Computational Statistics and Data Analysis 160 (2021) 107227

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Frequentist delta-variance approximations with mixed-effects models and TMB

Nan Zheng, Noel Cadigan [*]

Centre for Fisheries Ecosystems Research, Fisheries and Marine Institute of Memorial University of Newfoundland, St. John's, NL, Canada A1C 5R3

ARTICLE INFO

Article history:
Received 14 August 2020
Received in revised form 6 March 2021
Accepted 12 March 2021
Available online 23 March 2021

Keywords:
Delta variance
Mixed effects
Penalized estimation
Template model builder

ABSTRACT

Measures of uncertainty are investigated for estimates and predictions using nonlinear mixed-effects models including state–space models in particular. These nonlinear mixed-effects models include fixed parameters and random effects. Maximum likelihood estimation of the parameters and conditional mean predictors of random effects are commonly used to estimate important quantities for a wide spectrum of applications. These quantities of interest may be functions of the parameters and random effects. In this case, software packages such as TMB and glmmTMB use a generalized delta method to provide standard errors and statistical inference. In the frequentist framework, it is clarified that these packages actually provide estimates of mean squared errors (MSE's) based on a multivariate normal approximation of the distribution of the random effects given data. It is further shown that the MSE's are not the variance of estimates due to repeated sampling of the data and the random effects. Equations are provided for that variance, including orders of approximations. In many cases the MSE's will be more appropriate to use for statistical inference, but not always, and this is demonstrated for a simple random-walk state–space model example.

1. Introduction

Mixed-effects models that include both fixed parameters and random effects are commonly used to model complex dependencies in data. An example is state–space models which are a flexible class of latent variable models commonly used in analyzing time series data (Durbin and Koopman, 2000) and have numerous applications in many fields including ecology, econometrics, engineering and environmental sciences (e.g. Kantas et al., 2015). These models involve a stochastic process equation to describe how a latent process evolves, and observation equations that link data to the latent process. State–space models have become a favored approach for modeling time varying ecological phenomena (e.g. Pedersen et al., 2011; Auger-Méthé et al., 2020). For example, state–space fish stock assessment models (Schnute, 1994; Aanes et al., 2007; Nielsen and Berg, 2014; Cadigan, 2015; Aeberhard et al., 2018; Perreault et al., 2020) that integrate multiple sources of data related to stock productivity are increasingly being used and are considered by Punt et al. (2020) to be an essential part of the next generation stock assessment package. Currently the most notable example is the SAM stock assessment package (Nielsen and Berg, 2014; Berg and Nielsen, 2016) used by many working groups of the International Council for the Exploration of the Seas (ICES) (e.g. ICES, 2019a,b). Recent versions of SAM are implemented in the Template Model Builder (TMB, Kristensen et al., 2016) package within R (R Core Team, 2018). TMB provides built-in and easy-to-use procedures for implementing nonlinear mixed-effects and state–space models. A range of generalized linear mixed

[*] Corresponding author.
E-mail address: Noel.Cadigan@mi.mun.ca (N. Cadigan).

https://doi.org/10.1016/j.csda.2021.107227
0167-9473/© 2021 Elsevier B.V. All rights reserved.

models (e.g. Brooks et al., 2019; Petersson et al., 2019) can be even more easily implemented using the glmmTMB R package (Brooks et al., 2017) which is built on TMB. Although the general focus of our research is on fish stock assessment model practice, the results of this paper are developed for a generic nonlinear mixed model framework.

The basic mixed-effects model and estimation approach are briefly described as follows, but more general reviews of linear mixed-effects models are provided by, for example, McCulloch (2003), Tuerlinckx et al. (2006) and Bolker et al. (2009), and additional details on nonlinear mixed-effects models as implemented very generically in TMB are available in Skaug and Fournier (2006) and Kristensen et al. (2016). The random response data are collected in a $n \times 1$ vector D and are assumed to have a multivariate probability density/mass function (pdf/pmf) $f(D|\Psi, \theta)$, given values of the $(p \times 1)$ vector of fixed-effects parameters θ and the $(q \times 1)$ vector of random-effects Ψ. The pdf of Ψ is $f(\Psi|\theta)$. In an integrated analysis setting the elements of D may involve different types of observations (e.g. fish age, length, etc.) with different types of distributions (e.g. Normal, Negative Binomial, Multinomial) but these distributions share some common parameters. Hence, we simply collect all the responses into the D vector and do not develop additional notation for the responses which is consistent with the generic notation used in some of the literature (e.g. Skaug and Fournier, 2006; Kristensen et al., 2016; Thorson and Minto, 2015). The marginal distribution of D is

$$f(D|\theta) = \int \cdots \int_q f(D|\Psi, \theta) f(\Psi|\theta) d\Psi_1, \ldots, d\Psi_q, \tag{1}$$

where $\Psi_1, \ldots, \Psi_q$ are the elements of Ψ. For simplicity this q-fold integral is denoted as $\int f(D|\Psi, \theta) f(\Psi|\theta) d\Psi$. There may also be covariate information used in the model, but we do not develop notation for this. The maximum marginal likelihood estimates (MMLE) of θ are those values $\hat{\theta}$ that maximize $f(D|\theta)$. Throughout this paper we use $\hat{}$ to denote estimators. We can "estimate" Ψ with the conditional mean $\int \Psi f(\Psi|\hat{\theta}, D) d\Psi$, and when the joint pdf $f(D, \Psi|\theta) = f(D|\Psi, \theta) f(\Psi|\theta)$ is unimodal and approximately symmetric about Ψ then these estimates are equivalently those values $\hat{\Psi}$ that maximize $f(D, \Psi|\theta)$ when $\theta = \hat{\theta}$.

With ecological models, and in particular fish stock assessment models, the θ parameters are usually not of direct interest and many of the θ's are nuisance parameters. Rather, nonlinear functions of θ, the random effects Ψ, and covariates are of direct interest. We generically denote such a function as $g(\theta, \Psi)$. TMB provides standard errors of MMLE's of θ and also estimates of user-specified $g(\theta, \Psi)$. This is explained further in Section 2. Both SAM and glmmTMB utilize this feature of TMB to calculate standard errors. Kristensen et al. (2016) referred to this as the generalized delta method (GDM). However, the statistical basis for the GDM standard errors is not completely clear. Two references are given in Kristensen et al. (2016), but the first Ref. (i.e. Skaug and Fournier, 2006) does not mention the delta method while the second Ref. (i.e. Kass and Steffey, 1989) directly considered functions of only the random effects, $g(\hat{\Psi})$, and derived standard errors in a Bayesian setting.

In this paper we (1) develop frequentist variance approximations for $\hat{\Psi}$. The frequentist variance refers to the variability of $\hat{\Psi}$ derived from infinitely many data sets randomly drawn from $f(D, \Psi|\theta)$. This will include repeated sampling of Ψ from $f(\Psi|\theta)$ and D from $f(D|\Psi, \theta)$. Using standard asymptotic results for $\mathrm{Cov}(\hat{\theta})$, we (2) provide a GDM approximation of $\mathrm{Cov}\{g(\hat{\theta}, \hat{\Psi})\}$, where $g(\cdot)$ is a vector-valued function. We (3) clarify what is the statistical basis for the TMB GDM variance formula. We illustrate the utility of $\mathrm{Cov}\{g(\hat{\theta}, \hat{\Psi})\}$ and the TMB GDM variance equation using simulations from a simple state–space random-walk model. We develop results using a frequentist approach (i.e. MMLE) to statistical inference for mixed-effects models and we do not use Bayesian concepts. Nevertheless, sometimes an identical result may be reached independently from frequentist and Bayesian outlooks, which we discuss in Section 7. We develop novel theoretical results without normal distributional assumptions for random effects, but we derive more amenable approximations for the normal distributional assumptions commonly found in statistical packages including TMB and ADMB (Kristensen et al., 2016; Fournier et al., 2012).

2. Nonlinear mixed-effects models and TMB: variance of predicted random effects

The Template Model Builder (TMB, Kristensen et al., 2016) package within R (R Core Team, 2018) can be used to implement nonlinear mixed-effects models for random response data which are collected in a $n \times 1$ vector D that are assumed to have a multivariate pdf $f(D|\Psi, \theta)$, given values of the fixed-effects parameters θ $(p \times 1)$ and the random-effects Ψ $(q \times 1)$. The pdf of Ψ is $f(\Psi|\theta)$. The marginal distribution of D is given by Eq. (1). The user only has to code in a TMB C++ template the conditional data loglikelihood $l_c(\theta, \Psi) = \log\{f(D|\Psi, \theta)\}$, and the loglikelihood of the random effects, $l(\theta, \Psi) = \log\{f(\Psi|\theta)\}$. This is usually straight-forward. We denote the joint loglikelihood as

$$l_j(\theta, \Psi) = l_c(\theta, \Psi) + l(\theta, \Psi). \tag{2}$$

TMB calculates the marginal loglikelihood $l(\theta)$, which is the natural logarithm (i.e. log) of $f(D|\theta)$ in Eq. (1), using the Laplace approximation which will be good when $f(D, \Psi|\theta)$ is approximately MVN (multivariate normal) about Ψ. If Ψ has a MVN distribution then this requirement is easier to satisfy. For convenience we assume $\mathrm{E}(\Psi) = 0$, but all the results in this paper are correct when $\mathrm{E}(\Psi) \neq 0$. As part of the Laplace approximation TMB calculates $\hat{\Psi}(\theta)$ that maximizes $l_j(\theta, \Psi)$ for any value of θ; that is,

$$\dot{l}_j(\theta, \Psi)|_{\Psi=\hat{\Psi}(\theta)} = \left.\frac{\partial l_j(\theta, \Psi)}{\partial \Psi}\right|_{\Psi=\hat{\Psi}(\theta)} = 0. \tag{3}$$

TMB uses automatic differentiation to provide the gradient function of $l(\theta)$,

$$\dot{l}(\theta) = \frac{\partial l(\theta)}{\partial \theta}, \tag{4}$$

which is very useful for finding the maximum likelihood estimates of θ.

Let $\Omega = (\Psi', \theta')'$ be the $(p+q) \times 1$ vector of all random and fixed effects, and let the $b \times 1$ vector-valued function of interest be $g(\hat{\Omega})$. If $\dot{g}(\Omega) = \partial g(\Omega)/\partial \Omega'$ is a $b \times (p+q)$ matrix of derivatives then the generalized delta method (GDM) covariances used by TMB are

$$\text{Cov}\{g(\hat{\Omega})\} = \dot{g}(\hat{\Omega})\text{Cov}(\hat{\Omega})\dot{g}'(\hat{\Omega}), \tag{5}$$

where $\dot{g}'(\Omega)$ is the matrix transpose of $\dot{g}(\Omega)$. This is a standard application of the Delta-Method. The challenging part of Eq. (5) is $\text{Cov}(\hat{\Omega})$. The equation TMB uses is

$$\text{Cov}(\hat{\Omega}) = \begin{bmatrix} -\ddot{l}_j^{-1}(\theta, \Psi) & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \dfrac{\partial \hat{\Psi}(\theta)}{\partial \theta'} \\ I \end{bmatrix} \text{Cov}(\hat{\theta}) \begin{bmatrix} \dfrac{\partial \hat{\Psi}'(\theta)}{\partial \theta} & I \end{bmatrix}, \tag{6}$$

where $\ddot{l}_j(\theta, \Psi) = \partial^2 l_j(\theta, \Psi)/\partial \Psi \partial \Psi'$ and I is a $p \times p$ identity matrix. Note that $\partial \hat{\Psi}(\theta)/\partial \theta' = -\ddot{l}_j^{-1}\left\{\theta, \hat{\Psi}(\theta)\right\} \times \partial^2 l_j(\theta, \Psi)/\partial \Psi \partial \theta'|_{\Psi = \hat{\Psi}(\theta)}$ (e.g. see Section 4.1 in Kristensen et al., 2016). From standard asymptotic theory (e.g. Barndorff-Nielsen and Cox, 1994) based on the marginal loglikelihood, $\text{Cov}(\hat{\theta}) \approx -E\{\ddot{l}(\theta)\}^{-1}$ where $\ddot{l}(\theta) = \partial^2 l(\theta)/\partial \theta \partial \theta'$. $\text{Cov}(\hat{\theta})$ can be estimated using $-\ddot{l}^{-1}(\hat{\theta})$ and $\text{Cov}(\hat{\Omega})$ can be estimated by replacing θ and Ψ in Eq. (6) with their estimates $\hat{\theta}$ and $\hat{\Psi}$. The $\text{Cov}(\hat{\Psi})$ sub-matrix in Eq. (6) is basically the same as the covariance equation used in ADMB (Fournier et al., 2012), as described in Equation (3.1) in Skaug and Fournier (2020). ADMB is another software package that can be used to implement nonlinear mixed effects models.

As the sample size and information (i.e. $\ddot{l}(\theta)$) about θ gets large (i.e. $\lambda_{min}\ddot{l}(\theta) \to \infty$ where λ_{min} is the smallest eigenvalue; see Fahrmeir and Kaufmann, 1985) then the diagonal elements of $\text{Cov}(\hat{\theta}) \to 0$ which makes sense. However, if the sample size gets large such that the information about the random effects ($\ddot{l}_j(\theta, \Psi)$) also gets large then the diagonals of the $\text{Cov}(\hat{\Psi})$ sub-matrix in Eq. (6)→ 0. This is puzzling because with repeated sampling of Ψ and D one might naively expect $\text{Cov}(\hat{\Psi}) \to \text{Cov}(\Psi)$ as the sample size, including the information about Ψ, gets large. Hence there is some ambiguity about what the TMB covariance for $\hat{\Psi}$ represents. However, $\text{Cov}(\hat{\Psi})$ is different from $\text{Cov}(\Psi)$, and an estimate of $\text{Cov}(\hat{\Psi})$ is also not an estimate of $\text{Cov}(\Psi)$. The differences among $\text{Cov}(\hat{\Psi})$, $\text{Cov}(\Psi)$ and the TMB covariance equation for $\hat{\Psi}$ are illustrated in two examples in Section 5.

In the next two sections we derive $\text{Cov}(\hat{\Omega})$ approximations and clarify what Eq. (6) represents. We include some approximation orders. Similar to Kass and Steffey (1989) and Flores-Agreda and Cantoni (2019), for this purpose we assume that there are $i = 1, \ldots, T$ observational units and that there are n observations in each observational unit that share the same subset of random effects ($\Psi_i \subset \Psi$). For example, in a time-series setting T may indicate number of years. We make the n sample size assumption within units just to illustrate the accuracy of the approximations, but the exact details will depend on the actual sample sizes in observational units, which may vary from n. Our approximation orders will be conservative in some cases.

3. Cov($\hat{\Psi}$)

We consider that there is a true but unknown value of θ that we denote as θ_o. We use a first-order Taylor's series expansion of $\hat{\Psi}(\hat{\theta})$ about $\hat{\theta} = \theta_o$,

$$\hat{\Psi}(\hat{\theta}) = \hat{\Psi}(\theta_o) + \left.\frac{\partial \hat{\Psi}(\theta)}{\partial \theta'}\right|_{\theta=\theta_o} (\hat{\theta} - \theta_o) + O_p(T^{-1}). \tag{7}$$

We use $O(\cdot)$ and $o(\cdot)$ notations in a matrix sense, such that they apply to each element of $(\cdot)$. Using Eq. (7), we can show that

$$\begin{aligned} \text{Cov}\left\{\hat{\Psi}(\hat{\theta})\right\} &= \text{Cov}\left\{\hat{\Psi}(\theta_o)\right\} + \frac{\partial \hat{\Psi}(\theta_o)}{\partial \theta_o'} \text{Cov}(\hat{\theta}) \frac{\partial \hat{\Psi}'(\theta_o)}{\partial \theta_o} \\ &\quad + \frac{\partial \hat{\Psi}(\theta_o)}{\partial \theta_o'} \text{Cov}\left\{\hat{\theta}, \hat{\Psi}(\theta_o)\right\} + \text{Cov}\left\{\hat{\Psi}(\theta_o), \hat{\theta}\right\} \frac{\partial \hat{\Psi}'(\theta_o)}{\partial \theta_o} \\ &\quad + o(T^{-1}). \end{aligned} \tag{8}$$

Note that we have used the notation $\partial \hat{\Psi}(\theta_o)/\partial \theta_o'$ to denote $\partial \hat{\Psi}(\theta)/\partial \theta'|_{\theta=\theta_o}$. The $o(T^{-1})$ term in Eq. (8) comes from the covariance between $O_p(T^{-1})$ and the first two terms on the right-hand side of Eq. (7). We describe in Appendix A.2 why $\text{Cov}\{\hat{\Psi}(\theta_o), O_p(T^{-1})\}$ is $o(T^{-1})$. It is clear that $\text{Cov}\{\hat{\theta} - \theta_o, O_p(T^{-1})\}$ is $o(T^{-1})$ since $\hat{\theta} - \theta_o$ is $O_p(T^{-1/2})$. In addition, using Eqs. (22) and (32) in the Appendix, $\mathrm{Cov}\{\hat{\theta}, \hat{\Psi}(\theta_o)\} = \mathrm{Cov}\{(\hat{\theta} - \theta_o), \hat{\Psi}(\theta_o)\} = \mathcal{I}^{-1}(\theta_o)\mathrm{Cov}\{\dot{l}(\theta_o), \hat{\Psi}(\theta_o)\} + \mathrm{Cov}\{O_p(T^{-1}), \hat{\Psi}(\theta_o)\} = o(T^{-1})$. Therefore, we ignore $\mathrm{Cov}\{\hat{\theta}, \hat{\Psi}(\theta_o)\}$ in Eq. (8).

Using a similar derivation we can show that

$$\mathrm{Cov}\left\{\hat{\Psi}(\hat{\theta}), \hat{\theta}\right\} = \frac{\partial \hat{\Psi}(\theta_o)}{\partial \theta_o'}\mathrm{Cov}(\hat{\theta}) + o(T^{-1}). \tag{9}$$

We show in Appendix A.1 (see Eq. (26)) that $\mathrm{Cov}\{\hat{\Psi}(\theta_o)\} = \mathrm{Cov}(\Psi) - \mathrm{E}\{\mathrm{Cov}[\Psi|D, \theta_o]\}$. The diagonal elements of $\mathrm{E}\{\mathrm{Cov}[\Psi|D, \theta_o]\}$ will be positive so that $\mathrm{Var}\{\hat{\Psi}(\theta_o)\} < \mathrm{Var}(\Psi)$ which makes sense because $\hat{\Psi}$ will usually be smoother or less variable than Ψ, depending on $f(\Psi|\theta)$. Using these results and Eq. (8), we obtain the (frequentist) sampling variance of $\hat{\Psi}(\hat{\theta})$

$$\mathrm{Cov}\left\{\hat{\Psi}(\hat{\theta})\right\} = \mathrm{Cov}(\Psi) - \mathrm{E}\{\mathrm{Cov}[\Psi|D, \theta_o]\} + \frac{\partial \hat{\Psi}(\theta_o)}{\partial \theta_o'}\mathrm{Cov}(\hat{\theta})\frac{\partial \hat{\Psi}'(\theta_o)}{\partial \theta_o} + o(T^{-1}). \tag{10}$$

The third term in this equation represents the uncertainty due to estimating θ. Eq. (10) is derived without any distributional assumptions for Ψ. If $f(D, \Psi|\theta)$ is approximately MVN about Ψ,

$$\mathrm{Cov}\left\{\hat{\Psi}(\hat{\theta})\right\} \approx \mathrm{Cov}(\Psi) - \mathrm{E}\left\{-\ddot{l}_j^{-1}(\Psi, \theta_o)\right\} + \frac{\partial \hat{\Psi}(\theta_o)}{\partial \theta_o'}\mathrm{Cov}(\hat{\theta})\frac{\partial \hat{\Psi}'(\theta_o)}{\partial \theta_o}. \tag{11}$$

$\mathrm{Cov}\{\hat{\Psi}(\hat{\theta})\}$ can be estimated by replacing $\mathrm{E}\{\ddot{l}_j^{-1}(\Psi, \theta_o)\}$ with $\ddot{l}_j^{-1}(\Psi, \theta_o)$ and replacing Ψ and θ_o with estimates $\hat{\Psi}$ and $\hat{\theta}$. If estimating $\mathrm{E}\{\ddot{l}_j^{-1}(\Psi, \theta_o)\}$ with $\ddot{l}_j^{-1}(\Psi, \theta_o)$ does not bring about large bias, this estimator of $\mathrm{Cov}\{\hat{\Psi}(\hat{\theta})\}$ has an approximation order of $O(n^{-1}T^{-1})$, which comes from replacing θ_o in $\ddot{l}_j^{-1}(\Psi, \theta_o)$ with $\hat{\theta}$. If n is small then the data are less informative about Ψ and the first two terms in Eq. (11) tend to cancel each other, and Eq. (11) can achieve an approximation order of $o(T^{-1})$.

The estimate of Eq. (11) is different than the relevant part of the TMB covariance in (Eq. (6)) which does not include an estimate of $\mathrm{Cov}(\Psi)$ and the sign of the $\ddot{l}_j$ term is different. We can show that

$$\begin{aligned}\mathrm{Cov}\left\{\begin{bmatrix}\hat{\Psi}(\hat{\theta}) - \Psi \\ \hat{\theta}\end{bmatrix}\right\} &= \begin{bmatrix}\mathrm{E}\{\mathrm{Cov}[\Psi|D, \theta_o]\} & 0 \\ 0 & 0\end{bmatrix} \\ &+ \begin{bmatrix}\dfrac{\partial \hat{\Psi}(\theta_o)}{\partial \theta_o'} \\ I\end{bmatrix}\mathrm{Cov}(\hat{\theta})\begin{bmatrix}\dfrac{\partial \hat{\Psi}'(\theta_o)}{\partial \theta_o} & I\end{bmatrix} + O(T^{-3/2}).\end{aligned} \tag{12}$$

Again no distributional assumption is involved in deriving Eq. (12). $\mathrm{Cov}\{\hat{\Psi}(\hat{\theta}) - \Psi\}$ is the covariance of the prediction error of $\hat{\Psi}(\hat{\theta})$ and hence is referred to as the "prediction variance", to distinguish it from the sampling variance given by (10). If $f(D, \Psi|\theta)$ is approximately MVN about Ψ,

$$\begin{aligned}\mathrm{Cov}\left\{\begin{bmatrix}\hat{\Psi}(\hat{\theta}) - \Psi \\ \hat{\theta}\end{bmatrix}\right\} &\approx \begin{bmatrix}\mathrm{E}\left\{-\ddot{l}_j^{-1}(\Psi, \theta_o)\right\} & 0 \\ 0 & 0\end{bmatrix} \\ &+ \begin{bmatrix}\dfrac{\partial \hat{\Psi}(\theta_o)}{\partial \theta_o'} \\ I\end{bmatrix}\mathrm{Cov}(\hat{\theta})\begin{bmatrix}\dfrac{\partial \hat{\Psi}'(\theta_o)}{\partial \theta_o} & I\end{bmatrix}.\end{aligned} \tag{13}$$

The estimate of Eq. (13) is the same as Eq. (6).

Using Eq. (7),

$$\begin{aligned}\mathrm{E}\left\{\hat{\Psi}(\hat{\theta}) - \Psi\right\} &= \mathrm{E}\left\{\hat{\Psi}(\theta_o) - \Psi\right\} + \frac{\partial \hat{\Psi}(\theta_o)}{\partial \theta_o'}\mathcal{I}^{-1}(\theta_o)\mathrm{E}\left\{\dot{l}(\theta_o)\right\} + O(T^{-1}) \\ &= \mathrm{E}\left\{\mathrm{E}\left[\hat{\Psi}(\theta_o) - \Psi|D\right]\right\} + O(T^{-1}) \\ &= O(T^{-1}).\end{aligned} \tag{14}$$

Therefore $\mathrm{Cov}\{\hat{\Psi}(\hat{\theta}) - \Psi\} = \mathrm{E}\{(\hat{\Psi}(\hat{\theta}) - \Psi)^2\} - \mathrm{E}\{\hat{\Psi}(\hat{\theta}) - \Psi\}^2 = \mathrm{MSE}\{\hat{\Psi}(\hat{\theta})\} + O(T^{-2})$, and Eq. (13) is the mean squared error for the predictors of random effects and is commonly used to measure the uncertainty for $\hat{\Psi}$ (e.g. Kackar and Harville, 1984; Datta and Lahiri, 2000; Das et al., 2004; Flores-Agreda and Cantoni, 2019). This has recently been clarified in TMB documentation (see http://kaskr.github.io/adcomp/_book/Appendix.html#theory-underlying-sdreport). The TMB variance represents the variability of the difference between $\hat{\Psi}(\hat{\theta})$ and Ψ which makes sense because as the sample size gets large and the information about θ and Ψ gets large we expect the prediction covariance $\mathrm{Cov}\{\hat{\Psi}(\hat{\theta}) - \Psi\}$ to decrease to zero as suggested by Eq. (6), but we do not expect the sampling covariance, $\mathrm{Cov}\{\hat{\Psi}(\hat{\theta})\}$, to decrease to zero. As the sample size gets large then $\mathrm{Cov}\{\hat{\Psi}(\hat{\theta})\} \to \mathrm{Cov}(\Psi)$. However, the prediction variance (Eq. (13)) does not indicate how $\hat{\Psi}$ will vary for different realizations of D and Ψ, but the sampling variance (Eq. (11)) does.

For conditionally independent hierarchical models, Kass and Steffey (1989) derived the variance for the posterior distribution of a smooth function of Ψ at the estimated parameter values in a Bayesian setting, which agrees with Eqs. (5) and (6). However, the variance of the posterior distribution and the prediction variance are fundamentally two different concepts respectively in Bayes and frequentist settings, and hence a derivation of the latter within a fully frequentist framework as done in this paper, instead of just borrowing from Bayesian conclusions, is necessary. In addition, Kass and Steffey (1989) do not provide an explicit formula for $\mathrm{Cov}\{\hat{\Psi}(\hat{\theta}), \hat{\theta}\}$ as in Eq. (9), and their paper does not provide the full covariance matrix in Eq. (6).

It is also informative to examine Eqs. (11) and (13) when the data are uninformative about some Ψ, which may occur when data are missing or when the model is used to forecast random effects. Let Ψ_s be the subset of Ψ that the data are uninformative about. If the sample size is large enough so that the last terms in Eqs. (11) and (13) are negligible then $\mathrm{Cov}\{\hat{\Psi}_s(\hat{\theta})\}$ will be zero whereas $\mathrm{Cov}\{\hat{\Psi}_s(\hat{\theta}) - \Psi_s\} = \mathrm{Cov}(\Psi_s)$. This makes sense because when the data are uninformative about Ψ_s then $\hat{\Psi}_s(\hat{\theta})$ will always be zero which is why $\mathrm{Cov}\{\hat{\Psi}_s(\hat{\theta})\} = 0$ and $\mathrm{Cov}\{\hat{\Psi}_s(\hat{\theta}) - \Psi_s\} = \mathrm{Cov}(\Psi_s)$.

4. Generalized delta method covariances

The generalized delta method to approximate the covariance matrix of a vector-valued differentiable function $g(\hat{\Omega})$ of fixed- and random-effects (i.e. Ω) estimates was described by Eq. (5). The estimate of $\mathrm{Cov}(\hat{\Omega})$, denoted as lower-case $\mathrm{cov}(\hat{\Omega})$, that is appropriate to use is

$$\mathrm{cov}(\hat{\Omega}) = \begin{bmatrix} \mathrm{cov}(\Psi) + \ddot{l}_j^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \dot{\hat{\Psi}} \\ I \end{bmatrix} \mathrm{cov}(\hat{\theta}) \begin{bmatrix} \dot{\hat{\Psi}}' & I \end{bmatrix}. \tag{15}$$

Note that for simplicity we used the notation $\ddot{l}_j = \partial^2 l_j(\theta, \Psi)/\partial\Psi\partial\Psi'|_{\theta=\hat{\theta},\Psi=\hat{\Psi}}$ and $\dot{\hat{\Psi}} = \partial\hat{\Psi}(\theta)/\partial\theta'|_{\theta=\hat{\theta}}$. The estimate $\mathrm{cov}(\Psi)$ is based on the estimated parameters for the distribution of Ψ, and $\mathrm{cov}(\hat{\theta}) = -\ddot{l}^{-1}(\hat{\theta})$ which is the matrix inverse of the hessian of the negative marginal loglikelihood evaluated at $\hat{\theta}$.

The GDM standard errors that TMB gives, based on $\mathrm{Cov}\{g(\hat{\Omega})\}$ in Eq. (6), are estimates of $\mathrm{Cov}\{g(\hat{\Omega}) - g(\Omega)\}$.

5. Simple examples

5.1. Example 1

In this section we illustrate results using a very simple example. Consider n observation of $Y_i \overset{i.i.d}{\sim} N(\mu, \sigma^2), i = 1, \ldots, n$, which we denote as $y_1, \ldots, y_n$. The interest is to predict the value of a new observation, say Y_{n+1}, based on the observed data $y_1, \ldots, y_n$. Hence, the single random effect in this example is $\Psi = Y_{n+1} \sim N(\mu, \sigma^2)$. The joint loglikelihood (i.e. Eq. (2)) is

$$l_j(\theta, \Psi) = K - (n+1)log(\sigma) - \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2} - \frac{(\Psi - \mu)^2}{2\sigma^2},$$

where K is a collection of constant terms. The marginal loglikelihood is

$$l(\theta) = K - nlog(\sigma) - \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}.$$

Clearly $\hat{\Psi}(\theta) = \mu$. Because the conditional distribution of Ψ given data is normal with variance σ^2, $\mathrm{E}\{\mathrm{Cov}[\Psi|D, \theta]\} = \mathrm{Cov}[\Psi|D, \theta] = -\ddot{l}_j^{-1}(\theta, \Psi) = \sigma^2$. In this case the prediction covariance (Eq. (6)) and the more precise formula (Eq. (12)) are the same. $\mathrm{Cov}(\hat{\theta})$ in Eq. (6) is

$$-\ddot{l}^{-1}(\theta) = \frac{\sigma^2}{n}\begin{bmatrix} 1 & 0 \\ 0 & 2\sigma^2 \end{bmatrix}.$$

Let $\Omega = (\Psi, \mu, \sigma^2)'$. Using (12) and (15) we have

$$\mathrm{Cov}\{\hat{\Omega} - \Omega\} = \frac{\sigma^2}{n}\begin{bmatrix} n+1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2\sigma^2 \end{bmatrix}, \text{ and}$$

$$\mathrm{Cov}\{\hat{\Omega}\} = \frac{\sigma^2}{n}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2\sigma^2 \end{bmatrix}.$$

Since $\hat{\Psi} = \hat{\Psi}(\hat{\theta}) = \hat{\mu} = \bar{y}$, the estimate of the sampling variance of $\hat{\Psi}$ is $\hat{\sigma}^2/n$, agreeing with the estimate using Eq. (15). The prediction variance estimate based on Eq. (6) or (12) is $\hat{\sigma}^2(1 + 1/n)$ which is the standard estimate of the prediction variance of a new observation. However, this is not the variance of $\hat{\Psi}$.

5.2. Example 2

A simple state–space example is a random-walk with multiple observations per time-step. The random walk is $\Psi_t|\Psi_{t-1} \overset{indep}{\sim} N(\Psi_{t-1}, \sigma_\Psi^2)$, for $t = 2, \ldots, T$, and $\Psi_1 \sim N(\beta, \sigma_\Psi^2)$. At each time-step there are n independent observations of the process, $Y_{t,i}|\Psi_t \overset{i.i.d}{\sim} N(\Psi_t, \sigma_\epsilon^2)$, $i = 1, \ldots, n$ and $t = 1, \ldots, T$. The parameters are $\theta = (\beta, \sigma_\Psi, \sigma_\epsilon)'$ and the random effects are $\Psi = (\Psi_1, \ldots, \Psi_T)'$ which is a $T \times 1$ vector. The joint loglikelihood is

$$\begin{aligned} l_j(\theta, \Psi) =& K - nT\log(\sigma_\epsilon) - \frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{T}\sum_{i=1}^{n}(y_{ti} - \Psi_t)^2 \\ & -T\log(\sigma_\Psi) - \frac{1}{2\sigma_\Psi^2}\left\{(\Psi_1 - \beta)^2 - \sum_{t=2}^{T}(\Psi_t - \Psi_{t-1})^2\right\}. \end{aligned} \tag{16}$$

The marginal means and covariances of the random walk are $E(\Psi_t) = \beta$, $\text{Var}(\Psi_t) = t\sigma_\Psi^2$, and $\text{Cov}(\Psi_t, \Psi_{s<t}) = s\sigma_\Psi^2$. $\Psi \sim \text{MVN}(\beta 1_T, \sigma_\Psi^2 M)$ where 1_T is a $T \times 1$ vector of ones and the (i, j)'th element of M is $m_{i,j} = min(i, j)$. If Y is the $nT \times 1$ vector with elements $(Y_{1,1}, \ldots, Y_{1,n}, Y_{2,1}, \ldots, Y_{2,n}, \ldots, Y_{T,n})$, then the marginal distribution of Y is $\text{MVN}(\beta 1, \Sigma)$, where $\Sigma = \sigma_\epsilon^2 \text{I} + \sigma_\Psi^2 M \otimes \text{J}_n$, 1 is a $nT \times 1$ vector of ones, I is an $nT \times nT$ identity matrix, and J_n is a $n \times n$ matrix of ones. The MLE of β is $\hat{\beta} = 1'\Sigma^{-1}Y/1'\Sigma^{-1}1$.

To simplify estimation we assume $\tau = \sigma_\Psi/\sigma_\epsilon$ is known. In this case, the marginal distribution of Y is $\text{MVN}(\beta 1, \sigma_\epsilon^2 A)$ where $A = \text{I} + \tau^2 M \otimes \text{J}_n$, and the MLE of σ_ϵ^2 is $\hat{\sigma}_\epsilon^2 = (y - \hat{\beta})'A^{-1}(y - \hat{\beta})/nT$. Using the joint loglikelihood with $\sigma_\Psi = \sigma_\epsilon \tau$ we can show that the values of $\hat{\Psi}$ that maximize Eq. (16) are the solution to

$$\left\{nI_T + \tau^{-2}(I_T - B)'(I_T - B)\right\}\hat{\Psi} = y_+^*, \tag{17}$$

where B is the backshift operator matrix and y_+ is a $T \times 1$ vector of the sum of the $y_{t,i}$'s for each t, $\sum_{i=1}^n y_{t,i}$. Note that in y_+^* the first element is $\tau^{-2}\hat{\beta} + y_{1,+}$ and all other elements are the same as in y_+. If we denote $C = \left\{nI_T + \tau^{-2}(I_T - B)'(I_T - B)\right\}$ and assume β is estimated with negligible error then $\text{Cov}(\hat{\Psi}) = C^{-1}\Sigma_+ C^{-1}$ where Σ_+ is derived by summing the corresponding elements of $\sigma_\epsilon^2 A$. We can show that $\Sigma_+ = n\sigma_\epsilon^2(I_T + n\tau^2 M)$ and $C = n\left\{I_T + M^{-1}/n\tau^2\right\}$. If $M_* = n\tau^2 M$ then

$$\begin{aligned} \text{Cov}(\hat{\Psi}) =& \frac{\sigma_\epsilon^2}{n}\left\{(I_T + M_*^{-1})^{-1}(I_T + M_*)(I_T + M_*^{-1})^{-1}\right\} \\ =& \frac{\sigma_\epsilon^2}{n}\left\{M_*(I_T + M_*)^{-1}M_*\right\} \\ =& \frac{\sigma_\epsilon^2}{n}\left\{M_* - (I_T + M_*^{-1})\right\} \text{(via the Woodbury matrix identity)} \\ =& \sigma_\Psi^2 M - \sigma_\epsilon^2 C^{-1}. \end{aligned}$$

The first two terms in Eq. (11) produce the same result. If β is fixed at $\hat{\beta}$ then the first term in Eq. (13) gives $\text{Cov}(\hat{\Psi} - \Psi) = \sigma_\epsilon^2 C^{-1}$ which is what TMB reports. As the sample size n for each random effect increases, C^{-1} goes to 0 and hence $\text{Cov}(\hat{\Psi}) \to \sigma_\Psi^2 M = \text{Cov}(\Psi)$, while the prediction covariance, $\text{Cov}(\hat{\Psi} - \Psi)$, goes to 0.

We illustrate these results using a simulated data set. We generated random y responses from the random-walk model with $\beta = 0$, $\sigma_\Psi = 0.5$, $\tau = 2$ so that $\sigma_\epsilon = 0.25$, $n = 5$ and $T = 50$. We fixed $\tau = 2$ when estimating β, σ_ϵ and $\Psi_1, \ldots, \Psi_{50}$. The data, estimates of Ψ_t and confidence intervals (CIs) based on the prediction variance (Eq. (13)) and the sampling variance (Eq. (11)) are shown in Fig. 1. The prediction variance-based CIs cover the real values of Ψ in 47 of the 50 years, or 94% of the years, which is very close the nominal 95% coverage of the CIs. Sampling variance-based CIs are very wide and not useful for inferences about the Ψ values used to generate the data. However, to illustrate the utility of $\text{Var}(\hat{\Psi})$ we also used this with the generalized delta method to derive the standard error (SE) for an estimate of the difference in the means of $\Psi_1, \ldots, \Psi_{25}$ and $\Psi_{26}, \ldots, \Psi_{50}$. The Ψ_t random-walk has mean β for all t; therefore, the difference has mean zero. The estimated difference shown in Fig. 1 (i.e. the difference in the dashed lines) is 0.1791 and the prediction delta-variance SE is 0.033 which is small relative to the estimate and indicates the difference in means is statistically significant. The sampling SE is large (i.e. 2.15) and does not indicate a statistically significant difference in the mean of the Ψ_t's for the two time-periods, which agrees with our theoretical knowledge of the Ψ stochastic process. We generated a second data set in which β was increased by 5 for $t = 26, \ldots, 50$ (see Fig. 2). The estimate of the difference in Ψ means is 5.17, the prediction SE is 0.039, and the sampling SE is 2.54. Again the prediction SE indicates high statistical significance but the sampling SE indicates moderate significance, with a two-sided normal-based p-value of 0.042.

We repeated the simulation process 1000 times. In the first case when β is constant for all t, 97% of the prediction variance-based p-values for the difference in means were less than 0.05 and the average prediction p-value was 0.013. Hence, the prediction SE's do not provide reliable statistical inferences about the mean of the Ψ_t's for the random-walk model. The sampling variance-based p-values were greater than 0.05 in 95% of the simulations which correctly indicates

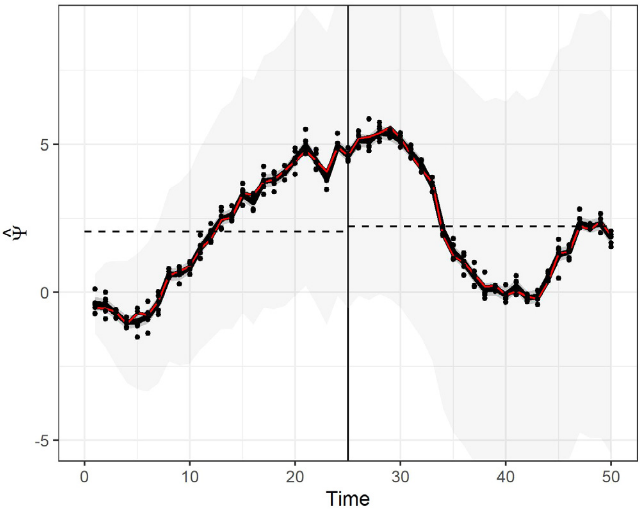

Fig. 1. Simulated y data (points) and estimates of Ψ_t (heavy solid line) for the random-walk example. The red line indicates the true values of Ψ_t used to generate the responses. The dark-gray shaded region indicates 95% confidence intervals based on the prediction variance (Eq. (13)). The light-gray shaded region indicates 95% confidence intervals based on Var($\hat{\Psi}$) (Eq. (11)). The vertical line indicates the mid-point of the time-series. The dashed horizontal line segments indicate the means of $\Psi_1, \ldots, \Psi_{25}$ and $\Psi_{26}, \ldots, \Psi_{50}$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

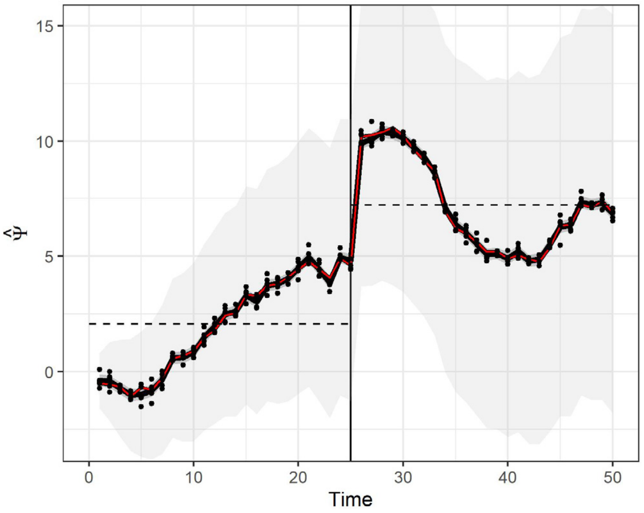

Fig. 2. Simulated y data (points) and estimates of Ψ_t (heavy solid lines) for the random-walk example with a mean shift of Ψ_t at $t = 26$. See Fig. 1 caption for other details.

no difference in means with the correct probability for Type 1 error. In the second case when β was increased by 5 for $t \geq 26$ the prediction p-values were less than 0.05 in all but one of the simulations. The sampling p-values were less than 0.05 in 58% of the simulations which indicates somewhat low statistical power to detect this magnitude of a departure in the random walk mean. We conducted two other sets of simulations with the β increased by 7.5 and 10 and the sampling p-values were less than 0.05 in 84% and 97% of the simulations, respectively. This gives a better indication of the types of mean shifts that can be reliably determined based on the generalized delta standard errors and Var($\hat{\Psi}$).

The simulations results have demonstrated that the prediction SE's for $\hat{\Psi}$ are useful for statistical inferences about the specific Ψ values associated with the data. In a sense they are conditioned on the data. We consider this further in the Discussion. SE's for $\hat{\Psi}$ based on sampling variance Var($\hat{\Psi}$) are not useful for inferences about the specific Ψ's associated with the data. However, the prediction SE's do not measure the variability in $\hat{\Psi}$ or functions of $\hat{\Psi}$ that will occur with repeated sampling, whereas sampling SE's for $\hat{\Psi}$ based on Var($\hat{\Psi}$) do. This makes the latter SE's more useful for statistical inferences about the distribution of Ψ, such as the mean or the difference in the mean for two time periods. Generalized delta standard errors for $g(\hat{\theta}, \hat{\Psi})$ based on the prediction covariance are appropriate for the particular data set and values of Ψ whereas the standard errors for $g(\hat{\theta}, \hat{\Psi})$ based on the sampling covariance $\mathrm{Var}(\hat{\Psi})$ will be appropriate to describe variability due to random sampling of the data D and random effects Ψ.

6. Comparison with linear mixed models literature

In this section we compare our variance results with those from some of the published literature on linear mixed effects models, primarily (Das et al., 2004). We use the same notation as Das et al. (2004) except that Ψ is used to denote random effects. The linear mixed model is

$$y = X\beta + Z\Psi + e,$$

where y is an $n \times 1$ vector of sample observations, X and Z are known matrices, β is a $p \times 1$ vector of unknown parameters (fixed effects) and Ψ and e are distributed independently with means 0 and covariance matrices G and R, respectively, depending on some unknown vector of parameters σ. We assume that p is fixed and X is of full rank $p < n$. Note that $\mathrm{Cov}(y) = \Sigma = R + ZGZ'$. For simplicity, we assume that all the distributions involved are MVN. Maximum likelihood estimation of β with empirical posterior means for random effects gives

$$\begin{aligned} &\hat{\beta} = \left(X' \Sigma^{-1} X\right)^{-1} X' \Sigma^{-1} y \text{ and} \\ &\hat{\Psi} = \left(Z' R^{-1} Z + G^{-1}\right)^{-1} Z' R^{-1} \left(y - X\hat{\beta}\right). \end{aligned}$$

Note that $\left(Z' R^{-1} Z + G^{-1}\right) GZ' \Sigma^{-1} = \left(Z'R^{-1}ZGZ' + Z'\right) \Sigma^{-1}$ and using $ZGZ' = \Sigma - R$ then $\left(Z'R^{-1}ZGZ' + Z'\right) \Sigma^{-1} = Z'R^{-1}$. Hence

$$\left(Z' R^{-1} Z + G^{-1}\right)^{-1} Z' R^{-1} = GZ' \Sigma^{-1},$$

and

$$\begin{aligned} &\hat{\beta} = \left(X' \Sigma^{-1} X\right)^{-1} X' \Sigma^{-1} y, \\ &\hat{\Psi} = GZ' \Sigma^{-1} \left(y - X\hat{\beta}\right). \end{aligned} \tag{18}$$

Eq. (18) is the same as the best linear unbiased estimator of β and the predictor of random effects given after Equation (1.3) in Das et al. (2004).

Das et al. (2004) and others (e.g. Datta and Lahiri, 2000; Prasad and Rao, 1990; Kackar and Harville, 1984) focused on the MSE of the estimator of a linear combination of $\hat{\beta}$ and $\hat{\Psi}$,

$$t(\sigma) = l'\hat{\beta} + m'\hat{\Psi}.$$

The asymptotic properties including the MSE for the MLE of $\hat{\beta}$ are standard, and both our results agree. Therefore, we only focus on whether our results for the MSE of $\hat{\Psi}$ agree with the literature. Hence, we assume $l = 0$ and $m = 1$ in the $t(\sigma)$ equation. The primary conclusion of Das et al. (2004) for ANOVA models and Datta and Lahiri (2000) for mixed-effects small area models is that

$$\mathrm{MSE}\left\{t(\sigma)\right\} \approx g_1(\sigma) + g_2(\sigma) + g_3(\sigma), \tag{19}$$

where

$$\begin{aligned} &g_1(\sigma) = G - GZ' \Sigma^{-1} ZG, \\ &g_2(\sigma) = GZ' \Sigma^{-1} X \left(X' \Sigma^{-1} X\right)^{-1} X' \Sigma^{-1} ZG \text{ and} \\ &g_3(\sigma) = \mathrm{tr}[L\, \Sigma\, L' Var(\hat{\sigma})] \text{ for } L = \partial(\Sigma^{-1} ZG)/\partial\sigma. \end{aligned}$$

In this case

$$-\ddot{j}^{-1}(\theta, \Psi) = \left(Z' R^{-1} Z + G^{-1}\right)^{-1}$$

in Eq. (13), where $\theta = (\beta', \sigma')'$. Also,

$$\mathrm{Cov}(\hat{\theta}) = \begin{bmatrix} (X'\Sigma^{-1}X)^{-1} & 0 \\ 0 & \mathrm{Var}(\hat{\sigma}) \end{bmatrix},$$

and

$$\frac{\partial \hat{\Psi}'}{\partial \theta} = \begin{bmatrix} -X' \Sigma^{-1} ZG \\ L(y - X\beta) \end{bmatrix}.$$

Using these results in Eq. (13), after some algebra, the MSE is given by

$$\mathrm{MSE}(\hat{\Psi}) \approx \left(Z' R^{-1} Z + G^{-1}\right)^{-1} + g_2(\sigma) + (y - X\beta)' L' Var(\hat{\sigma}) L (y - X\beta). \tag{20}$$

Note that

$$\begin{aligned}\left(Z' R^{-1}Z + G^{-1}\right) g_1 &= Z' R^{-1}ZG + I - Z' R^{-1}ZGZ' \Sigma^{-1}ZG - Z' \Sigma^{-1}ZG \\ &= Z' R^{-1}ZG + I - Z' R^{-1}(ZGZ' + R - R)\, \Sigma^{-1}ZG - Z' \Sigma^{-1}ZG \\ &= Z' R^{-1}ZG + I - Z' R^{-1}(\Sigma - R)\, \Sigma^{-1}ZG - Z' \Sigma^{-1}ZG \\ &= I,\end{aligned}$$

so that the first term in (20), $\left(Z' R^{-1}Z + G^{-1}\right)^{-1} = g_1(\sigma)$. The third term in (20) is

$$\begin{aligned}\mathrm{E}\{(y - X\beta)' L' Var(\hat{\sigma})L(y - X\beta)\} &= \mathrm{E}\{y - X\beta\}' L' Var(\hat{\sigma})L\, \mathrm{E}\{y - X\beta\} \\ &\quad + \mathrm{tr}[L' Var(\hat{\sigma})L\, Var(y - X\beta)] \\ &= \mathrm{tr}[L' Var(\hat{\sigma})L\, \Sigma] = \mathrm{tr}[L\, \Sigma\, L' Var(\hat{\sigma})] \\ &= g_3.\end{aligned}$$

Therefore, Eq. (20) based on our results agrees with Eq. (19) based on Das et al. (2004) and Datta and Lahiri (2000).

Let T and n be similarly defined as at the end of Section 2. Eq. (20) is correct to the order $o(T^{-1})$. Das et al. (2004), Datta and Lahiri (2000) and others further derived estimators of MSE to the order $o(T^{-1})$. For TMB,

$$\begin{aligned}\mathrm{MSE}(\hat{\Psi})(\sigma) &\approx \mathrm{MSE}(\hat{\Psi})(\hat{\sigma}) + \frac{\partial \mathrm{MSE}(\hat{\Psi})(\hat{\sigma})}{\partial \hat{\sigma}'}(\sigma - \hat{\sigma}) \\ &\approx \mathrm{MSE}(\hat{\Psi})(\hat{\sigma}) + \frac{\partial \mathrm{MSE}(\hat{\Psi})(\hat{\sigma})}{\partial \hat{\sigma}'} \frac{\partial^2 l(\sigma)}{\partial \sigma \partial \sigma'} \frac{\partial l(\sigma)}{\partial \sigma} \\ &= \mathrm{MSE}(\hat{\Psi})(\hat{\sigma}).\end{aligned}$$

Here the $\partial l(\sigma)/\partial \sigma$ term can be ignored because $\mathrm{E}\{\partial l(\sigma)/\partial \sigma\} = 0$. The bias of this estimator is $O(n^{-1}T^{-1})$, which is the same as Das et al. (2004), Datta and Lahiri (2000) and others, except for additional higher order expansion terms in their $o(T^{-1})$ approximations. $\mathrm{MSE}(\hat{\Psi})(\hat{\sigma})$ is $O(n^{-1})$. If $n \ll T$, $\mathrm{MSE}(\hat{\Psi})(\hat{\sigma})$ dominates and the bias $O(n^{-1}T^{-1})$ can be neglected. If n and T are of the same magnitude or $n \gg T$, $\mathrm{MSE}(\hat{\Psi})(\hat{\sigma})$ has an approximation order of $o(T^{-1})$. Therefore this estimator is appropriate.

7. Discussion

We developed frequentist variance approximations for predictions of random effects (i.e. Ψ) and functions of Ψ and fixed effects (i.e. θ) in nonlinear mixed effects models. We focus on maximum likelihood estimators of θ (i.e. $\hat{\theta}$) and the conditional mean predictor of Ψ given data (i.e. $\hat{\Psi}$). Our purpose was to better understand the generalized delta method (GDM) variance equation used by TMB to provide standard errors. We showed that the TMB variance equation is an approximation of the mean squared error (MSE) between $\hat{\Psi}$ and Ψ which is commonly used approach to measure the variability of predictors of random effects (e.g. Kackar and Harville, 1984; Datta and Lahiri, 2000; Das et al., 2004; Flores-Agreda and Cantoni, 2019). We referred to this as the prediction variance. The TMB variance is not an approximation of the variance of $\hat{\Psi}$ that occurs because of random sampling of Ψ and the data conditional on Ψ. We derived an equation that is appropriate for that variance.

We highlighted that when data are missing for some Ψ_s subset of Ψ, the prediction variance for $\hat{\Psi}_s$ increases to $\mathrm{Var}(\Psi_s)$. This does not describe what happens with the variability of $\hat{\Psi}_s$, which will go to zero as the information about Ψ_s decreases because of the missing data. When this happens we know that $\hat{\Psi}_s$ will be approximately zero or whatever is the estimated mean of Ψ_s if this was non-zero in the model, and $\hat{\Psi}_s$ will have little variability. However, this does not mean that $\hat{\Psi}_s$ is a good estimate of Ψ_s when data are missing; it just means that with repeated sampling that includes missing data, we will get about the same value for $\hat{\Psi}_s$ in every sample.

Which variance formula should be used? In a sense the prediction variance (i.e. TMB variance) is conditional on the observed data. This is described in more detail by Flores-Agreda and Cantoni (2019). What we mean is that in a frequentist sense, the prediction variance is based on the idea that each randomly sampled data set has a different set of random effects associated with it, and inferences about the random effects based on the prediction variance will be appropriate for the specific values of the random effects for the observed data. If the objective of the statistical modeling involves the specific but unobserved values of the random effects associated with the data then the prediction variance is appropriate to use. The variance formulae we developed accurately describes the variability of $\hat{\Psi}$ caused by repeated sampling of Ψ and the data conditional on Ψ. If the objective of the modeling is to make inferences about the stochastic process that generates the Ψ's then the repeated sampling variance equation we developed is more appropriate to use. A practical example of inference about the stochastic process is to detect regime shifts in fisheries stock assessment. If the null hypothesis is "no regime shift", Example 2 in Section 5.2 suggests that the prediction variance can lead to substantial Type I error, while $\mathrm{Cov}(\hat{\Psi})$ does not. Another application of $\mathrm{Cov}(\hat{\Psi})$ is to identify outliers among the random effect predictions.

The MSE of a parameter estimator is equal to its variance plus its squared bias (e.g. Section 7.3.1 of Casella and Berger, 2002), and hence the MSE is larger than the variance. However, for random effects this may not be true. If Ψ

is simply a scalar variable with mean 0, $\mathrm{E}\{(\hat{\Psi}-\Psi)^2\}=\mathrm{Var}\{\hat{\Psi}\}+\mathrm{E}\{(\mathrm{E}[\hat{\Psi}]-\Psi)^2\}-2\mathrm{E}\{\hat{\Psi}\Psi\}$, and $\hat{\Psi}$ and Ψ are usually positively correlated, making the third term negate the second positive term. Alternatively, $\mathrm{Cov}\{\hat{\Psi}\}=\mathrm{Cov}\{\hat{\Psi}-\Psi+\Psi\}=\mathrm{Cov}\{\hat{\Psi}-\Psi\}+\mathrm{Cov}\{\hat{\Psi},\Psi\}+\mathrm{Cov}\{\Psi,\hat{\Psi}\}-\mathrm{Cov}\{\Psi\}$. If the data are highly informative about the parameters and random effects, $\mathrm{Cov}\{\Psi,\hat{\Psi}\}$ will be close to $\mathrm{Cov}\{\Psi\}$ and thus $\mathrm{Cov}\{\hat{\Psi}\}>\mathrm{Cov}\{\hat{\Psi}-\Psi\}$ with $\mathrm{Cov}\{\hat{\Psi}-\Psi\}$ being approximately the MSE as previously explained. This agrees with the observations based on (11) and (13).

The variance of the prediction error of random effects can be derived as $\mathrm{Var}\{\hat{\Psi}(\hat{\theta})-\Psi\}=\mathrm{E}\{\mathrm{Var}[\hat{\Psi}(\hat{\theta})-\Psi|D]\}+\mathrm{Var}\{\mathrm{E}[\hat{\Psi}(\hat{\theta})-\Psi|D]\}=\mathrm{E}\{\mathrm{Var}[\Psi|D]\}+\mathrm{Var}\{\hat{\Psi}(\hat{\theta})-\hat{\Psi}(\theta_0)\}\approx\mathrm{E}\{\mathrm{Var}[\Psi|D]\}+(\partial\hat{\Psi}(\theta_o)/\partial\theta_o')\mathrm{Cov}\{\hat{\theta}\}(\partial\hat{\Psi}'(\theta_o)/\partial\theta_o)$, which gives the relevant sub-matrix in Eq. (12). A further approximation is $\mathrm{Var}\{\hat{\Psi}(\hat{\theta})-\Psi\}\approx\mathrm{Var}[\Psi|D,\hat{\theta}]+(\partial\hat{\Psi}(\hat{\theta})/\partial\hat{\theta}')$ $\mathrm{Cov}\{\hat{\theta}\}(\partial\hat{\Psi}'(\hat{\theta})/\partial\hat{\theta})$, which is the same as the parametric empirical Bayes formula for the posterior variance of random effects in Equation (3.8) of Kass and Steffey (1989) when assuming a uniform prior on θ. This explains why TMB and ADMB can borrow the empirical Bayes formula to estimate the prediction error of random effects in a frequentist framework (Kristensen et al., 2016; Fournier et al., 2012). Nevertheless, the variance of prediction error in a frequentist framework and the variance of posterior distribution in a Bayesian framework are fundamentally two different concepts, and hence there inevitably are some differences in their respective evaluations. For example, when the approximation $\mathrm{E}\{\mathrm{Var}[\Psi|D]\}\approx\mathrm{Var}[\Psi|D]$ does not hold, Eq. (12) is not the same as the empirical Bayes formula (3.8) in Kass and Steffey (1989). In addition, the TMB MSE equation further requires $\mathrm{Cov}[\Psi|D,\theta_o]\approx-\ddot{l}_j^{-1}(\Psi,\theta_o)$, namely, the multivariate normal assumption of $f(D,\Psi|\theta)$ about Ψ. Practical implementation of Eqs. (10) and (12) when these approximations are not valid deserves further investigation. In this regard, Monte-Carlo and bootstrap methods can be effective. For this topic we refer to Flores-Agreda and Cantoni (2019) who provide a good review of the estimators and bootstrap schemes for the MSE of random effects prediction, and proposed for the entire class of generalized linear mixed model an implementation of the random weighted Laplace bootstrap which exhibited good performance in their simulation study.

Finally, we note that when applying the generalized delta method (Section 4), non-linearity of $g(\hat{\Omega})$ can introduce bias. In this case we recommend the epsilon-method for bias correction proposed by Thorson and Kristensen (2016).

Acknowledgments

We are particularly thankful to the anonymous referees and the Associate Editor for their comments which helped improve the quality of this paper.

Funding

Research funding to NC was provided by the Ocean Choice International Industry Research Chair program at the Marine Institute of Memorial University of Newfoundland. Research funding to NC and NZ was provided by the Ocean Frontier Institute, through an award from the Canada First Research Excellence Fund.

Appendix

We first summarize the asymptotic properties of $\hat{\theta}$ which is used to derive $\mathrm{Cov}\{\hat{\theta},\hat{\Psi}(\theta_o)\}$. Since $\dot{l}(\hat{\theta})=0$, a first-order Taylor's series expansion of $\dot{l}(\hat{\theta})$ at θ_o gives

$$\hat{\theta}-\theta_o\approx-\ddot{l}^{-1}(\theta_o)\dot{l}(\theta_o).$$

$\dot{l}(\theta_o)$ is commonly referred to in statistics as a score function. For any score function, under certain regularity conditions on the density function of the associated random variables, one can show that

$$\mathrm{E}\{\dot{l}(\theta_o)|\theta_o\}=0 \text{ and } \mathrm{Var}\{\dot{l}(\theta_o)|\theta_o\}=-\mathrm{E}\{\ddot{l}(\theta_o)|\theta_o\}=\mathcal{I}(\theta_o), \tag{21}$$

where $\mathcal{I}(\theta_o)$ is the Fisher information matrix for θ_o. Eq. (21) is involved when showing that $\hat{\theta}-\theta_o\sim N\left\{0,\mathcal{I}^{-1}(\theta_o)\right\}$. For large sample sizes $-\ddot{l}(\theta_o)\approx\mathcal{I}(\theta_o)$. If $f(D|\theta_o)$ is a normal distribution then $-\ddot{l}(\theta_o)$ is exactly equal to $\mathcal{I}(\theta_o)$. Hence, the approximation we use is

$$\hat{\theta}-\theta_o\approx\mathcal{I}^{-1}(\theta_o)\dot{l}(\theta_o). \tag{22}$$

A.1. $\mathrm{Cov}\{\hat{\Psi}(\theta_o)\}$

$$\begin{aligned}\mathrm{Cov}\left\{\hat{\Psi}(\theta_o)-\Psi\right\}&=\mathrm{Cov}\left\{\hat{\Psi}(\theta_o)\right\}+\mathrm{Cov}\left\{\Psi\right\}\\&\quad-\mathrm{Cov}\left\{\hat{\Psi}(\theta_o),\Psi\right\}-\mathrm{Cov}\left\{\Psi,\hat{\Psi}(\theta_o)\right\}.\end{aligned} \tag{23}$$

$$\begin{aligned}\mathrm{Cov}\left\{\hat{\Psi}(\theta_o)-\Psi\right\}&=\mathrm{E}\left\{\mathrm{Cov}\left[\hat{\Psi}(\theta_o)-\Psi|D\right]\right\}+\mathrm{Cov}\left\{\mathrm{E}\left[\hat{\Psi}(\theta_o)-\Psi|D\right]\right\}\\&=\mathrm{E}\left\{\mathrm{Cov}\left[\Psi|D\right]\right\}\end{aligned} \tag{24}$$

because conditional on D $\hat{\Psi}(\theta_o)$ is a constant, and $\hat{\Psi}(\theta_o) = \mathrm{E}[\Psi|D]$. Similarly,

$$\begin{aligned}\mathrm{Cov}\left\{\hat{\Psi}(\theta_o), \Psi\right\} &= \mathrm{E}\left\{\mathrm{Cov}\left[\hat{\Psi}(\theta_o), \Psi|D\right]\right\} + \mathrm{Cov}\left\{\mathrm{E}\left[\hat{\Psi}(\theta_o)|D\right], \mathrm{E}\left[\Psi|D\right]\right\}\\ &= \mathrm{Cov}\left\{\hat{\Psi}(\theta_o)\right\}.\end{aligned}\tag{25}$$

Substituting (24) and (25) into (23) and rearranging terms,

$$\mathrm{Cov}\left\{\hat{\Psi}(\theta_o)\right\} = \mathrm{Cov}(\Psi) - \mathrm{E}\left\{\mathrm{Cov}[\Psi|D, \theta_o]\right\}.\tag{26}$$

Note that the diagonals of $\mathrm{E}\{\mathrm{Cov}[\Psi|D, \theta_o]\}$ will be non-negative in which case $\mathrm{Var}\{\hat{\Psi}(\theta_o)\} \le \mathrm{Var}(\Psi)$ which makes sense because we expect $\hat{\Psi}$ to be smoother or less variable than Ψ.

A.2. Approximation order of $\mathrm{Cov}\{\hat{\Psi}(\theta_o), O_p(T^{-1})\}$ *in* (8)

Note that $O_p(T^{-1})$ in (8) is a function of data only.

$$\begin{aligned}\mathrm{Cov}\left\{\Psi, O_p(T^{-1})\right\} &= \mathrm{Cov}\left\{\mathrm{E}\left[\Psi|D\right], \mathrm{E}\left[O_p(T^{-1})|D\right]\right\}\\ &\quad + \mathrm{E}\left\{\mathrm{Cov}\left[\Psi, O_p(T^{-1})|D\right]\right\}\\ &= \mathrm{Cov}\{\hat{\Psi}(\theta_o), O_p(T^{-1})\}.\end{aligned}\tag{27}$$

$$\begin{aligned}\mathrm{Cov}\left\{\Psi, O_p(T^{-1})\right\} &= \mathrm{E}\left\{\Psi\, O_p(T^{-1})\right\} - \mathrm{E}\left\{\Psi\right\}\mathrm{E}\left\{O_p(T^{-1})\right\}\\ &= \mathrm{E}\left\{\Psi\, O_p(T^{-1})\right\},\end{aligned}\tag{28}$$

where without loss of generality we assume $\mathrm{E}\{\Psi\} = 0$. $\{\Psi\, O_p(T^{-1})\}$ is $O_p(T^{-1})$, and hence $\mathrm{E}\{\left|\Psi\, O_p(T^{-1})\right|\}$ is $O(T^{-1})$ if $\{\Psi\, O_p(T^{-1})\}$ is uniformly integrable, which we assume true. Because $\mathrm{E}\{\Psi\} = 0$, it is reasonable that $\mathrm{E}\{\Psi\, O_p(T^{-1})\}$ is smaller than $O(T^{-1})$, namely, is $o(T^{-1})$. This can be further seen as follows. Let Ψ_i be the elements of Ψ for the i'th unit. $O_p(T^{-1})$ term is mainly $(\hat{\theta} - \theta_o)^2$. $\hat{\theta}$ is based on the data in all the T units, and hence as T increases $(\hat{\theta} - \theta_o)^2$ becomes less correlated with Ψ_i. As a result,

$$\frac{\mathrm{E}\left\{\Psi_i\, O_p(T^{-1})\right\}}{\mathrm{E}\left\{O_p(T^{-1})\right\}} \xrightarrow{T\to\infty} \frac{\mathrm{E}\left\{\Psi_i\right\}\mathrm{E}\left\{O_p(T^{-1})\right\}}{\mathrm{E}\left\{O_p(T^{-1})\right\}} = 0,$$

which implies that $\mathrm{E}\{\Psi_i\, O_p(T^{-1})\}$ is $o(T^{-1})$ because $\mathrm{E}\{O_p(T^{-1})\}$ is $O(T^{-1})$. Applying this result to (27) and (28) we have that $\mathrm{Cov}\{\hat{\Psi}(\theta_o), O_p(T^{-1})\}$ in (8) is $o(T^{-1})$.

A.3. Proof that $\mathcal{I}^{-1}(\theta_o)\mathrm{Cov}\{\dot{l}(\theta_o), \hat{\Psi}(\theta_o)\} = O(T^{-3/2})$

With a similar procedure as in A.2 we can prove

$$\mathrm{Cov}\left\{\dot{l}(\theta_o), \hat{\Psi}(\theta_o)\right\} = \mathrm{E}\left\{\dot{l}(\theta_o)\Psi\right\}.\tag{29}$$

For simplicity we write $\dot{l}(\theta_o)$ and Ψ as scalars, but actually we are considering each element of matrix $\dot{l}(\theta_o)\Psi'$.

$$\begin{aligned}\mathrm{E}\left\{\dot{l}(\theta_o)\Psi\right\} &= \int \dot{l}(\theta_o)\Psi f(D|\Psi, \theta_o) f(\Psi|\theta_o) d\Psi dD\\ &= \int \frac{df(D|\theta_o)/d\theta_o}{f(D|\theta_o)} \int \Psi f(D|\Psi, \theta_o) f(\Psi|\theta_o) d\Psi dD\\ &= \int \frac{df(D|\theta_o)}{d\theta_o}\hat{\Psi}(\theta_o) dD = -\int \frac{d\hat{\Psi}(\theta_o)}{d\theta_o} f(D|\theta_o) dD\\ &= -\mathrm{E}\left\{\frac{d\hat{\Psi}(\theta_o)}{d\theta_o}\right\}.\end{aligned}\tag{30}$$

(30) indicates that $\mathrm{Cov}\{\dot{l}(\theta_o), \hat{\Psi}(\theta_o)\} = \mathrm{E}\{\dot{l}(\theta_o)\Psi\}$ is $O(1)$. Therefore,

$$\begin{aligned}\mathrm{Cov}\left\{\dot{l}(\theta_o), \hat{\Psi}(\theta_o)\right\} &= \mathrm{E}\left\{\dot{l}(\theta_o)\Psi\right\} = \mathrm{E}\left\{\dot{l}(\hat{\theta})\Psi\right\} + O(T^{-1/2})\\ &= O(T^{-1/2})\end{aligned}\tag{31}$$

since $\dot{l}(\hat{\theta}) \equiv 0$. Because $\mathcal{I}^{-1}(\theta_o)$ is $O(T^{-1})$, we proved

$$\mathcal{I}^{-1}(\theta_o)\,\mathrm{Cov}\{\dot{l}(\theta_o), \hat{\Psi}(\theta_o)\} = O(T^{-3/2}).\tag{32}$$

References

Aanes, S., Engen, S., Sæther, B.-E., Aanes, R., 2007. Estimation of the parameters of fish stock dynamics from catch-at-age data and indices of abundance: can natural and fishing mortality be separated?. Can. J. Fish. Aquat. Sci. 64 (8), 1130–1142.

Aeberhard, W.H., Mills Flemming, J., Nielsen, A., 2018. Review of state-space models for fisheries science. Annu. Rev. Stat. Appl. 5, 215–235.

Auger-Méthé, M., Newman, K., Cole, D., Empacher, F., Gryba, R., King, A.A., Leos-Barajas, V., Flemming, J.M., Nielsen, A., Petris, G., et al., 2020. An introduction to state-space modeling of ecological time series. arXiv preprint arXiv:2002.02001.

Barndorff-Nielsen, O.E., Cox, D.R., 1994. Inference and Asymptotics. Chapman & Hall.

Berg, C.W., Nielsen, A., 2016. Accounting for correlated observations in an age-based state-space stock assessment model. ICES J. Mar. Sci. 73 (7), 1788–1797.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.-S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24 (3), 127–135.

Brooks, M.E., Kristensen, K., Darrigo, M.R., Rubim, P., Uriarte, M., Bruna, E., Bolker, B.M., 2019. Statistical modeling of patterns in annual reproductive rates. Ecology 100 (7), e02706.

Brooks, M.E., Kristensen, K., van Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Maechler, M., Bolker, B.M., 2017. Glmmtmb balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. R J. 9 (2), 378–400, URL: https://journal.r-project.org/archive/2017/RJ-2017-066/index.html.

Cadigan, N.G., 2015. A state-space stock assessment model for northern cod, including under-reported catches and variable natural mortality rates. Can. J. Fish. Aquat. Sci. 73 (2), 296–308.

Casella, G., Berger, R.L., 2002. Statistical Inference, Vol. 2. Duxbury Pacific Grove, CA.

Das, K., Jiang, J., Rao, J., 2004. Mean squared error of empirical predictor. Ann. Statist. 32 (2), 818–840.

Datta, G.S., Lahiri, P., 2000. A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Statist. Sinica 10, 613–627.

Durbin, J., 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 Stat. Methodol. 62 (1), 3–56.

Fahrmeir, L., Kaufmann, H., 1985. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Ann. Statist. 13 (1), 342–368.

Flores-Agreda, D., Cantoni, E., 2019. Bootstrap estimation of uncertainty in prediction for generalized linear mixed models. Comput. Statist. Data Anal. 130, 1–17.

Fournier, D.A., Skaug, H.J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M.N., Nielsen, A., Sibert, J., 2012. Ad model builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optim. Methods Softw. 27 (2), 233–249.

ICES, 2019a. Arctic Fisheries Working Group (AFWG). Techreport 1:30, ICES Scientific Reports, p. 934. http://dx.doi.org/10.17895/ices.pub.5292.

ICES, 2019b. North Western Working Group (NWWG). Techreport 1:14, ICES Scientific Reports, p. 830. http://dx.doi.org/10.17895/ices.pub.5298.

Kackar, R.N., Harville, D.A., 1984. Approximations for standard errors of estimators of fixed and random effects in mixed linear models. J. Amer. Statist. Assoc. 79 (388), 853–862.

Kantas, N., Doucet, A., Singh, S.S., Maciejowski, J., Chopin, N., et al., 2015. On particle methods for parameter estimation in state-space models. Statist. Sci. 30 (3), 328–351.

Kass, R.E., Steffey, D., 1989. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). J. Amer. Statist. Assoc. 84 (407), 717–726.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H.J., Bell, B., 2016. Tmb: Automatic differentiation and laplace approximation. J. Stat. Softw. 70 (5), 1–21.

McCulloch, C.E., 2003. Generalized Linear Mixed Models. In: NSF-CBMS Regional Conference Series in Probability and Statistics, JSTOR, pp. i–84.

Nielsen, A., Berg, C.W., 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fish. Res. 158, 96–101.

Pedersen, M.W., Berg, C.W., Thygesen, U.H., Nielsen, A., Madsen, H., 2011. Estimation methods for nonlinear state-space models in ecology. Ecol. Model. 222 (8), 1394–1400.

Perreault, A.M., Wheeland, L.J., Morgan, M.J., Cadigan, N.G., 2020. A state-space stock assessment model for American plaice on the grand bank of newfoundland. J. Northwest Atl. Fish. Sci. 51, 45–104.

Petersson, L.K., Milberg, P., Bergstedt, J., Dahlgren, J., Felton, A.M., Götmark, F., Salk, C., Löf, M., 2019. Changing land use and increasing abundance of deer cause natural regeneration failure of oaks: Six decades of landscape-scale evidence. Forest Ecol. Manag. 444, 299–307.

Prasad, N.N., Rao, J.N., 1990. The estimation of the mean squared error of small-area estimators. J. Amer. Statist. Assoc. 85 (409), 163–171.

Punt, A.E., Dunn, A., Elvarsson, B.Þ., Hampton, J., Hoyle, S.D., Maunder, M.N., Methot, R.D., Nielsen, A., 2020. Essential features of the next-generation integrated fisheries stock assessment package: A perspective. Fish. Res. 229, 105617.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL: https://www.R-project.org/.

Schnute, J.T., 1994. A general framework for developing sequential fisheries models. Can. J. Fish. Aquat. Sci. 51 (8), 1676–1688.

Skaug, H.J., Fournier, D.A., 2006. Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. Comput. Statist. Data Anal. 51 (2), 699–709.

Skaug, H.J., Fournier, D.A., 2020. Random Effects in AD Model Builder. ADMB-RE User Guide. Version 12.1 ed., ADMB Foundation, Honolulu.

Thorson, J.T., Kristensen, K., 2016. Implementing a generic method for bias correction in statistical models using random effects, with spatial and population dynamics examples. Fish. Res. 175, 66–74.

Thorson, J.T., Minto, C., 2015. Mixed effects: a unifying framework for statistical modelling in fisheries biology. ICES J. Mar. Sci. 72 (5), 1245–1256.

Tuerlinckx, F., Rijmen, F., Verbeke, G., De Boeck, P., 2006. Statistical inference in generalized linear mixed models: A review. Br. J. Math. Stat. Psychol. 59 (2), 225–255.