* × *M. paradoxus*) example using errors-in-variables (EV; λ known) and Monte Carlo Kernel Likelihood (MCKL; λ known vs. unknown).

	Parameter used for simulation		
	$\boldsymbol{\Theta}_{.01}$[a]	$\boldsymbol{\Theta}_{.05}$[b]	$\boldsymbol{\Theta}_{.10}$[c]
EV (λ known)			
Mean	−2.61	−2.83	−3.30
Std. Dev.	0.359	0.307	0.261
RMSE	0.698	0.757	0.760
MCKL (λ known)			
Mean	−1.97	−2.21	−2.65
Std. Dev.	0.0896	0.180	0.175
RMSE	0.0980	0.193	0.187
MCKL (λ unknown)			
Mean	−2.76	−3.02	−3.82
Std. Dev.	0.951	1.05	0.959
RMSE	1.21	1.37	1.56

[a] $\log(\sigma_v) = -2.01$.
[b] $\log(\sigma_v) = -2.14$.
[c] $\log(\sigma_v) = -2.60$.

Table 7. Confidence interval (CI) performance of errors-in-variables (EV; λ known) and Monte Carlo Kernel Likelihood (MCKL; λ known vs. unknown) for the hake (*Merluccius capensis* × *M. paradoxus*) example.

	Parameters used for simulation		
Expected percentile	$\boldsymbol{\Theta}_{.01}$	$\boldsymbol{\Theta}_{.05}$	$\boldsymbol{\Theta}_{.10}$
EV (λ known)			
0.05	1	1	1
0.1	1	1	1
0.25	1	1	1
0.5	1	1	1
MCKL (λ known)			
0.05	0.02	0.03	0.05
0.1	0.02	0.05	0.06
0.25	0.06	0.13	0.18
0.5	0.28	0.35	0.36
MCKL (λ unknown)			
0.05	0.06	0.09	0.06
0.1	0.12	0.14	0.11
0.25	0.27	0.30	0.31
0.5	0.41	0.48	0.49

Note: Row labels give nominal p value cutoff and table entries give actual number of simulations for which the true parameters were rejected using that cutoff.

To compare performance of MCKL and EV, we simulated 100 data sets using the real catch values and each of three sets of assumed parameter values, which were themselves estimated from the data using the MCKL method. For MCKL, minimum values for σ_ε and σ_v were set to 0.01. The three sets of parameters were (*i*) the MCKL maximum likelihood estimate (MLE) for all parameters, (*ii*) the MCKL MLE of (r, K, q, σ_v) with σ_ε fixed at 0.05, and (*iii*) the MCKL MLE of (r, K, q, σ_v) with σ_ε fixed at 0.10. The full MLE yielded $\hat{\sigma}_\varepsilon = 0.01$, the minimum value allowed because MCMC sampling is inefficient for smaller $\hat{\sigma}_\varepsilon$ values, so the latter two cases were used to include meaningful amounts of observation error in the simulations. This model-fitting problem showed a difficulty in distinguishing process and observation error similar to, but not as severe as, the linear autoregressive example. We label the cases as $\boldsymbol{\Theta}_{.01}$, $\boldsymbol{\Theta}_{.05}$ and $\boldsymbol{\Theta}_{.10}$, with subscript indicating σ_ε.

For each of the three parameter scenarios, we simulated 100 sets of survey index time series using the true catch records. Then we estimated all parameters in three ways: EV with the correct λ, MCKL with λ assumed known and correct, and MCKL with λ unknown. As before, EV and MCKL with λ known differ only by using different likelihoods but share the same parameter space, while EV with λ unknown has a larger parameter space because it estimates both error variances.

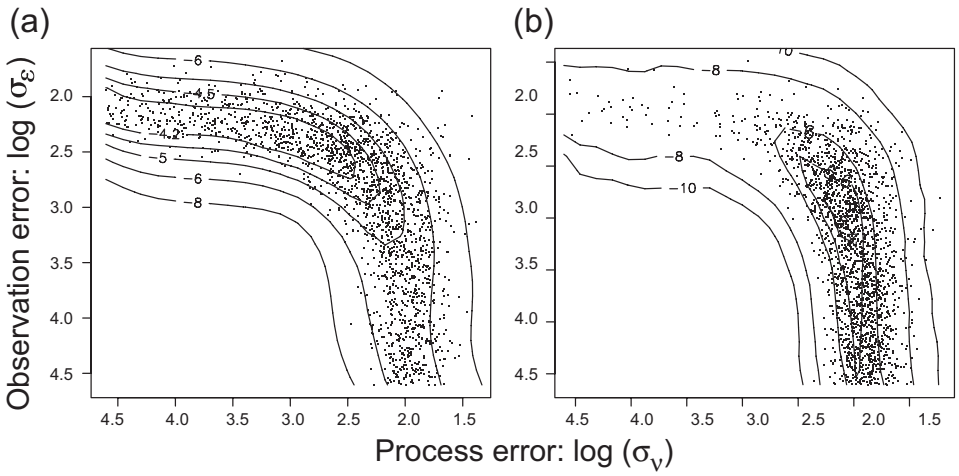
Implementation details for the hake example

We used a standard Nelder–Mead (Press et al. 1992; *R* "optim" function) algorithm for all optimizations, with as many restarts as necessary to ensure convergence. For EV, we optimized from two sets of initial conditions: one based on a no-process-error (no-PE) state trajectory, the other based on a no-observation-error (no-OE) state trajectory, given the data and true parameters in both cases. Starting from each of these initial state trajectories, we optimized the parameters only, and then from those parameters we jointly optimized parameters and states. For the final estimates, we used the highest EV likelihood result from either the no-OE or no-PE initial conditions. Using the true parameters to estimate initial state conditions for the full maximization problem was cheating in the sense that we used the correct parameters to initialize the algorithm, but this allowed a high chance of locating the global EV maximum likelihood for proper comparison with MCKL, which was the primary goal here.

For MCKL, we used a nearly uninformative prior and a Metropolis–Hastings algorithm (Gilks et al. 1996; Robert and Casella 1999) to sample from $P[\log(\boldsymbol{\Theta}), \log(\mathbf{B}^T)|\mathbf{I}^T]$ (Appendix B). We kept every 100th sample to obtain $m = 40\,000$ points. For a single analysis, we would use a much larger sample, but here we used smaller samples to facilitate comparison of hundreds of replicated data series. For estimation of the three sets of parameters from which we simulated, we used samples of $m = 5 \times 10^6$.

In MCMC posterior sampling, the issue arises of how to treat the distribution of initial state values. One objective approach would be to use the stationary distribution of the model, but this is not generally analytically available. We used an approach that approximates the stationary distribution by including 10 years of unknown state variables before the first data, with the state at the first year (i.e., $t = -10$ if the data start at $t = 0$) distributed as $N(K, \sigma_v^2)$, where K is the equilibrium value in the absence of harvest. This trick is limited to the case where the harvest data begin at the onset of substantial harvest, so that pre-data harvest can be assumed to be 0, which is approximately true for the hake case. This approach works because given any fixed initial state, the distribution of states will be the stationary distribution after a reasonable amount of time. Since we allow 10 years of stochastic state dynamics, unconstrained by observations, the resulting distribution for the tenth year (i.e., first year of data)

Fig. 6. Two example profile likelihood contours for $\log(\sigma_v)$ and $\log(\sigma_\varepsilon)$ plotted over (thinned) marginal samples. Each figure shows 2000 scatter-plot points of a marginal posterior sample of $\log(\sigma_v)$ and $\log(\sigma_\varepsilon)$ from simulated hake data. Contours show the profile likelihood (not the marginal density). For each point on a 25×25 grid of $\log(\sigma_v)$ and $\log(\sigma_\varepsilon)$, the Monte Carlo Kernel Likelihood (MCKL) was maximized over the remaining three parameter dimensions, and the contours show approximate surfaces of equal maximum likelihood. Contour labels are log likelihoods plus an unknown constant. In (*a*), the MCKL maximum has the minimum allowed value for σ_v, but in (*b*), both σ_v and σ_ε have intermediate values at the MCKL maximum. In both cases, the likelihood ridge over a wide range of σ_v and σ_ε illustrates the difficulty of separating process and observation error.



will approximate the stationary distribution.

Maximization for MCKL also used the Nelder–Mead algorithm with multiple restarts. The algorithm was started at the mean posterior $\boldsymbol{\Theta}$ value, without any input based on true parameters. For MCKL with λ known, we used the full posterior sample but maximized under the constraint of known λ. For kernel density estimation, the kernel covariance was set to $\boldsymbol{\Sigma}_\mathbf{h} = 0.1\hat{\boldsymbol{\Sigma}}_{\pi(\boldsymbol{\Theta}|I)}$, where $\hat{\boldsymbol{\Sigma}}_{\pi(\boldsymbol{\Theta}|I)}$ is the estimated covariance matrix of the posterior $\boldsymbol{\Theta}$ sample. The kernel density sometimes overlapped $\log(\sigma_v) < \log(0.01)$ or $\log(\sigma_\varepsilon) < \log(0.01)$ (from which no points were sampled), so we renormalized the kernel density accordingly, allowing MCKL estimation of $\hat{\sigma_v}$ or $\hat{\sigma_\varepsilon}$ near their minimum values.

Results for the hake example

Comparing EV and MCKL with λ known shows that EV is more biased, has higher RMSE, and has very inaccurate confidence regions, while MCKL with λ known has low bias and RMSE and usefully approximates confidence regions (Tables 4–7; Figs. 3–5). Comparing MCKL with λ known vs. unknown shows that estimates of ρ are similar, but when λ is unknown, separation of process and observation error is again difficult.

The difficulty of distinguishing process vs. observation error is illustrated by two examples of likelihood profiles (i.e. maximized over other parameters) in the $\log(\sigma_v) \times \log(\sigma_\varepsilon)$ plane, plotted over a sample of the marginal posterior for illustration (Fig. 6). As illustrated in the first example, MCKL with unknown λ sometimes estimates an all-process-error or all-observation-error model. Nevertheless, the confidence regions for MCKL with λ unknown (reflected in p values of the true parameters) provide a useful estimate of uncertainty, even if the uncertainty is large. In many cases, the confidence region would include a span from all-process-error to all-observation-error parameters, reflecting the high level of uncertainty in these parameters.

Discussion

MCKL provides a method to calculate likelihoods from nonlinear state-space models with accuracy that increases as Monte Carlo sample size (computational effort) increases. Implementation of MCKL requires a method for obtaining a Bayesian posterior sample, generally by MCMC. Relevant MCMC samplers have been under active development in recent years. The additional steps for MCKL of calculating and maximizing a weighted kernel density estimate of a Bayesian posterior is relatively simple to implement.

The errors-in-variables method has long been recognized to be biased and requires difficult assumptions about the relative magnitudes of process and observation errors. Our study further explores these difficulties by comparison with state-space likelihoods that integrate over the distribution of unknown state trajectories. For both a linear and nonlinear model, integrated state-space likelihoods outperformed EV likelihoods. Integrated state-space likelihoods do not require assumptions about relative error magnitudes, although separation of process and observation error is an inherently noisy problem. Comparison of methods using other data and more complex models, such as age- or size-structured models, is an important topic for further investigation.

Fisheries scientists have increasingly turned to Bayesian methods in recent years. Our purpose here is not to compare and contrast Bayesian and classical methods but rather to separate the Bayesian vs. classical choice from model complexity, so that the best aspects of each approach are available. Bayesian methods have been highly associated with full incorporation of process and observation error for complex fisheries models largely because the boom in MCMC methods made posterior

sampling feasible. With MCKL, classical likelihood analysis can also fully incorporate both types of error, so the pros and cons of each approach may be brought to bear on the same problems.

Acknowledgements

We thank Jon Schnute, an anonymous reviewer, and the Associate Editor for helpful comments. The bulk of this work was conducted while the first author was a Postdoctoral Associate at the National Center for Ecological Analysis and Synthesis, a center funded by the National Science Foundation (grant No. DEB-0072909), the University of California, and the Santa Barbara campus.

References

Booth, J.G., and Hobert, J.P. 1999. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. J.R. Stat. Soc. Ser. B, **61**: 265–285.

Chan, K., and Ledolter, J. 1995. Monte Carlo EM estimation for time series models involving counts. J. Am. Stat. Assoc. **90**: 242–252.

Collie, J.S., and Sissenwine, M.P. 1983. Estimating population size from relative abundance data measured with error. Can. J. Fish. Aquat. Sci. **40**: 1871–1879.

de Valpine, P. 2003. Better inferences from population dynamics experiments using Monte Carlo state-space likelihoods. Ecology, **84**: 3064–3077.

de Valpine, P. 2004. Monte Carlo state-space likelihoods by weighted posterior kernel density estimation. J. Am. Stat. Assoc. **99**: 523–536.

de Valpine, P., and Hastings, A. 2002. Fitting population models with process noise and observation error. Ecol. Monogr. **72**: 51–76.

Durbin, J., and Koopman, S.J. 1997. Monte Carlo maximum likelihood estimation for non-Gaussian state space models. Biometrika, **84**: 669–684.

Durbin, J., and Koopman, S.J. 2000. Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. J.R. Stat. Soc. Ser. B, **62**: 3–56.

Freeman, S., and Kirkwood, G. 1994. On a structural time series method for estimating stock biomass and recruitment from catch and effort data. Fish. Res. **22**: 77–98.

Geyer, C.J., and Thompson, E.A. 1992. Constrained Monte Carlo maximum likelihood for dependent data. J.R. Stat. Soc. Ser. B, **54**: 657–699.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (*Editors*). 1996. Markov Chain Monte Carlo in practice. Chapman & Hall, New York.

Gordon, N.J., Salmond, D.J., and Smith, A.F.M. 1993. Novel approach to nonlinear non-Gaussian Bayesian state estimation. IEE Proceedings F, **140**: 107–113.

Gudmundsson, G. 1994. Time series analysis of catch-at-age observations. Appl. Stat. **43**: 117–126.

Harvey, A.C. 1989. Forecasting, structural time series models, and the Kalman filter. Cambridge University Press, Cambridge, UK.

Harvey, A.C. 1993. Time series models. 2nd ed. MIT Press, Cambridge, Mass.

Hilborn, R., and Mangel, M. 1997. The ecological detective: confronting models with data. Princeton University Press, Princeton, N.J.

Hürzeler, M., and Künsch, H.R. 2001. Approximating and maximising the likelihood for a general state-space model. *In* Sequential Monte Carlo methods in practice. *Edited by* A. Doucet, N. de Freitas, and N. Gordon. Springer, New York. pp. 159–175.

Johnson, R.A., and Wichern, D.W. 1992. Applied multivariate statistical methods. 3rd ed. Prentice-Hall, Inc., Upper Saddle River, N.J.

Kimura, D.K., Balsiger, J.W., and Ito, D.H. 1996. Kalman filtering the delay-difference equation: practical approaches and simulations. Fish. Bull. **94**: 678–691.

Kitagawa, G. 1987. Non-Gaussian state-space modeling of nonstationary time series (with Discussion). J. Am. Stat. Assoc. **82**: 1032–1063.

Ludwig, D., and Walters, C.J. 1981. Measurement errors and uncertainty in parameter estimates for stock and recruitment. Can. J. Fish. Aquat. Sci. **38**: 711–720.

Ludwig, D., and Walters, C.J. 1989. A robust method for parameter estimation from catch and effort data. Can. J. Fish. Aquat. Sci. **46**: 137–144.

Ludwig, D., Walters, C.J., and Cooke, J. 1988.Comparison of two models and two estimation methods for catch and effort data. Nat. Res. Model. **2**: 457–498.

McAllister, M.K., Pikitch, E.K., Punt, A.E., and Hilborn, R. 1994. A Bayesian approach to stock assessment and harvest decisions using the sampling/importance resampling algorithm. Can. J. Fish. Aquat. Sci. **51**: 2673–2687.

McCulloch, C.E. 1997. Maximum likelihood algorithms for generalized linear models. J. Am. Stat. Assoc. **92**: 162–170.

Mendelssohn, R. 1988. Some problems in estimating population sizes from catch-at-age data. Fish. Bull. **86**: 617–630.

Meyer, R., and Millar, R.B. 1999. Bayesian stock assessment using a state-space implementation of the delay difference model. Can. J. Fish. Aquat. Sci. **56**: 37–52.

Millar, R.B., and Meyer, R. 2000. Non-linear state space modelling of fisheries biomass dynamics by using Metropolis-Hastings within-Gibbs sampling. Appl. Stat. **49**: 327–342.

Otter Research. 2000. An introduction to AD Model Builder. Version 4. Otter Research, Sidney, B.C.

Pella, J.J. 1993. Utility of structural time series models and the Kalman Filter for predicting consequences of fishery actions. *In* Proceedings of the International Symposium on Management Strategies for Exploited Fish Population. *Edited by* G. Kruse, D. Eggers, R. Marasco, C. Pautzke, and T.J. Quinn, II. Alaska Sea Grant College Program Report No. 93-02. University of Alaska, Fairbanks. pp. 571–593.

Pitt, M.K., and Shephard, N. 1999. Filtering via simulation: auxiliary particle filters. J. Am. Stat. Assoc. **94**: 590–599.

Polacheck, T., Hilborn, R., and Punt, A.E. 1993. Fitting surplus production models: comparing methods and measuring uncertainty. Can. J. Fish. Aquat. Sci. **50**: 2597–2607.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. 1992. Numerical recipes in C: the art of scientific computing. 2nd ed. Cambridge University Press, Cambridge, UK.

Punt, A.E., and Hilborn, R. 1997. Fisheries stock assessment and decision analysis: the Bayesian approach. Rev. Fish Biol. Fish. **7**: 35–63.

Richards, L.J., Schnute, J.T., and Olsen, N. 1997. Visualizing catch-age analysis: a case study. Can. J. Fish. Aquat. Sci. **54**: 1646–1658.

Robert, C.P., and Casella, G. 1999. Monte Carlo statistical methods. Springer, New York.

Schnute, J.T. 1994. A general framework for developing sequential fisheries models. Can. J. Fish. Aquat. Sci. **51**: 1676–1688.

Schnute, J.T., and Kronlund, A. 2002. Estimating salmon stock–recruitment relationships from catch and escapement data. Can. J. Fish. Aquat. Sci. **59**: 433–449.

Schnute, J.T., Mulligan, T., and Kuhn, B. 1990. An errors-in-variables bias model with an application to salmon hatchery data. Can. J. Fish. Aquat. Sci. **47**: 1453–1467.

Schnute, J.T., and Richards, L.J. 1995. The influence of error on population estimates from catch-age models. Can. J. Fish. Aquat. Sci. **52**: 2063–2077.

Scott, D. 1992. Multivariate density estimation. John Wiley & Sons, New York.

Severini, T.A. 2000. Likelihood methods in statistics. Oxford University Press, New York.

Silverman, B. 1986. Density estimation for statistics and data analysis. Chapman and Hall, New York.

Sullivan, P.J. 1992. A Kalman Filter approach to catch-at-length analysis. Biometrics, **48**: 237–257.

Wei, G.C.G., and Tanner, M.A. 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J. Am. Stat. Assoc. **85**: 699–704.

Appendix A. EV estimation of autoregressive model.

Much of this appendix depends on assuming μ known. Since estimates of μ are well-behaved for both KF and EV methods, this assumption does not greatly confound the analytical results developed.

Derivation of l^{EV} for the AR (Example 1) case

We begin by deriving a relatively simple expression for the log of the EV likelihoods, $l^{\mathrm{EV}} = \log(L^{\mathrm{EV}})$. Throughout this appendix, we drop the superscript "T" notation, so $\boldsymbol{X} \equiv \boldsymbol{X}^T$ and $\boldsymbol{Y} \equiv \boldsymbol{Y}^T$, and we drop $\boldsymbol{\Theta}$ from the probability notation, so $P(\cdot) \equiv P(\cdot|\boldsymbol{\Theta})$.

The covariance matrix of $\boldsymbol{X}$ is $\boldsymbol{\Sigma_{XX}}$, where $(\boldsymbol{\Sigma_{XX}})_{i,j} = \frac{\sigma_v^2 \rho^{|i-j|}}{1-\rho^2}$. The covariance matrix of $\boldsymbol{Y}$ is $\boldsymbol{\Sigma_{YY}} = \boldsymbol{\Sigma_{XX}} + \sigma_\varepsilon^2 \mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

For the linear Gaussian case, the EV likelihood is

$$\text{(A1)} \quad L^{\mathrm{EV}} = P[\boldsymbol{Y}|(\boldsymbol{X} = \hat{\boldsymbol{X}})]P(\boldsymbol{X} = \hat{\boldsymbol{X}})$$

where $\hat{\boldsymbol{X}}$ is the estimated state trajectory that maximizes L^{EV}. Because the distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are multivariate normal, the estimated states are equal to the mean of the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y}$ (cf. Schnute 1994):

$$\text{(A2)} \quad \hat{\boldsymbol{X}} - \mu_{\boldsymbol{X}} = \boldsymbol{\Sigma_{XY}} \boldsymbol{\Sigma_{YY}^{-1}} (\boldsymbol{Y} - \mu_{\boldsymbol{Y}})$$

where $\boldsymbol{\Sigma_{XY}}$ is the matrix of covariances between elements of $\boldsymbol{X}$ and $\boldsymbol{Y}$: $(\boldsymbol{\Sigma_{XY}})_{ij} = \mathrm{cov}(X_i, Y_j)$. Note that $\boldsymbol{\Sigma_{XY}} = \boldsymbol{\Sigma_{XX}}$.

The distribution of $\boldsymbol{Y} - \mu_{\boldsymbol{Y}}$ given $\boldsymbol{X}$ is

$$\text{(A3)} \quad (\boldsymbol{Y} - \mu_{\boldsymbol{Y}})|\boldsymbol{X} \sim N[\boldsymbol{X} - \mu_{\boldsymbol{X}}, \boldsymbol{\Sigma_{YY}} - \boldsymbol{\Sigma_{YX}} \boldsymbol{\Sigma_{XX}^{-1}} \boldsymbol{\Sigma_{XY}}]$$

where $N(\boldsymbol{a}, \mathbf{b})$ denotes a normal distribution with mean vector $\boldsymbol{a}$ and covariance matrix $\mathbf{b}$. Note the covariance matrix of $(\boldsymbol{Y} - \mu_{\boldsymbol{Y}})|\boldsymbol{X}$ reduces to $\sigma_\varepsilon^2 \mathbf{I}$.

Using eqs. A2 and A3, the first term of l^{EV} is

$$\text{(A4)} \quad \log\{P[\boldsymbol{Y}|(\boldsymbol{X} = \hat{\boldsymbol{X}})]\} = -n \log(\sigma_\varepsilon) - \frac{\sigma_\varepsilon^2 (\boldsymbol{Y} - \mu_{\boldsymbol{Y}})' (\boldsymbol{\Sigma_{YY}^{-1}})^2 (\boldsymbol{Y} - \mu_{\boldsymbol{Y}})}{2}$$

Together with

$$\text{(A5)} \quad \log P(\boldsymbol{X} = \hat{\boldsymbol{X}}) = -\frac{\log(|\boldsymbol{\Sigma_{XX}}|)}{2} - \frac{1}{2}(\boldsymbol{Y} - \mu_{\boldsymbol{Y}})' [\boldsymbol{\Sigma_{YY}^{-1}} - \sigma_\varepsilon^2 (\boldsymbol{\Sigma_{YY}^{-1}})^2] (\boldsymbol{Y} - \mu_{\boldsymbol{Y}})$$

we have for the log EV likelihood:

$$\text{(A6)} \quad l^{\mathrm{EV}} = -\frac{1}{2} \log(|\sigma_\varepsilon^2 \boldsymbol{\Sigma_{XX}}|) - \frac{1}{2}(\boldsymbol{Y} - \mu_{\boldsymbol{Y}})' \boldsymbol{\Sigma_{YY}^{-1}} (\boldsymbol{Y} - \mu_{\boldsymbol{Y}})$$

The log true likelihood, l^{KF}, is given simply by the normal distribution for $\boldsymbol{Y}$:

$$\text{(A7)} \quad l^{\mathrm{KF}} = -\frac{1}{2} \log(|\boldsymbol{\Sigma_{YY}}|) - \frac{1}{2}(\boldsymbol{Y} - \mu_{\boldsymbol{Y}})' \boldsymbol{\Sigma_{YY}^{-1}} (\boldsymbol{Y} - \mu_{\boldsymbol{Y}})$$

which is equivalent to the KF.

Comparing eqs. A6 and A7, we see that l^{EV} mishandles the variances, since $|\boldsymbol{\Sigma_{YY}}| = |\boldsymbol{\Sigma_{XX}} + \sigma_\varepsilon^2 \mathbf{I}|$, so

$$\text{(A8)} \quad |\boldsymbol{\Sigma_{YY}}| \neq |\sigma_\varepsilon^2 \boldsymbol{\Sigma_{XX}}|$$

Note that $|\sigma_\varepsilon^2 \boldsymbol{\Sigma_{XX}}|$ is $\sigma_\varepsilon^{2n} \sigma_v^{2n}$ times a function of ρ, while $|\boldsymbol{\Sigma_{YY}}|$ is σ_ε^{2n} times a function of ρ and λ; therefore, l^{EV} includes twice as many variance terms in its normalization constant than does the true likelihood l^{KF}.

Asymptotic EV parameter estimates

We can obtain asymptotic expected parameter estimates for the EV method by maximizing the expected log likelihood. This is justified because as long as the random parts of l^{EV} converge to fixed values, we can take expectations of the EV likelihood and then maximize in the parameters, rather than maximizing in the parameters and then taking expectations. We do this with Matlab (The MathWorks Inc., Natick, Massachusetts) using a simplified expression for l^{EV}.

We simplify l^{EV} by simplifying $|\boldsymbol{\Sigma_{XX}}|$ and $E[(\boldsymbol{Y} - \mu_{\boldsymbol{Y}})' \times \boldsymbol{\Sigma_{YY}^{-1}} (\boldsymbol{Y} - \mu_{\boldsymbol{Y}})]$. The notation $E[h(z)]$ denotes the expected value of $h(z)$. For the first part, we have

$$\text{(A9)} \quad |\boldsymbol{\Sigma_{XX}}| = \frac{1}{|\boldsymbol{\Sigma_{XX}^{-1}}|} = \frac{\sigma_v^{2n}}{1-\rho^2}$$

To obtain this, consider the case $\sigma_v^2 = 1$, so $\boldsymbol{\Sigma_{XX}^{-1}}$ has 1 in the upper left and lower right entries, $1 + \rho^2$ on all other diagonal entries, $-\rho$ above and below the diagonal, and 0 elsewhere. The determinant of this matrix is $1 - \rho^2$.

For the other piece of l^{EV}, we can use the following relations. If A is a multivariate, normal, random variable with covariance matrix $\boldsymbol{\Sigma}_{AA}$, we will be interested in the expected value of $A' \boldsymbol{\Sigma_{BB}^{-1}} A$, where $\boldsymbol{\Sigma}_{BB}$ is some covariance matrix. Write

$\mathbf{\Sigma}_{BB}^{-1} = Q'_B \mathbf{\Lambda}_B^{-1/2} \mathbf{\Lambda}_B^{-1/2} \mathbf{Q}_B$. This is a standard factorization of covariance matrices (e.g., Johnson and Wichern 1992), where $\mathbf{Q}_B$ is an orthonormal rotation matrix, and $\mathbf{\Lambda}_B^{-1/2}$ is a diagonal matrix with entries that are the square roots of the eigenvalues. Now $Z = \mathbf{\Lambda}_B^{-1/2} \mathbf{Q}_B A$ is a random variable with covariance

$$\text{(A10)} \quad \mathbf{\Sigma}_{ZZ} = \mathbf{\Lambda}_B^{-1/2} Q'_B \mathbf{\Sigma}_{AA} \mathbf{Q}_B \mathbf{\Lambda}_B^{-1/2}$$

We want the expected value of $Z'Z$. This is given by

$$\begin{aligned} \text{(A11)} \quad E[Z'Z] &= \text{tr}(\mathbf{\Sigma}_{ZZ}) \\ &= \text{tr}(\mathbf{\Lambda}_B^{-1/2} Q'_B \mathbf{\Sigma}_{AA} \mathbf{Q}_B \mathbf{\Lambda}_B^{-1/2}) \\ &= \text{tr}(\mathbf{\Sigma}_{AA} \mathbf{\Sigma}_{BB}^{-1}) \end{aligned}$$

where for the last step tr denotes trace, and the relation $\text{tr}(CD) = \text{tr}(DC)$ has been used with $C = \mathbf{\Sigma}_{AA} \mathbf{Q}_B \mathbf{\Lambda}_B^{-1/2}$, $D = \mathbf{\Lambda}_B^{-1/2} Q'_B$.

Applied to the equation for l^{EV}, we get that if candidate parameters are given by $\hat{\mathbf{\Theta}}$, then

$$\text{(A12)} \quad E[(\boldsymbol{Y} - \mu)' \mathbf{\Sigma}_{\boldsymbol{YY},\hat{\mathbf{\Theta}}}^{-1} (\boldsymbol{Y} - \mu)] = \text{tr}(\mathbf{\Sigma}_{\boldsymbol{Ym}} \mathbf{\Sigma}_{\boldsymbol{YY},\hat{\mathbf{\Theta}}}^{-1})$$

Putting these pieces together gives

$$\begin{aligned} \text{(A13)} \quad E[l(\mathbf{\Theta})] = -\frac{1}{2}\{n[\log(\hat{\sigma}_\varepsilon^2 \hat{\sigma}_\nu^2)] - \log(1 - \hat{\rho}^2) \\ - \text{tr}(\mathbf{\Sigma}_{\boldsymbol{YY}} \mathbf{\Sigma}_{\boldsymbol{YY},\hat{\mathbf{\Theta}}}^{-1})\} \end{aligned}$$

The usefulness of this equation is that $\text{tr}(\mathbf{\Sigma}_{\boldsymbol{YY}} \mathbf{\Sigma}_{\boldsymbol{YY},\hat{\mathbf{\Theta}}}^{-1})$ can be easily computed and eq. A13 easily maximized computationally. Even with small values of n (e.g., 20), and hence computationally quick matrix sizes, the answer is only negligibly different than the answer with much larger matrix sizes. As a final note on the justification of this calculation, the trace term actually increases with n, but if one divides by n, the distribution of $\frac{1}{n}(\boldsymbol{Y} - \mu)' \mathbf{\Sigma}_{\boldsymbol{YY},\hat{\mathbf{\Theta}}}^{-1} (\boldsymbol{Y} - \mu)$ converges appropriately. Severini (2000) discusses this more generally and formally.

Asymptotic estimate of variance for true ρ

To obtain insight into EVs handling of variance estimation, consider the artificial case of known autocorrelation $\hat{\rho} = \rho$. Denote $\sigma^2 = \sigma_\varepsilon^2$ and define $\mathbf{\Sigma}^*$ by $\mathbf{\Sigma} = \sigma^2 \mathbf{\Sigma}^*$. Then

$$\begin{aligned} \text{(A14)} \quad l^{\text{EV}} = -2n \log(\sigma) - \\ \frac{\log(|\mathbf{\Sigma}^*_{\boldsymbol{XX}}|)}{2} - \frac{(\boldsymbol{Y} - \mu_{\boldsymbol{Y}})(\mathbf{\Sigma}^*_{\boldsymbol{YY}})^{-1}(\boldsymbol{Y} - \mu_{\boldsymbol{Y}})}{2\sigma^2} \end{aligned}$$

Setting the derivative with respect to σ equal to zero gives the l^{EV} estimate:

$$\text{(A15)} \quad \hat{\sigma}^2 = \frac{(\boldsymbol{Y} - \mu_{\boldsymbol{Y}})(\mathbf{\Sigma}^*_{\boldsymbol{YY}})^{-1}(\boldsymbol{Y} - \mu_{\boldsymbol{Y}})}{2n}$$

Even if we know the correct values of ρ and μ, so that $E[(\boldsymbol{Y} - \mu_{\boldsymbol{Y}})(\mathbf{\Sigma}^*_{\boldsymbol{YY}})^{-1}(\boldsymbol{Y} - \mu_{\boldsymbol{Y}})] = n\sigma^2$, this estimator is biased by 50%:

$$\text{(A16)} \quad E(\hat{\sigma}^2) = \sigma^2/2$$

Although we will not generally know ρ, this result provides insight into the manner in which the EV method underestimates the true variance by using twice as many variances in the normalization constant compared with the true likelihood. Related results are discussed by Ludwig and Walters (1981), Schnute and Kronlund (2002), and references therein.

Asymptotic EV distribution of $\hat{\boldsymbol{X}}$

Given estimates $\hat{\rho}$ and $\hat{\sigma}^2$, we can calculate the distribution of $\hat{\boldsymbol{X}}$ from eq. A2. This is

$$\text{(A17)} \quad \hat{\boldsymbol{X}} - \mu_{\boldsymbol{X}} \sim N(0, \hat{\mathbf{\Sigma}}_{\boldsymbol{XX}} \hat{\mathbf{\Sigma}}_{\boldsymbol{YY}}^{-1} \mathbf{\Sigma}_{\boldsymbol{YY}} \hat{\mathbf{\Sigma}}_{\boldsymbol{YY}}^{-1} \hat{\mathbf{\Sigma}}_{\boldsymbol{XX}})$$

The covariances between $(\hat{\boldsymbol{X}} - \mu_{\boldsymbol{X}})$ and $(\boldsymbol{X} - \mu_{\boldsymbol{X}})$ are given by the matrix $\hat{\mathbf{\Sigma}}_{\boldsymbol{XX}} \hat{\mathbf{\Sigma}}_{\boldsymbol{YY}}^{-1} \mathbf{\Sigma}_{\boldsymbol{XX}}$

This matrix can be computed easily with mathematical software, and the appropriate terms can be used to calculate the slope of $\hat{\boldsymbol{X}}_T$ on $\boldsymbol{X}_T$:

$$\text{(A18)} \quad b = \frac{\text{cov}(\hat{\boldsymbol{X}}_T, \boldsymbol{X}_T)}{\text{var}(\hat{\mathbf{X}}_T)}$$

Hessians

In standard likelihood-based methods, one can obtain approximate CIs by using the fact that parameter estimates are asymptotically normally distributed. Some investigators (see text) use this result in the different context where parameters and states are included together and are hoped to be asymptotically normal in a similar way. Following the example of Richards et al. (1997), let $\mathbf{\Phi} = (\mathbf{\Theta}, \boldsymbol{X})$ (they use $\mathbf{\Theta}$ where we have used $\mathbf{\Phi}$). They propose that

$$\text{(A19)} \quad l(\mathbf{\Phi}) \approx l(\hat{\mathbf{\Phi}}) - \frac{1}{2}(\mathbf{\Phi} - \hat{\mathbf{\Phi}})' J(\hat{\mathbf{\Phi}})(\mathbf{\Phi} - \hat{\mathbf{\Phi}})$$

where

$$\text{(A20)} \quad J(\hat{\mathbf{\Phi}}) = -\frac{\partial^2 l}{\partial \mathbf{\Phi}^2}|_{\mathbf{\Phi} = \hat{\mathbf{\Phi}}}$$

Under standard likelihood theory, this approximation is not valid, but would be valid if $\mathbf{\Phi}$ were replaced with $\mathbf{\Theta}$ (i.e., the parameters but not the states).

For the states, the elements of the Hessian are

$$\text{(A21)} \quad -\frac{\partial^2 l}{\partial \hat{\boldsymbol{X}}_i^2} = \frac{1}{\sigma_\nu^2}\left(\frac{1}{\gamma} + 1 + \rho^2\right), \quad i = 1, ..., T-1$$

$$\text{(A22)} \quad -\frac{\partial^2 l}{\partial \hat{\boldsymbol{X}}_i \partial \hat{\boldsymbol{X}}_{i+1}} = -\frac{\rho}{\sigma_\nu^2}, \quad i = 1, ..., T-1$$

$$\text{(A23)} \quad -\frac{\partial^2 l}{\partial \hat{\boldsymbol{X}}_T^2} = \frac{1}{\sigma_\nu^2}\left(\frac{1}{\gamma} + 1\right)$$

where $\gamma = \sigma_\varepsilon^2/\sigma_\nu^2$. Since these values are constant even as the amount of data increases, this is not the type of likelihood situation to which the usual asymptotic results apply. One could obtain analytic expressions for other elements of the Hessian, but for simplicity we calculated Hessians numerically using standard finite element methods.

Appendix B. MCMC sampler for the hake model.

The MCMC algorithm sampled from the density $P(\boldsymbol{\Theta}, B_{-10}, \ldots, B_0, \ldots, B_T | I_0, \ldots, I_T) = P(\boldsymbol{\Theta}, \mathbf{B}|\mathbf{I})$, where the catch and index record starts at time 0. We used the log of all parameters, $\boldsymbol{\Theta} = (\log r, \log K, \log q, \log \sigma_v, \log \sigma_\varepsilon)$, as the sampling space. Biomass values for $t < 0$ are included in the model so that B_0 is distributed approximately as the stationary distribution of the model, which works because harvest levels were minor for $t < 0$ (see text). We define $C_t = 0$ for $t < 0$. For MCMC sampling, one only needs to be able to calculate the target density up to an unknown constant factor. In this case, we have $P(\boldsymbol{\Theta}, \mathbf{B}|\mathbf{I}) \propto P(\boldsymbol{\Theta}, \mathbf{B}, \mathbf{I})$, and the latter density was used to calculate MCMC acceptance probabilities. We write $\phi(x, \mu, \sigma)$ for the normal distribution density for value x given mean μ and standard deviation σ.

The total probability can be factored as

$$\text{(B1)} \quad P(\boldsymbol{\Theta}, \mathbf{B}, \mathbf{I}) = P_{\text{Prior}}(\boldsymbol{\Theta}) P(B_{-10}|\boldsymbol{\Theta}) \times \left[\prod_{t=-9}^{T} P(B_t|B_{t-1})\right]\left[\prod_{t=0}^{T} P(I_t|B_t)\right]$$

where

$$\text{(B2)} \quad P(B_t|B_{t-1}, \boldsymbol{\Theta}) = \phi\{\log(B_t) - \log[B_{t-1} + r B_{t-1}(1 - \frac{B_{t-1}}{K}) - C_{t-1}], 0, \sigma_v\}, \quad t > -10$$

$$\text{(B3)} \quad P(B_{-10}|\boldsymbol{\Theta}) = \phi[\log(B_{-10}), \log(K), \sigma_v]$$

$$\text{(B4)} \quad P(I_t|B_t) = \phi[\log(I_t) - \log(q B_t), 0, \sigma_\varepsilon]$$

The prior $P_{\text{Prior}}(\boldsymbol{\Theta})$ was the product of independent priors for each parameter. For MCMC sampling, the priors need only be specified up to a constant. Priors for $\log(r)$, $\log(\sigma_v)$, and $\log(\sigma_\varepsilon)$ were each exponential with rate 0.0001 and minimum value $\log(0.01)$. For example

$$\text{(B5)} \quad P_{\text{Prior}}[\log(r)] = D \exp[-0.0001 \log(r)], \quad \log(r) > 0.01$$

where D is a normalizing constant that does not matter in the Metropolis–Hastings acceptance probability. The prior for $\log(K)$ was constant between 0 and $\log(8000)$, and that for $\log(q)$ was constant between $\log(10^{-5})$ and $\log(10^{-2})$.

The sampler iterated through a sequence of MCMC steps with one-dimensional proposals densities as well a joint [$\log(q)$, **B**] proposal. Here, prime indicates a proposal value, $N(\mu, \sigma)$ indicates a normal distribution, and $N_r(\mu, \sigma, \text{min or max})$ indicates a reflected normal distribution with mean μ, standard deviation σ, and minimum or maximum values specified. For example, if $(x'|x) \sim N_r(x, 1, 10)$ and $x < 10$, then $P(9.5|9) = \phi(0.5, 0, 1) + \phi(1.5, 0, 1)$ (i.e., a normal draw of 1.5 given $x = 9$ would be reflected at $x' = 10$ to give a proposal value of $x' = 9.5$). In practice, only the backwards-to-forwards ratio $P(x|x')/P(x'|x)$ enters the MCMC acceptance probability, and this is one for all of our proposal densities, including reflected normals. To obtain good mixing, we included for some parameters multiple proposal densities with different standard deviations to allow mixing at different scales for different parts of parameter space. One iteration of our sampler included Metropolis–Hastings steps with the following proposals:

(1) $[\log B_t' | \log(B_{t-1})] \sim N(\log B_{t-1}, 0.05)$

(2) $[\log(q') | \log(q)] \sim N_r[\log(q), 0.001, \log(0.01), \log(0.00001)]$

(3) $[\log(r') | \log(r)] \sim N_r[\log(r), 0.05, \log(0.01)]$

(4) $[\log(q') | \log(q)] \sim N_r[\log(q), 0.01, \log(0.01), \log(0.00001)]$

(5) $[\log(K') | \log(K)] \sim N_r[\log(K), 0.05, 0, \log(8000)]$

(6) $[\log(q') | \log(q)] \sim N_r[\log(q), 0.05, \log(0.01), \log(0.00001)]$

(7) $[\log(\sigma_v') | \log(\sigma_v)] \sim N_r[\log(\sigma_v), 0.05, \log(0.01)]$

(8) $[\log(\sigma_\varepsilon') | \log(\sigma_\varepsilon)] \sim N_r[\log(\sigma_\varepsilon), 0.05, \log(0.01)]$

(9) $[\log(q') | \log(q)] \sim N_r[\log(q), 0.1, \log(0.01), \log(0.00001)]$

and

(10) $[\log(q') | \log(q)] \sim N_r[\log(q), 0.05, \log(0.01), \log(0.00001)]$,

with $\log(B_t') = \log(B_t) + [\log(q) - \log(q')]$ for all t

Step 10 requires care because it samples in a transformed coordinate space, for which a Jacobian transformation of proposal probabilities is necessary, but in this case the transformation is linear and the Jacobian is constant, so the backwards-to-forwards ratio remains constant.