Sensitivity to Model Misspecification of the Von Bertanlanffy Growth Model with Measurement Error in Age

by

© **Rajib Dey**

A thesis submitted to the School of Gradate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Department of Mathematics and Statistics

Memorial University

September 2017

St. John's, Newfoundland and Labrador, Canada

Abstract

The Von Bertalanffy growth function (VonB) specifies the length of a fish as a function of its age. However, in practice, age is measured with error. We study the structural errors-in-variables (SEV) approach to account for measurement error (ME) in age. Cope and Punt (2007) also proposed this approach for fish growth data. They assumed unobserved age had a simple Gamma distribution. In this study, we investigate whether SEV VonB parameter estimators are robust to the Gamma approximation of true unobserved ages. By robust we mean lack of bias due to ME and model misspecification. Our results demonstrate that this method is not robust. We propose a flexible parametric Normal mixture distribution for the unobserved true ages to reduce this bias when estimating the length-age relationship with a VonB model. We investigate the performance of this approach in comparison to the Gamma age model through extensive simulation studies and a real-life data set.

To My Parents

Acknowledgements

I would like to express my most sincere and deepest gratitude to my supervisor, Dr. Noel Cadigan. This thesis would not have been completed without his excellent guidance, generous support and continuous motivation. I would also like to thank him for his exceptional generosity with his time and energy in my PhD application.

I offer my very special thanks to Dr. Taraneh Abarin, my co-supervisor, for her suggestions and guidance throughout this study. I am very grateful for the hospitality I was offered during many holiday parties over the last two years.

I gratefully acknowledge the funding received towards my MSc from School of Graduate Studies, the Department of Mathematics and Statistics, and my supervisor in the form of graduate fellowships and teaching assistantships.

My sincere thanks go to all the professors from the Statistics wing, MUN for all their help and support. I am grateful to Dr. JC Loredo-Osti and Dr. Zhaozhi Fan, for giving me suggestions and help in my PhD application.

Furthermore, I would like to thank Center for Fisheries Ecosystems Research, Fisheries and Marine Institute of MUN for creating a friendly and comfortable working environment. Thanks to Karen Dwyer at the Department of Fisheries and Oceans in St. John's, NL, for providing the data used in this thesis and for helpful discussions about my research.

Last but most importantly, I am very grateful to my parents and my younger sister for their strong belief in my capabilities and for all of their eternal love, support and encouragement during my stay far away from them.

Table of contents

List of tables

List of figures

Chapter 1

Introduction

1.1 Measurement Error

The purpose of regression analysis is to make inference about a mathematical model expressed in terms of an explanatory variable X. However, due to different reasons, the most obvious being the inaccuracy of measurements, X may not be observable. Hence, we consider X_T to be the true but unobservable covariate and instead of it we observe X as a proxy variable. The substitution of X for X_T complicates statistical analysis and creates problems due to the error in the measurement of X for X_T. Error in covariates is a problem in many scientific areas. For example, in fisheries science, error in age estimates for individual fish could be a consequence of misinterpretation by readers of ageing structures (e.g. scales and otoliths) or inability of ageing structures to accurately record growth sequence information.

Special estimation methods are needed when a covariate in a model is measured with error. Regression analysis ignoring this error is known to produce biased and falsely precise estimates of the regression parameters (e.g. Fuller, 2009 [21]; Carroll et al., 2006 [12]). Further effects are unreliable coverage levels of confidence intervals and reduced power of hypothesis tests.

1.1.1 Measurement Error Models

Specification of a model for the measurement error (ME) process is necessary for analyzing a ME problem. There are two general approaches:

Classical ME Model: This is appropriate when a quantity is measured by some device and repeated measurements vary around the true value. The model can be specified as

$$X = X_T + U,$$

where U, the ME, is assumed to be independent of X_T with mean zero and variance σ_u^2, which is the ME variance.

Berkson Error Model: This is appropriate in a situation when a group's average is assigned to each individual suiting the group's characteristics. The group's average is thus the measured value; i.e., the value that enters the analysis, and the individual value is the true value. This type of error model can be defined as

$$X_T = X + U,$$

where U is assumed to be independent of X with mean zero and variance σ_u^2.

Finally, a very important difference between classical ME and Berkson error models is that in the classical model, the variability of the observed X is larger than the variability of the true X_T. In the Berkson model it is the other way around. One needs information about the data structure in order to perform a ME analysis.

1.1.2 Sources of Data

The data sources can be separated into two main categories, 1) internal subsets of the primary data, and 2) data from external or independent sources. Within each of these broad categories, there are three types of data described in Carroll et al. (2006) [12] which are as follows:

1. **Validation Data:** In which X_T is observed directly.

2. **Replication Data:** In which replicates of X are available.

3. **Instrumental Data:** If the ME is unknown then one needs to estimate it with the validation data or replicate measurements of X. However, it is not always possible to obtain replicates, and thus estimation of ME variance σ_u is sometimes impossible. When there is no information about the σ_u, the estimation of regression parameters is still possible if the data contains an instrumental variable I in addition to X. I must be observed and correlated with X_T. Furthermore, it must be uncorrelated with the ME, $U = X - X_T$.

1.1.3 Differential and Non-differential Errors

It is important to make a distinction between differential and non-differential MEs. Non-differential ME occurs in a broad sense when one would not be concerned with X if X_T were available. This is the type of error in a fish growth model, e.g. when the true age (X_T) of a fish is known then the observed age (X) measured with error does not have any information about the length of fish (Y). Let the conditional pdf of Y given $X_T = x_T$ be $f_{Y|X_T}(y \mid x_T)$, and the conditional pdf of X given $X_T = x_T$ be $f_{X|X_T}(x \mid x_T)$. We use the abbreviation pdf to stand for probability density function.

When ME is non-differential then the joint pdf of (Y,X) given $X_T = x_T$ is

$$f_{Y,X|X_T}(y, x \mid x_T) = f_{Y|X_T}(y \mid x_T) f_{X|X_T}(x \mid x_T). \tag{1.1}$$

However, when ME is differential, using standard conditioning arguments, the joint pdf of (Y,X) given $X_T = x_T$ becomes

$$f_{Y,X|X_T}(y, x \mid x_T) = f_{Y|X_T}(y \mid x_T) f_{X|Y,X_T}(x \mid y, x_T). \tag{1.2}$$

Note that the only difference between Eqns. (1.1) and (1.2) is in the ME term. In the former, under non-differential ME, X and Y are independent when X_T is given. Therefore, one can estimate parameters in models for responses even when the true covariates are not observable. However, parameter estimation is difficult when ME is differential, since we must determine the conditional distribution of X given X_T and the response Y. This is essentially impossible to do in practice unless one has a subset of the data in which all of $(Y; X; X_T)$ are observed, i.e., a validation data set.

1.1.4 Model Identification

Model identification is an important aspect of ME modeling. A model is identified if all its parameters can be uniquely estimated from the data. According to Fuller (2009) [21], the parameter θ of the distribution of a random variable Z, with distribution function $F_Z(z; \theta)$, is identified if, for any two parameters in the parameter space, $\theta_1 \neq \theta_2$, $(\theta_1, \theta_2) \in \Theta$, $F_Z(z; \theta_1) \neq F_Z(z; \theta_2)$ for at least one value of z. If the parameters of a model are identified then the model is said to be identified.

1.1.5 The Effect of ME in Simple Linear Regression

The linear model specifies Y as a function of X_T,

$$Y = \beta_o + \beta_1 X_T + \epsilon, \tag{1.3}$$

where β_0 and β_1 are the intercept and slope parameters, respectively. Suppose X_T is measured with the classical ME model and defined as

$$X = X_T + U. \tag{1.4}$$

Assume that ϵ and U follow $N(0, \sigma_e^2)$ and $N(0, \sigma_u^2)$, respectively. Further, assume that X_T follows a $N(\mu_{x_T}, \sigma_{x_T}^2)$ and that X_T, ϵ and U are mutually independent. Then the multivariate Normal distribution (MVN) of (X_T, U, ϵ) is

$$\begin{pmatrix} X_T \\ U \\ \epsilon \end{pmatrix} \sim MVN \left[\begin{pmatrix} \mu_{x_T} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{x_T}^2 & 0 & 0 \\ 0 & \sigma_u^2 & 0 \\ 0 & 0 & \sigma_e^2 \end{pmatrix} \right]. \tag{1.5}$$

Therefore, the bivariate Normal distribution (BVN) of (ϵ, X_T) is

$$\begin{pmatrix} \epsilon \\ X_T \end{pmatrix} = \begin{pmatrix} Y - (\beta_o + \beta_1 X_T) \\ X - U \end{pmatrix} \sim \text{BVN} \left[\begin{pmatrix} 0 \\ \mu_{x_T} \end{pmatrix}, \begin{pmatrix} \sigma_e^2 & 0 \\ 0 & \sigma_{x_T}^2 \end{pmatrix} \right]. \tag{1.6}$$

By rearranging Eqn. (1.6), the BVN of (Y, X) is

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \text{BVN} \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \right], \tag{1.7}$$

where $\mu_1 = E(Y) = \beta_o + \beta_1 \mu_{x_T}$, $\mu_2 = E(X) = \mu_{x_T}$, and the elements of the variance-covariance matrix of the BVN of (Y, X) are

$$\sigma_{11} = \mathrm{V}(\beta_o + \beta_1 X_T + \epsilon) = \beta_1^2 \sigma_{x_T}^2 + \sigma_e^2,$$

$$\sigma_{22} = \mathrm{V}(X) = \mathrm{V}(X_T + U) = \sigma_{x_T}^2 + \sigma_u^2,$$

$$\sigma_{12} = \mathrm{Cov}(Y, X) = \mathrm{Cov}(\beta_o + \beta_1 X_T + \epsilon, X_T + U) = \mathrm{Cov}(\beta_1 X_T, X_T) = \beta_1 \sigma_{x_T}^2.$$

Finally, the BVN in Eqn. (1.7) is

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \mathrm{BVN} \left[\begin{pmatrix} \beta_0 + \beta_1 \mu_{x_T} \\ \mu_{x_T} \end{pmatrix}, \begin{pmatrix} \beta_1^2 \sigma_{x_T}^2 + \sigma_e^2 & \beta_1 \sigma_{x_T}^2 \\ \beta_1 \sigma_{x_T}^2 & \sigma_{x_T}^2 + \sigma_u^2 \end{pmatrix} \right]. \tag{1.8}$$

Attenuation Bias: If the true covariate X_T were observed in a sample of size n observations (Y_i, X_{T_i}) of (Y, X_T), the ordinary least squares (OLS) estimator of β_1 of the linear model in Eqn. (1.3) is

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^n (X_{T_i} - \bar{X}_T)(Y_i - \bar{Y})}{\sum_{i=1}^n (X_{T_i} - \bar{X}_T)^2}. \tag{1.9}$$

In the presence of ME, the model involving the measured value of X_T, X, is

$$Y = \beta_0 + \beta_1 X + \epsilon^*, \tag{1.10}$$

where $\epsilon^* = -\beta_1 U + \epsilon$. Given a random sample of n observations (Y_i, X_i) of (Y, X), the OLS estimator of β_1 for the model in Eqn. (1.10) is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}. \tag{1.11}$$

According to Fuller (2009) [21] if (Y, X) is a bivariate Normal vector (Eqn. 1.8), then the expected value of $\hat{\beta}_1$ is given by

$$\begin{aligned} E[\hat{\beta}_1] &= \frac{\mathrm{Cov}(X, Y)}{\mathrm{V}(X)} \\ &= \frac{\mathrm{Cov}(X_T + U, Y)}{\mathrm{V}(X_T + U)} \\ &= \frac{\mathrm{Cov}(X_T + U, \beta_0 + \beta_1 X_T + \epsilon)}{\mathrm{V}(X_T) + \mathrm{V}(U)} \\ &= \beta_1 \frac{\sigma^2_{x_T}}{\sigma^2_{x_T} + \sigma^2_u} \\ &= \beta_1 (1 - \frac{\sigma^2_u}{\sigma^2_{x_T} + \sigma^2_u}). \end{aligned} \tag{1.12}$$

In addition, X and ϵ^* are correlated with each other, i.e.,

$$\mathrm{Cov}[X, \epsilon^*] = \mathrm{Cov}[X_T + U, \epsilon - \beta_1 U] = -\beta_1 \sigma^2_u \neq 0.$$

Hence, from Eqn. (1.12), $E[\hat{\beta}_1] \neq \beta_1$; consequently, the OLS estimator of β_1 is biased and inconsistent. The expected value of $\hat{\beta}_1$ is attenuated toward zero when the ME is large. The bias does not reduce with increasing sample size, n. The extent of the attenuation is measured by the quantity $\Lambda = \frac{\sigma^2_{x_T}}{\sigma^2_{x_T} + \sigma^2_u}$, which is known as the reliability ratio.

Model Identification: Model identification is a key issue here. All the parameters in the model can be represented by the vector, $\theta = (\beta_0, \beta_1, \mu_{x_T}, \sigma^2_{x_T}, \sigma^2_u, \sigma^2_e)$. The BVN distribution in Eqn. (1.8) is completely determined by its mean and variance-covariance matrix, which involve $\mathrm{E}(Y), \mathrm{E}(X), \mathrm{V}(X), \mathrm{V}(Y)$, and $\mathrm{Cov}(X, Y)$. Each of these are functions of the six elements of θ. Hence, there are many parameter vectors θ that produce the same mean vector and variance-covariance matrix of the BVN. Thus by definition, the model defined in Eqns. (1.3)-(1.4) is not identified. This ME model is identified if one of the parameter values is known in addition to the information in the sample. For example, if σ_u^2 is known then the other values of the parameters uniquely determine the BVN distribution of X and Y.

The OLS of β_1 can be corrected for bias to get the best linear unbiased estimator of β_1. This can be achieved when the reliability ratio Λ is known. By rearranging Eqn. (1.12), a bias-corrected estimator of β_1 is

$$E[\hat{\beta}_1 \Lambda^{-1}] = \beta_1. \tag{1.13}$$

1.1.6 ME Methods

In the previous section, we showed that the OLS estimator is typically biased when there is covariate ME, and the direction as well as severity of the bias increases with the magnitude of the ME.

Two basic methods have been used in ME models (e.g. Stefanski, 2000 [43]). We review the functional and structural based methods in the following sections.

Functional Errors-in-Variables

In functional errors-in-variables (FEV) models, the unobserved true covariates are modeled as unknown, nonrandom constants (i.e. parameters) (e.g. Carroll et al., 2006 [12]). In this model, with a sample of size n, we have n measurements of unobserved true covariates; hence, the parameter vector includes $X_{T_1}, X_{T_2}, ..., X_{T_n}$ and the number of parameters increases linearly with n. When the number of nuisance parameters is large relative to n it is well known that maximum likelihood estimators (MLEs) for some parameters, particularly for variance parameters, can be largely biased and inconsistent (e.g. section 4.3 in Barndorff and Cox, 1994 [3]; Berger et al., 1999 [4]).

Finding a conditional distribution or score function that does not depend on nuisance parameters is one approach to deal with this bias problem due to many nuisance parameters. Methods based on FEV models can be divided into approximately consistent (remove most bias) and fully consistent methods (remove all bias as $n \to \infty$). Fully consistent methods for nonlinear regression models typically require assumptions on the distribution of the ME. Regression calibration and SIMEX are examples of approximately consistent methods while corrected scores and conditional scores methods are fully consistent for large classes of models.

Simulation-Extrapolation (SIMEX): The underlying concept of SIMEX, proposed in Cook and Stefanski (1994) [14] and further developed by Stefanski and Cook (1995) [42], is to determine the effect of ME on the parameter estimator experimentally via simulation, assuming that the ME variance is known or estimated from validation data. A detail discussion of the method can be found in Carroll et al. (2006) [12]. The SIMEX algorithm is executed in two steps: Simulation and Extrapolation.

1. **Simulation Step:** Computer simulated ME (i.e. random error) is added to the observed measurement X and the biased parameter estimate is computed with the additional simulated pseudo error terms. This process is repeated for several increments in the value of the simulated ME.

2. **Extrapolation Step:** It consists of modeling the trend between the biased parameter estimates and the corresponding size of the simulated ME. A nearly unbiased SIMEX estimator is then obtained by extrapolating the trend back to the point of zero ME.

We explain this concept with the simple linear model. We generated 1000 observations of X_T from a standard Normal distribution. Then, we generated 1000 observations of Y assuming a linear model (Eqn. 1.3) using the $\beta_0 = 0$, $\beta_1 = 1$, $\sigma_e = 1$ and X_T. We use $\sigma_u = 0.3$ in the Eqn. (1.4) to generate 1000 observations of X. We have considered the simulated ME λ such as 0.5, 1, 1.5 and 2. We randomly generate b = 100 independent datasets for each λ. For each (say b^{th}) dataset we follow the following steps:

- **Step 1:** We generated simulated pseudo errors $Z_{b,i}$ (i = 1, 2,. . . , 1000) from a standard Normal distribution.

- **Step 2:** We generate pseudo predictors $X_{b,i}(\lambda) = X_i + \lambda^{0.5}\ \sigma_u\ Z_{b,i}$ (i = 1, 2,. . . , 1000) for a specific value of λ.

- **Step 3:** Fit the linear model to $(Y_i,\ X_{b,i}(\lambda))$ for a specific value of λ. We used OLS method to estimate the slope parameter which is $\hat{\beta}_{1,b}(\lambda)$.

Repeat these steps b times and calculate the average simulation estimate of β_1, $\hat{\beta}_1(\lambda) = \frac{1}{b}\sum_{i=1}^{b}\hat{\beta}_{1,i}(\lambda)$. We plot the $\hat{\beta}_1(\lambda)$ against each value of $(1 + \lambda)$. We use a quadratic extrapolation fitting method to the trend back to the point of zero ME, i.e., $\lambda = -1$. Figure 1.1 illustrates that the SIMEX estimate $\hat{\beta}_1(\lambda)$ changes as added simulated ME increases. To estimate $\hat{\beta}_1$ if there is no ME we use a quadratic extrapolation fitting method to the trend back to the point of zero ME, i.e., $\lambda = -1$. The estimated value of β_1 is 0.985 while the true value of β_1 was 1.

Regression Calibration: Regression Calibration (RC) was introduced by Prentice (1982) [33] in a proportional hazards model application and generalized by Carroll and Stefanski (2006) [10] to increase its scope. The method and its applications are discussed extensively in Carroll et al. (2006) [12]. The RC algorithm is as follows:

- **Step 1:** Estimate the regression of X_T on X using replication or validation data. This is called the calibration function.

- **Step 2:** Replace the unobserved X_T by its estimate from the regression model, and then run a standard analysis to obtain parameter estimates.

- **Step 3:** Adjust the resulting standard errors to account for the estimation at the step 1, using either the bootstrap or asymptotic methods.

There are two important drawbacks of RC, 1) for linear and log-linear models RC provides asymptotically unbiased estimators whereas for nonlinear models the RC estimator is approximately unbiased (e.g. Buzas et al., 2003 [7]); 2) standard errors for parameter estimates are obtained from bootstrapping or asymptotic normal assumptions; consequently, statistical inference is not exact.

Structural Errors-in-Variables

In structural errors-in-variables (SEV) models the observed X and true unobserved covariate X_T are jointly considered to be random and vary in repeated sampling. Let Y be the response variable, and we define Y_T as a function of X_T,

$$Y_T = g(X_T; \theta_R), \tag{1.14}$$

where the mean function $g(.)$ is a continuous real valued function and θ_R is the parameters of the mean function. We define the classical ME models as

$$Y = Y_T + \epsilon, \tag{1.15}$$

$$X = X_T + U, \tag{1.16}$$

where ϵ follows $N(0, \sigma_e^2)$, and U follows $N(0, \sigma_u^2)$ with σ_u^2 is known. We assume that the conditional pdf of Y given $X_T = x_T$ is

$$f_{Y|X_T}(y \mid x_T; \theta_R),$$

and the conditional pdf of X given $X_T = x_T$ is

$$f_{X|X_T}(x \mid x_T; \sigma_u).$$

The joint pdf of (Y,X) is

$$f_{Y,X}(y, x; \theta) = \int \int f_{Y,X|Y_T,X_T}(y, x \mid y_T, x_T; \theta_R, \sigma_u)\ f_{X_T,Y_T}(x_T, y_T; \theta_R, \theta_E)\ \partial y_T\ \partial x_T,$$

where $\theta = (\theta_R, \theta_E)$. We assume that MEs in X and Y are independent so that

$$f_{Y,X|Y_T,X_T}(y, x \mid y_T, x_T; \theta_R, \sigma_u) = f_{X|X_T}(x \mid x_T; \sigma_u)\ f_{Y|Y_T}(y \mid y_T; \theta_R). \tag{1.17}$$

The joint pdf of (Y_T,X_T) can be expressed as

$$f_{X_T,Y_T}(x_T, y_T; \theta_R, \theta_E) = f_{Y_T|X_T}(y_T \mid x_T; \theta_R)\ f_{X_T}(x_T; \theta_E). \tag{1.18}$$

Combining Eqns. (1.17)-(1.18), the joint pdf of (Y,X) is

$$f_{Y,X}(y, x; \theta) = \int \int f_{X|X_T}(x \mid x_T; \sigma_u)\ f_{Y|Y_T}(y \mid y_T; \theta_R)\ f_{Y_T|X_T}(y_T \mid x_T; \theta_R)\ f_{X_T}(x_T; \theta_E)\ \partial y_T\ \partial x_T.$$

If we know the true covariate X_T, then we assume that we would know the true response exactly, i.e., $Y_T \mid X_T = g(X_T; \theta_R)$ is the regression model value of response for a true covariate X_T; hence,

$$f_{Y_T|X_T}(y_T \mid x_T; \theta_R) = \begin{cases} 1 & y_T = g(x_T; \theta_R) \\ 0 & \text{otherwise.} \end{cases} \tag{1.19}$$

Eventually, under the above assumption,

$$f_{Y|Y_T}(y \mid y_T; \theta_R) = f_{Y|X_T}(y \mid x_T; \theta_R).$$

Therefore, the joint pdf of (Y, X) is

$$f_{Y,X}(y, x; \theta) = \int f_{Y|X_T}(y \mid x_T; \theta_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f_{X_T}(x_T; \theta_E)\ dx_T. \tag{1.20}$$

The integral is replaced by a sum if X_T is a discrete random variable. In Eqn. (1.20),

1. **Response Model:** $f_{Y|X_T}(y \mid x_T; \theta_R)$ describes the relationship between Y and X_T;

2. **ME Model:** $f_{X|X_T}(x \mid x_T; \sigma_u)$ describes the relationship between X and X_T;

3. **Covariate Model:** The true unobserved covariate X_T is considered as a random variable with pdf $f_{X_T}(x_T; \theta_E)$.

The likelihood for the observed data is then maximized over all the parameters in two of the above three component distributions, i.e., response model and covariate model, to obtain MLEs of the parameters.

The SEV approach requires parametric distributional assumptions for the unobserved covariate, X_T, which none of the preceding FEV methods required. It is common to assume a Normal distribution for the covariate model, but unless there are validation data, it is not possible to assess the adequacy of the covariate model using the data. In SEV models, the distribution of X_T is speculative and could be quite different from it's true underlying distribution leading to model misspecification. Hence, an important concern is whether the SEV estimates are robust to misspecification of the distribution of the true covariate. The sensitivity of the regression parameters estimators to covariate model misspecification in SEV models has been illustrated in Carroll, Roeder and Wasserman (1999) [11].

Semi-parametric and flexible parametric modeling are two approaches that have been explored to address potential robustness issues in specifying a pdf for X_T. Semi-parametric methods leave the pdf for X_T unspecified, and the pdf for X_T is essentially considered as another parameter that needs to be estimated. These models have the advantage of model robustness but may lack efficiency relative to the full likelihood (e.g. Suh and Schafer, 2002 [45]). A flexible parametric model is described in Carroll, Roeder and Wasserman, 1999 [11], where mixtures of Normals were used to approximate the covariate model (since the true covariate model is generally unknown) to estimate the parameters of a linear errors-in-variables model and a change-point Berkson model. Flexible parametric approaches have been studied in, e.g. Kuchenhoff and Carroll (1997) [30], Schafer (2002) [39].

The choice between functional or structural models depends both on the assumptions one is willing to make and, in a few cases, the form of the model relating Y to X. To explain this point we have taken one of the examples stated in Fuller (2009) [21]. Consider the relationship between yield of corn, say Y, and available nitrogen, say X_T, in the soil. To estimate the available soil nitrogen, it is necessary to sample the soil of the experimental plot and to perform a laboratory analysis on the selected sample. As a result of the sampling and of the laboratory analysis, we observe X instead of X_T. The description of the collection of the soil nitrogen data allows two interpretations of the true values X_T. First, assume that the fields are a set of experimental fields managed by the experiment station in ways that produce different levels of soil nitrogen in the different fields. In such a situation, one would treat the true, but unknown, nitrogen levels in the different fields as fixed. In this case the FEV approach is a natural choice. On the other hand, if the fields were a random sample of farmers fields, the X_T could be treated as random variables. In this case SEV is preferable. In addition, the type and amount of data available also play role. For example, validation data provides information on the distribution of X_T, and may make structural modeling more preferable.

To sum up, the SEV approach can yield high efficiency and allow construction of likelihood-based inference, whereas this may be more difficult in FEV models due to the large number of parameters. However, robustness in SEV estimates is an important issue due to misspecification of the covariate model. Despite this shortcoming, SEV models are more common and usually preferable to FEV models because of the general applicability of SEV models, and their simple computation and potential gain in efficiency.

1.2 Fish Growth and ME

Body growth is an important factor in fish population dynamics and determining sustainable levels of fishing. Growth varies from species to species, for different populations of a species (i.e. stocks), and different year-classes within a population (e.g. Chen and Mello, 1999 [13]). Growth information is essential for fish stock assessment which is a process that produces scientific advice on the health of a fish stock and the impacts of fisheries. Fishing quotas are usually based on weight whereas mortality involves the number of fish in a stock. Good information about body growth rates are required to predict the impacts of future fishing quotas on stock mortality and when deciding what are good and sustainable harvest rates. Poor information on body growth rates may lead to incorrect prediction of the mortality impacts of fishing and other population dynamics (e.g. Vincenzi et al., 2014 [49]). Therefore, estimation of growth rates is a common and important part of fisheries stock assessment studies.

Generally, two basic types of growth data are available from commercial fisheries: 1) age-to-length data in which one age and length measurement per fish is collected from a large sample of fish, and 2) multiple measurements of the same fish via capture-recapture or other repeated measures studies (e.g. Francis, 1988 [19]). The first type of data is more common. Ages are determined by counting growth bands in ear bones (otoliths) and this can involve substantial error especially for older fish because the bands get harder to differentiate as a fish gets older. There are several growth models available in literature; however, the Von Bertanlanffy growth model is the most widely used and its parameters are useful for describing a fish growth curve as discussed in Von Bertalanffy (1960) [50].

1.2.1 Von Bertanlanffy (VonB) Growth Model

The theory of the VonB model is based on the assumption that the change in length per unit time, $\frac{dY_T}{dX_T}$, declines with length. That is, the growth rate of large individuals is less than the growth rate of small individuals. If Y_T denotes the length at age X_T, then the VonB growth rate model is based on the differential equation

$$\frac{dY_T}{dX_T} = k(L_\infty - Y_T), \tag{1.21}$$

and $\frac{dY_T}{dX_T} = 0$ when $Y_T = L_\infty$. Thus, the growth rate of fish will get smaller and eventually becomes zero as a fish nears its maximum possible length L_∞. The parameter

L_∞ is the asymptotic length at which the growth rate is zero and k is the growth rate parameter. Assuming that $Y_T = 0$ when $X_T = 0$, the solution of Eqn. (1.21) is

$$Y_T = L_\infty(1 - e^{-kX_T}).$$

We illustrate this model in Figure 1.2 (top panel) when $L_\infty = 120$ and $k = 0.2$. This figure demonstrates that for the VonB model, fish grow more quickly when they are young, growth slows gradually as the individual fish ages, and eventually stops growing at length, $L_\infty = 120$.

Generally, the length of a fish during its first year (age zero) is not zero, i.e., $Y_T > 0$ at $X_T = 0$. To account for this, we use the following form of the VonB growth model,

$$Y_T(X_T; L_\infty, a_o, k) = L_\infty(1 - e^{-k(X_T - a_o)}), \tag{1.22}$$

where $a_o < 0$ is the theoretical age at which a fish has zero length. In practical terms age cannot be negative, but if $Y_T > 0$ at age $X_T = 0$ and we extrapolate the growth curve back to when $Y_T = 0$, we obtain a negative age (see Figure 1.2, bottom panel). The VonB model (Eqn. 1.22) is used to describe the mean growth of a population where L_∞, k and a_0 are the population mean growth parameters.

1.2.2 The Effect of ME in VonB Model

In reality the length of fish can usually be measured fairly accurately, however, error in measuring age (i.e. covariate ME) is very common in age-to-length data. Age reading errors may be due to 1) misinterpretation by readers of ageing structures to record growth sequence information, 2) different readers provides different age measurements. Therefore, in practice, X_T is not observed, and instead of it we observe X.

The standard method used in fish stock assessments to fit VonB models, i.e., Eqn. (1.22), to data is either by nonlinear least squares (e.g. Tomilnson and Abramson, 1961 [46]) or maximum likelihood (e.g. Kimura, 1980 [29]). In both model fitting procedures Y_T is thought of as the expected or mean length of a fish with age X, and the fitting procedure simply selects parameter values to minimize the difference between the observed and the expected values of Y_T for each observed age X. This approach assumes that all the deviations between the model and the data are due to variation in length measurements. ME in age may cause bias, and we will investigate the potential magnitude of the bias using a simulation experiment.

In a simulation study, we randomly generated 1000 independent datasets. For each dataset we follow the following steps:

- **Step 1:** We generated 1000 true ages X_T from a Gamma distribution with $\alpha = 7$ and $\beta = 1$. The pdf of the Gamma distribution with parameters α and β is

$$f_{X_T}(x_T; \alpha, \beta) = \frac{1}{\beta^{\alpha}\, \Gamma(\alpha)}\; x_T^{\alpha-1}\; e^{-\frac{x_T}{\beta}}.$$

- **Step 2:** We generated 1000 true lengths Y_T assuming a VonB growth model (Eqn. 1.22) using $L_\infty = 120, k = 0.2$ and $a_o = -0.1$ and the true age X_T generated in step 1.

- **Step 3:** We considered the classical error model of Y defined in Eqn. (1.15) to generate 1000 observed lengths Y with $\sigma_e = 0.1$. We used the classical error models of X defined in Eqn. (1.16) to generate 1000 observed ages X with a range of σ_u values vary from 0 to 0.8.

- **Step 4:** For each data set we used nonlinear least squares to estimate the VonB growth parameters.

Finally, we calculate the average simulated estimates of growth model parameters. Figure 1.3 displays the change in estimates of growth parameters L_∞, k, and a_0 with the magnitude of ME. The results illustrate that as the ME in age increases the change in average simulated estimates of growth parameters is substantial. When there is no ME in age, i.e., $\sigma_u = 0$ the average simulated estimates are the same as their corresponding true values. For example, the average estimated value of L_∞ is 120. However, as σ_u increases the average estimated values of L_∞ increases substantially. For instance, when $\sigma_u = 0.8$ then the average estimated value of L_∞ is 127, which is very different from its true value 120. Therefore, ME in age has a substantial effect on the estimates of growth parameters.

MEs in age and length of fish have substantial consequences on estimates of growth, mortality, recruitment and yield (e.g. Bradford, 1991 [6]; Reeves, 2003 [35]). Covariate ME has long been recognized as important in fisheries science when fitting linear regression models (Schnute, 1990 [38]), stock-recruit models (Walters and Ludwig, 1981 [47]), simple biomass production models (Uhler, 1980 [48]), and growth models (e.g. Kitakado, 2000 [28], Cope and Punt, 2007 [15]), which is the focus of this study. It is difficult to estimate the ME in both age (i.e. the covariate) and length (i.e. the response) without additional information on the accuracy of at least one of these sources. An SEV model approach was used by Suh and Schafer (2001) [45]. They considered a situation where individual growth measurements were available, with error in both length and age, but also a smaller validation sample in which there were no aging errors. Such validation data are not commonly available. Cope and Punt (2007) [15] considered the more common situation in which estimates of the ageing error variance are available from multiple age measurements of a sample of fish. They suggested an SEV model with a Gamma distribution for the unobserved true ages, and they showed using simulations that this approach provided more precise estimates of VonB model parameters compared to the conventional “errors in length” nonlinear least squares method.

1.3 Organizations of Subsequent Chapters

The following chapters are organized as follows. In Chapter 2, we approximate the unobserved true age distribution using a simple Gamma distribution in an SEV VonB model to estimate the VonB model parameters. This approach is commonly used for age-to-length data (e.g. Cope and Punt, 2007 [15]). In this study, we call it the SEV VonB Gamma model. We investigate whether SEV VonB Gamma model parameter estimators are robust to misspecification of the true unobserved age distribution or not. We consider robustness (Huang, X. (2006) [53], and Huang, Stefanski, and Davidian (2006) [26]) to mean lack of bias in estimators for the parameters.

In Chapter 3, we propose an SEV VonB growth model that involves mixtures of Normal distributions for the unobserved ages, which is more robust to misspecification of the true unobserved age distribution. For our purposes, we call it the SEV VonB G-Normal mixture model. We compare the estimators based on the SEV VonB G-Normal mixture model with that based on the SEV VonB Gamma model in terms of large sample bias.

In Chapter 4, we extend the SEV VonB G-Normal mixture model in Chapter 3 to account for between-individual (BI) variation in growth. We assume BI variation in growth appears because individuals achieve different asymptotic sizes (L_∞). Here, we call it the SEV VonB G-Normal mixture BI model. We compare the estimators based on the SEV VonB G-Normal mixture BI model with that based on the SEV VonB Gamma BI model in terms of finite sample bias.

In Chapter 5, we apply the SEV VonB G-Normal mixture BI model to the length-at-age Greenland Hailbut data in the NAFO management unit Subarea 2 + Divisions 3KLMNO provided by Dwyer et al. (2016)[17].

In Chapter 6, we conclude by summarizing the main ideas proposed in this thesis and the main results obtained.

1.4 Figures

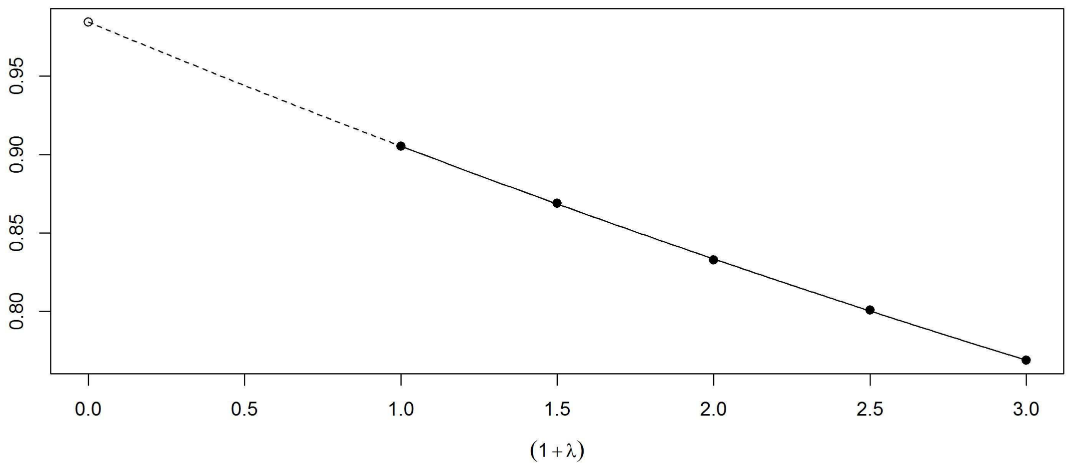

Figure 1.1: Simulation-Extrapolation (SIMEX) estimate $\hat{\beta}_1(\lambda)$ of slope parameter of simple linear model. The simulated measurement error denoted by λ.

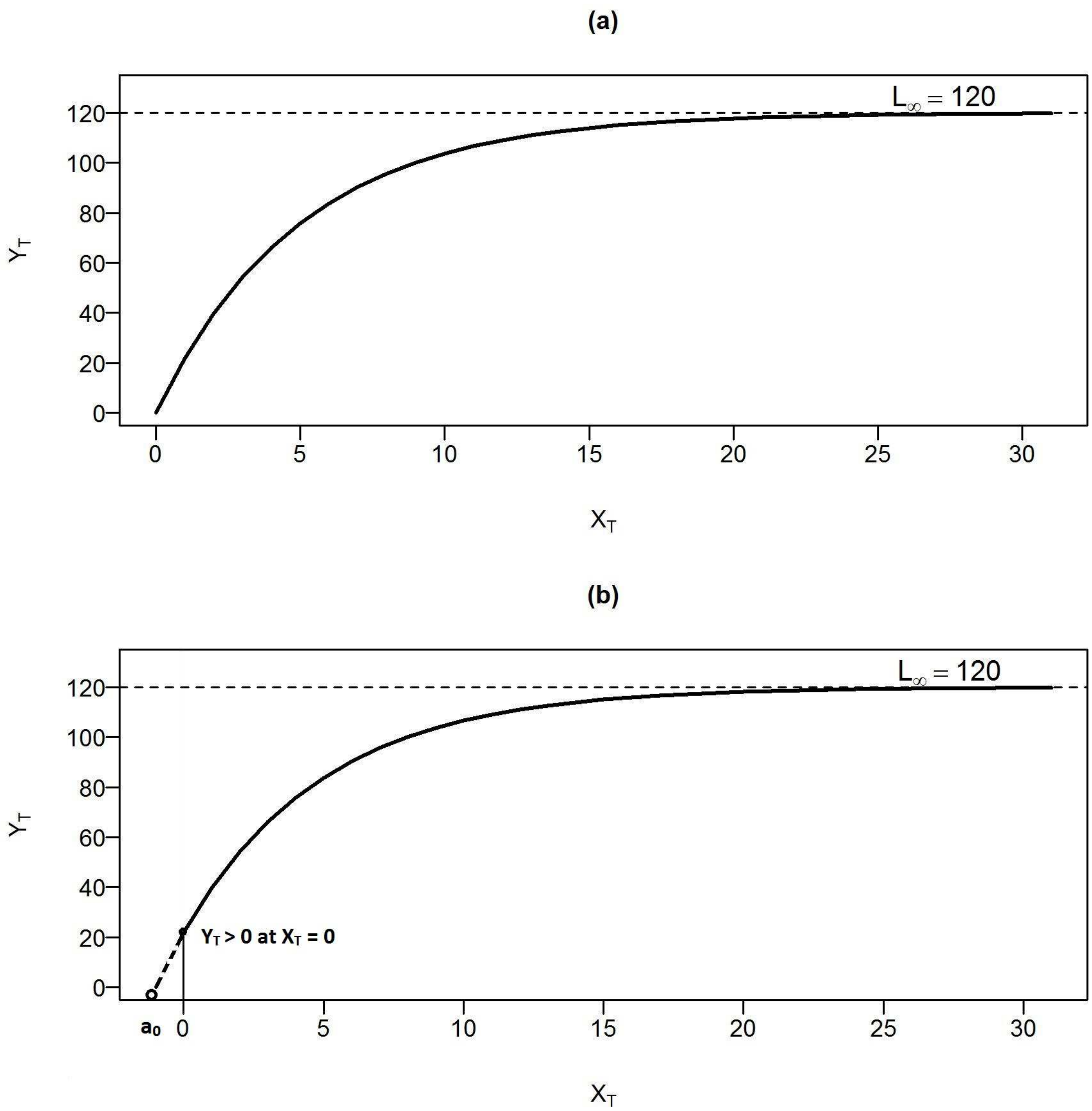

Figure 1.2: Application of the Von Bertalanffy (VonB) model for $L_\infty = 120$, k = 0.2. The top panel illustrates the growth curve when $a_0 = 0$, and the bottom panel illustrates when $a_0 < 0$. Y_T and X_T denote the true length and true age of fish, respectively.

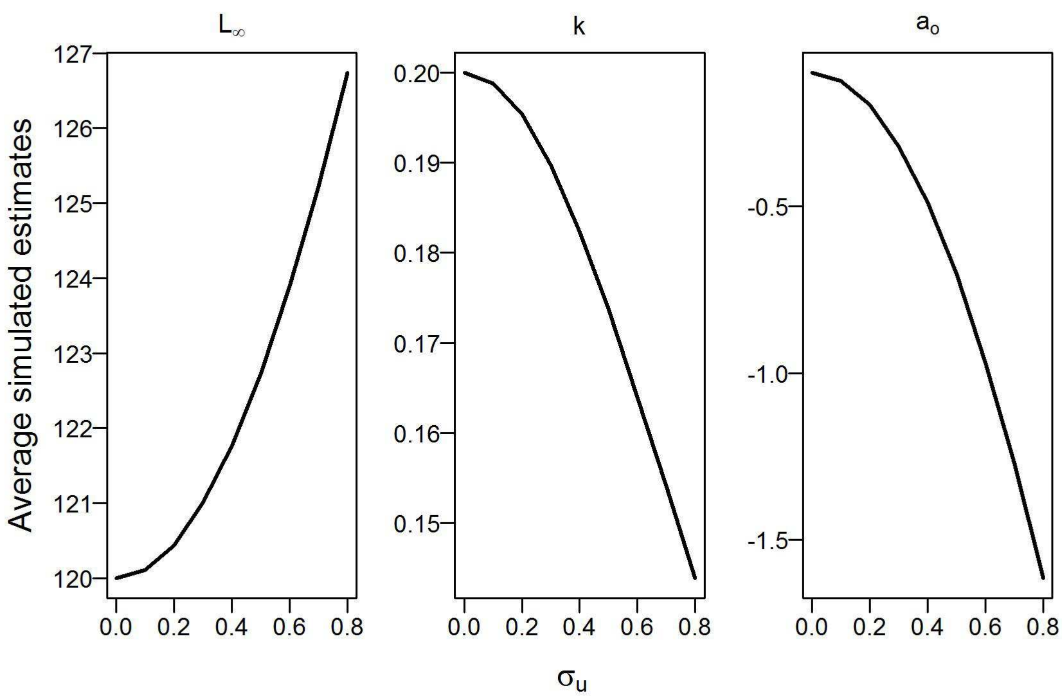

Figure 1.3: Average simulated values of Von Bertalanffy (VonB) parameters L_∞, k & a_o estimated using the nonlinear least squares method. The measurement error variance is denoted by σ_u.

Chapter 2

Robustness of Structural Errors-in-Variables Model

2.1 Introduction

In Chapter 1 we introduced the structural errors-in-variables (SEV) model. This approach was proposed to account for age ME in VonB growth models in fisheries science. Cope and Punt (2007) [15] suggested an SEV model with a Gamma distribution for the unobserved true ages, X_T, and their results showed that this approach provided more precise estimates of VonB model parameters compared to the nonlinear least squares method.

Mohammed (2015) [32] extended the model in Cope and Punt (2007) [15] to include between-individual variation in growth, but still assuming that unobserved ages have a simple Gamma distribution. The simulation studies in Mohammed (2015) [32] indicated that misspecification of the distribution of unobserved age did not result in much bias in the estimates of VonB parameters. However, we note that the age reading error variance considered in Mohammed (2015) [32] was very small. Cadigan and Campana (2016) [8] also assumed a Gamma age distribution in their hierarchical modelling of growth for many fish populations, and they showed that parameter estimates did not change much when a more flexible parametric age distribution was used. Huang, Stefanski and Davidian (2006) [26] showed that when the covariate ME is low the asymptotic bias of parameter estimators is close to zero for SEV models. Therefore, potential bias is due to the joint effect of model misspecification and the magnitude of covariate ME. This was not studied much in Cope and Punt (2007) [15], Mohammed (2015) [32], or Cadigan and Campana (2016) [8].

This is the motivation for this chapter. We investigate whether SEV VonB model parameter estimators based on a simple Gamma distribution for unobserved ages are robust to misspecification of the true unobserved age distribution. We consider robustness (e.g. Huang, X. (2006) [53], and Huang, Stefanski, and Davidian (2006) [26]) to mean lack of bias in estimators for the parameters of interest. In practice the true age distribution may be quite complicated, varying from unimodal to multimodal. The age distribution of a fish population will depend on the reproduction and survival of previous cohorts, and reproduction rates and early life-stage survival for fish are known to vary widely from year to year. This can lead to potentially complicated and multi-modal age distributions; hence, robustness to such misspecifications is practically relevant. Heagerty and Kurland (2001) [23] proposed a method for evaluating large sample bias due to misspecification of the random effects distribution in generalized linear mixed models. A general framework to quantify the bias due to covariate ME misspecification was proposed by Hossain and Gustafson (2011) [24]. We used their approaches to compute the large sample bias in SEV VonB model parameter estimators to investigate how this is jointly affected by misspecification of the distribution of unobserved true ages and the magnitude of the ME in age.

2.2 Structural Errors-in-Variables Model

In this section, we investigate a method to assess the robustness of the estimates of parameters in the SEV model. Recall from Chapter 1 that the joint pdf of (Y,X) defined in Eqn. (1.20) is

$$f_{Y,X}(y, x; \theta) = \int f_{Y|X_T}(y \mid x_T; \theta_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f_{X_T}(x_T; \theta_E)\ dx_T,$$

where

1. **Response Model:** $f_{Y|X_T}(y \mid x_T; \theta_R)$ describes the relationship between Y and X_T;

2. **ME Model:** $f_{X|X_T}(x \mid x_T; \sigma_u)$ describes the relationship between X and X_T;

3. **Covariate Model:** the true covariate X_T is considered as a random variable with pdf $f_{X_T}(x_T; \theta_E)$.

Let $f^T_{X_T}(x_T; \Theta_E)$ denote the true pdf of X_T and $f^A_{X_T}(x_T; \theta_E)$ denote the assumed pdf of X_T that we use in the SEV model. If the true covariate model with pdf $f^T_{X_T}(x_T; \Theta_E)$ is misspecified as the assumed covariate model with pdf $f^A_{X_T}(x_T; \theta_E)$, then under such misspecification θ_E may no longer be meaningful; however, θ_R will still be meaningful. In this chapter, our interest is in how sensitive the bias in the estimate of θ_R is to such misspecification.

Note that the true response model and the true ME model are the same as the corresponding assumed models because we are assuming they are correctly specified; only the covariate model is misspecified.

2.2.1 Covariate Model Correct Specification

Under the correct model, the observed data likelihood is

$$L^T(\Theta \mid Y, X) = \int f_{Y|X_T}(y \mid x_T; \theta_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f^T_{X_T}(x_T; \Theta_E)\ dx_T, \tag{2.1}$$

where $\Theta = (\theta_R, \Theta_E)$. The log-likelihood function for Θ based on Y and X under correct specification is

$$\mathbf{l}(\Theta \mid Y, X) = \log\{L^T(\Theta \mid Y, X)\}.$$

The score function for Θ is

$$\mathbf{S}(\Theta \mid Y, X) = \frac{\partial}{\partial\Theta}\mathbf{l}(\Theta \mid Y, X) = \frac{\frac{\partial}{\partial\Theta}L^T(\Theta \mid Y, X)}{L^T(\Theta \mid Y, X)}.$$

The expected score function, $E[\mathbf{S}(\Theta \mid Y, X)]$, evaluated at the true parameter value θ^* of Θ is

$$E[\mathbf{S}(\theta^* \mid Y, X)] = 0. \tag{2.2}$$

The proof is as follows:

$$\begin{aligned}
E[\mathbf{S}(\theta^* \mid Y, X)] &= \int\int \frac{\frac{\partial}{\partial\Theta}L^T(\theta^* \mid Y, X)}{L^T(\theta^* \mid Y, X)} L^T(\theta^* \mid Y, X)\ dy\ dx \\
&= \int\int \frac{\partial}{\partial\Theta} L^T(\theta^* \mid Y, X)\ dy\ dx \\
&= \frac{\partial}{\partial\Theta}\int\int L^T(\theta^* \mid Y, X)\ dy\ dx \\
&= \frac{\partial}{\partial\Theta}\int\int\int f_{Y|X_T}(y \mid x_T; \theta^*_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f^T_{X_T}(x_T; \Theta^*_E)\ dx_T\ dy\ dx \\
&= \frac{\partial}{\partial\Theta}\ (1) \\
&= 0.
\end{aligned}$$

2.2.2 Covariate Model Misspecification

Under the misspecified (assumed) covariate model, the observed likelihood in Eqn. (2.1) is

$$L^A(\theta \mid Y, X) = \int f_{Y|X_T}(y \mid x_T; \theta_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f^A_{X_T}(x_T; \theta_E)\ dx_T, \tag{2.3}$$

where $\theta = (\theta_R, \theta_E)$ be a $(P \times 1)$ vector. The log-likelihood function for θ based on Y and X under misspecification is

$$l(\theta \mid Y, X) = \log\{L^A(\theta \mid Y, X)\}.$$

The score function for θ for the misspecified model is

$$S(\theta \mid Y, X) = \frac{\partial}{\partial\theta} l(\theta \mid Y, X) = \frac{\frac{\partial}{\partial\theta} L^A(\theta \mid Y, X)}{L^A(\theta \mid Y, X)}.$$

The SEV MLE of θ under the misspecified covariate model are the values maximizing Eqn. (2.3). Let $\hat{\theta}$ be the SEV MLE of θ. Hence, $\hat{\theta}_R$ is the SEV MLE of θ_R.

Define $\theta(\sigma_u)$ as a function of σ_u implicitly via

$$E[S\{\theta(\sigma_u) \mid Y, X\}] = 0. \tag{2.4}$$

The expectation is taken with respect to the joint pdf of (Y,X) (Eqn. 2.1) for the correct specification of X_T. Let $\theta(\sigma_u) = (\theta_R(\sigma_u), \theta_E(\sigma_u))$ be the analytical solution that makes the expected score equation under misspecification (Eqn. 2.4) equal zero.

Theoretical Robustness of MLE of θ_R : Huang, X. (2006) [53], and Huang, Stefanski, and Davidian (2006) [26] proposed a method to study the robustness to model specification of X_T. For studying robustness, we assume that ME exists with known variance σ_u and that the unknown pdf of X_T is possibly misspecified. The SEV MLE $\hat{\theta}_R$ is robust if

$$\theta_R(\sigma_u) \text{ is approximately equal to } \theta_R^* \text{ for } \sigma_u \geq 0,$$

where θ_R^* is the true parameter value that makes the correctly specified mean score function equal to zero. Therefore, the asymptotic bias in the SEV MLE of θ_R that arises due to covariate model misspecification is

$$\text{Asymptotic Bias}(\hat{\theta}_R) = \theta_R(\sigma_u) - \theta_R^*. \tag{2.5}$$

The main difficulty in finding the asymptotic bias of $\hat{\theta}_R$ is finding the analytical solution $\theta_R(\sigma_u)$ when there is no closed form of Eqn. (2.4). Therefore, to approximate the solution of $\theta(\sigma_u)$ one idea is to generate a large sample of size n and apply the large sample theory of estimation. The large sample estimation process for $\theta(\sigma_u)$ is described below.

2.2.3 Estimating Method of the SEV Model Parameters

Let $D = (Y_i, X_i)_{i=1}^n$ denote independent realizations from the ME models defined in Eqns. (1.15) and (1.16) and $D_i = (Y_i, X_i)$ be the i$^{\text{th}}$ realization. By the law of large numbers we have

$$\frac{1}{n}\sum_{i=1}^{n} S\{\theta(\sigma_u) \mid D_i\} \xrightarrow{p} E[S\{\theta(\sigma_u) \mid Y, X\}],$$

in probability when n is large. Therefore, for large sample estimation of $\theta(\sigma_u)$ we need to solve

$$\sum_{i=1}^{n} S\{\theta(\sigma_u) \mid D_i\} = 0, \tag{2.6}$$

for a large sample of size n. Eqn. (2.6) is the estimating equation for $\theta(\sigma_u)$ under the misspecified model. Let $\tilde{\theta}(\sigma_u)$ be the solution to the estimating Eqn. (2.6) for large n. We used the Newton-Raphson (N-R) iterative method to solve the estimating Eqn. (2.6) for $\theta(\sigma_u)$, where at any particular iteration (r+1) we have

$$\tilde{\theta}^{(r+1)}(\sigma_u) = \tilde{\theta}^{(r)}(\sigma_u) - [I\{\tilde{\theta}^{(r)}(\sigma_u)\}]^{-1}[S\{\tilde{\theta}^{(r)}(\sigma_u)\}], \tag{2.7}$$

where $S\{\tilde{\theta}^{(r)}(\sigma_u)\}$ is a $(\mathrm{P}\times 1)$ score vector and $I\{\tilde{\theta}^{(r)}(\sigma_u)\}$ is a $(\mathrm{P}\times\mathrm{P})$ Hessian matrix at the r$^{\text{th}}$ iteration. The j$^{\text{th}}$ element of $S\{\tilde{\theta}^{(r)}(\sigma_u)\}$ is

$$S_{\mathrm{j}}\{\tilde{\theta}^{(\mathrm{r})}(\sigma_u)\} = \sum_{\mathrm{i}=1}^{n} S_{\mathrm{j}}\{\tilde{\theta}^{(\mathrm{r})}(\sigma_u) \mid D_{\mathrm{i}}\} = \sum_{\mathrm{i}=1}^{n} \frac{\partial}{\partial\theta_{\mathrm{j}}} l\{\theta(\sigma_u) \mid D_{\mathrm{i}}\}\Big|_{\theta(\sigma_u)=\tilde{\theta}^{(\mathrm{r})}(\sigma_u)}, \tag{2.8}$$

and the (k, j)$^{\text{th}}$ element of $I\{\tilde{\theta}^{(r)}(\sigma_u)\}$ is

$$I_{\mathrm{k,j}}\{\tilde{\theta}^{(\mathrm{r})}(\sigma_u)\} = \sum_{\mathrm{i}=1}^{n} I_{\mathrm{k,j}}\{\tilde{\theta}^{(\mathrm{r})}(\sigma_u) \mid D_{\mathrm{i}}\} = \sum_{\mathrm{i}=1}^{n} \frac{\partial^2}{\partial\theta_{\mathrm{k}}\partial\theta_{\mathrm{j}}} l\{\theta(\sigma_u) \mid D_{\mathrm{i}}\}\Big|_{\theta(\sigma_u)=\tilde{\theta}^{(\mathrm{r})}(\sigma_u)}, \tag{2.9}$$

for j, k $= 1, 2, \ldots, \mathrm{P}$. Evaluation of the terms in Eqns. (2.8)-(2.9) requires evaluation of integrals with no closed form except in the case of the linear regression model. Therefore, numerical integration is required. The estimation procedure discussed in Hossain and Gustafson (2011) [24] is as follows:

1. Generate a sample of the $D = (Y_{\mathrm{i}}, X_{\mathrm{i}})_{\mathrm{i}=1}^{n}$ using Eqns. (1.15) and (1.16);

2. Integrate out the X_T numerically in order to evaluate $S\{\theta(\sigma_u) \mid D_{\mathrm{i}}\}$ and $I\{\theta(\sigma_u) \mid D_{\mathrm{i}}\}$;

3. Use the N-R iterative method discussed in the Eqn. (2.7) to find $\tilde{\theta}(\sigma_u)$.

Large Sample Bias of MLE of θ_R : Let $\tilde{\theta}(\sigma_u)$ be the numerically approximating $\theta(\sigma_u)$ of Eqn. (2.6) for large n. The large sample bias in the SEV MLE $\hat{\theta}_R$ under covariate model misspecification is

$$\text{Bias}(\hat{\theta}_R) = \tilde{\theta}_R(\sigma_u) - \theta_R^*,$$

which is a numerical approximation of the asymptotic bias defined in Eqn. (2.5). Therefore, the SEV MLE for θ_R, i.e., $\hat{\theta}_R$ is robust if

$$\tilde{\theta}_R(\sigma_u) \text{ approximately equal to } \theta_R^* \text{ for } \sigma_u \geq 0.$$

2.2.4 Example: Assessing Bias due to Covariate Model Misspecification in a Simple Linear Model

Huang, Stefanski, and Davidian (2006) [26] investigated the bias in the estimates of the parameters in a simple linear model due to covariate model misspecification. We have reproduced their results to check our bias estimation procedure for the SEV model. The linear model specifies Y as a function of X_T,

$$Y = \beta_o + \beta_1 X_T + \epsilon,$$

with β_0 and β_1 are the intercept and slope parameters, respectively. Suppose, X_T is measured with the classical ME model,

$$X = X_T + U.$$

Assume that ϵ and U are $N(0, \sigma_e^2)$ and $N(0, \sigma_u^2)$, respectively.

1. **Response Model:** The Normal pdf of Y given X_T is

$$f_{Y|X_T}(y \mid x_T; \theta_R) = \frac{1}{\sigma_e\sqrt{2\pi}} e^{-\left(y-\eta(x^T;\beta_0,\beta_1)\right)^2/2\sigma_e^2}, \tag{2.10}$$

   where $\eta(x^T; \beta_0, \beta_1) = \beta_o + \beta_1 x_T$ and $\theta_R = (\beta_o, \beta_1, \sigma_e^2)$.

2. **ME Model:** The Normal pdf of X given X_T is

$$f_{X|X_T}(x \mid x_T; \sigma_u) = \frac{1}{\sigma_u\sqrt{2\pi}} e^{-(x-x_T)^2/2\sigma_u^2}. \tag{2.11}$$

3. **Covariate Model:** The assumed Normal pdf of X_T is

$$f_{X_T}^A(x_T, \theta_E) = \frac{1}{\mu\sqrt{2\pi}} e^{-(x_T-\mu)^2/2\mu^2}. \tag{2.12}$$

   where the parameter, $\theta_E = \mu$. The mean, $\mathrm{E}(X_T) = \mu$, variance, $\mathrm{V}(X_T) = \mu^2$, and the coefficient of variation, $\mathrm{CV}(X_T)= 1$.

To estimate the parameters of interest (β_0 and β_1) we follow the estimating method discussed in Section 2.2.3. We used the simulation procedure described below to compute the large sample bias in the SEV MLE of β_0 and β_1.

A large random sample (n = 50,000) of responses Y were generated with parameters fixed at $\beta_0 = 0, \beta_1 = 1$ and $\sigma_e = 1$. Three true distributions of X_T were investigated: $N(0.5, 1), N(1, 1)$, and $N(1.5, 1)$. When the true distribution of X_T is $N(1, 1)$ then the assumed model $N(\mu, \mu^2)$ is correctly specified, while for the other two cases the assumed model is incorrect. Figure 2.1 displays the bias in the estimates of β_0 and β_1 against σ_u for three true distributions of X_T, two of which were misspecified as $N(\mu, \mu^2)$. The misspecification and lack of flexibility in modelling X_T results in bias in the estimates of β_0 and β_1 that increase in magnitude with σ_u. Virtually identical results were provided by Huang, Stefanski and Davidian (2006) [26].

2.3 SEV VonB Model

Let Y be the measured length of a fish, Y_T be the unknown true length, X_T be the unobserved true age, and X be the observed age. For simplicity we assume that measurements of both length and age of fish are continuous similar to Cope and Punt (2007) [15]. The VonB growth model specifies Y_T as a function of X_T,

$$Y_T(X_T; L_\infty, a_o, k) = L_\infty(1 - e^{-k(X_T - a_o)}).$$

The parameter L_∞ is the asymptotic length (as $X_T \to \infty$) at which the growth rate is zero, k is the growth rate parameter, and $a_o < 0$ is the theoretical age at which a fish has zero length. We assume that the observed length (Y) and age of fish (X) have independent multiplicative MEs,

$$Y = Y_T(X_T; L_\infty, a_o, k)\ e^{\epsilon}, \tag{2.13}$$

$$X = X_T\ e^{U}, \tag{2.14}$$

where ϵ is $N(0, \sigma_e^2)$ and U is $N(0, \sigma_u^2)$. Since U is $N(0, \sigma_u^2)$, e^U is a Lognormal$(0, \sigma_u)$ distribution. The coefficient of variation of a Lognormal$(0, \sigma_u)$ distribution is $\sqrt{e^{\sigma_u^2} - 1}$, which is approximately σ_u when σ_u is small. Therefore, e.g. $\sigma_u = 0.3$ will be regarded as 30 percent ME variance in age.

We assume multiplicative ME in length because in practice errors will be smaller for small fish compared to larger sizes. Errors in age will also usually increase with age because it is more difficult to count annual otolith growth increments for older fish; therefore, the multiplicative error in age is valid as described in Cope and Punt (2007) [15] and Cadigan and Campana (2016)[8].

1. **Length Model:** The Lognormal pdf of Y given X_T is

$$f_{Y|X_T}(y \mid x_T; \theta_R) = \frac{1}{y\sigma_e\sqrt{2\pi}} e^{-(\log(y)-\eta(x_T;L_\infty,k,a_o))^2 / 2\sigma_e^2}, \tag{2.15}$$

   where $\eta(x_T; L_\infty, k, a_o) = \log\{L_\infty(1 - e^{-k(x_T - a_o)})\}$ and $\theta_R = (L_\infty, k, a_o, \sigma_e^2)$.

2. **ME Model:** The Lognormal pdf of X given X_T is

$$f_{X|X_T}(x \mid x_T; \sigma_u) = \frac{1}{x\sigma_u\sqrt{2\pi}} e^{-(\log(x)-\log(x_T))^2 / 2\sigma_u^2}. \tag{2.16}$$

3. **Age Model:** The assumed pdf of X_T is Gamma,

$$f^A_{X_T}(x_T; \theta_E) = \frac{1}{\beta^\alpha \; \Gamma(\alpha)} \; x_T^{\alpha-1} \; e^{-\frac{x_T}{\beta}}, \tag{2.17}$$

   where the parameter vector $\theta_E = (\alpha, \beta)$. The mean is $\mathrm{E}(X_T) = \alpha\beta$ and the variance is $\mathrm{V}(X_T) = \alpha\beta^2$.

In this study, we called an SEV VonB model with a simple Gamma distribution for unobserved ages a SEV VonB Gamma model. For simplicity, we assumed σ_e was known. To estimate the parameters, i.e. $\theta = (L_\infty, k, a_o, \alpha, \beta)$, we follow the estimating method discussed in Section 2.2.3. The score vector and Hessian matrix of θ for this model are provided in Appendix A. The R procedure “integrate” is used to integrate out the X_T numerically in order to evaluate the score vector and Hessian matrix of θ.

In the next section, we conduct a simulation study to investigate the impact of mis-specifying the true unobserved age distribution as Gamma when it is actually something else, like a Lognormal or mixture of distributions.

2.3.1 Simulation Design and Settings

The response Y is generated assuming a VonB growth model (Eqn. 2.13) with the parameters fixed at $L_\infty = 120, k = 0.2, a_o = -0.1$ and $\sigma_e = 0.1$. We assume that only the true unobserved age distribution is misspecified. For all simulation setups we consider σ_u varies from 0 to 0.3. When $\sigma_u = 0$, there is no ME in age. We use a sample of size $n = 50{,}000$ for studying the bias in the estimates of $\theta_R = (L_\infty, k, a_o)$.

2.3.2 Simulation Design 1: Lognormal versus Gamma Distribution for True Age

The true distribution of unobserved age X_T is generated from a Lognormal(μ, σ) distribution with pdf

$$f^T_{X_T}(x_T; \Theta_E) = \frac{1}{x_T\ \sigma\sqrt{2\pi}} e^{-(\log(x_T)-\mu)^2/2\sigma^2}. \tag{2.18}$$

The parameter vector includes $\Theta_E = (\mu, \sigma)$. The mean is $\mathrm{E}(X_T) = e^{\mu+\frac{\sigma^2}{2}}$, the variance is $\mathrm{V}(X_T) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$, the coefficient of variation is $\mathrm{CV}(X_T) = \sqrt{e^{\sigma^2} - 1}$, and the skewness is $\mathrm{Sk} = (e^{\sigma^2} + 2)\sqrt{e^{\sigma^2} - 1}$. We consider different degrees of skewness and heavy tailedness for the distribution of X_T. Different simulation settings for the true unobserved age distribution of fish are:

1. $\mathrm{E}(X_T) = 4$, $\mathrm{CV}(X_T) = 0.5$; therefore, $\mu = 1.275$, $\sigma = 0.4723$ and $\mathrm{Sk} = 1.6$;

2. $\mathrm{E}(X_T) = 7$, $\mathrm{CV}(X_T) = 0.5$; therefore, $\mu = 1.834$, $\sigma = 0.4723$ and $\mathrm{Sk} = 1.6$;

3. $\mathrm{E}(X_T) = 10$, $\mathrm{CV}(X_T) = 0.5$; therefore, $\mu = 2.2$, $\sigma = 0.4723$ and $\mathrm{Sk} = 1.6$;

4. $\mathrm{E}(X_T) = 4$, $\mathrm{CV}(X_T) = 1.5$; therefore, $\mu = 0.7946$, $\sigma = 1.085$ and $\mathrm{Sk} = 7.9$;

5. $\mathrm{E}(X_T) = 7$, $\mathrm{CV}(X_T) = 1.5$; therefore, $\mu = 1.35$, $\sigma = 1.085$ and $\mathrm{Sk} = 7.9$;

6. $\mathrm{E}(X_T) = 10$, $\mathrm{CV}(X_T) = 1.5$; therefore, $\mu = 1.713$, $\sigma = 1.085$ and $\mathrm{Sk} = 7.9$.

Figure 2.2 illustrates the simulated true ages. Cases 1-3 represent situations where the Lognormal distributions have less heavy tails and cases 4-6 have heavier tails. Cases 1-3 and 4-6 are contrasting situations where the levels of skewness are different for the same corresponding mean age. The Gamma distribution is expected to pick up the shape of the true unobserved age distributions in cases like 4-6. Cases 1-3 are considered to examine whether light tails in the age distribution are a factor contributing to estimator bias.

As shown in Figure 2.3, in case of correct specification, i.e., when the true distribution of X_T follows Gamma$(\alpha = 7, \beta = 1)$, there is little bias in the estimators of L_∞, k and a_o. However, mis-specifying the distribution of X_T as Gamma (when it was actually Lognormal) may result in large bias in estimators of L_∞, k and a_o. When there is no ME in age, i.e. $\sigma_u = 0$, all the estimated values are about the same as the true values of the corresponding parameters irrespective of the different levels of skewness in the true age distribution. The estimates of the VonB growth parameters are fairly close to their true values when σ_u is less than 5 percent. Hence, misspecification of the true age distribution does not cause much bias when ME in age is low. However, as σ_u increases the bias of these estimators increases substantially in cases like 1-2 where skewness is comparatively small. Notice that across all the situations L_∞ is overestimated while k and a_o are underestimated. The bias in L_∞ is negatively correlated with the bias in k, which is expected as described in (e.g. Quinn and Deriso, 1999 [51]). Interestingly, when skewness increases to 7.9 from 1.6, the bias in L_∞, k and a_o becomes low.

It is well known that a broad range of ages is required to estimate VonB model parameters reliably. This should include young fish whose growth rates provide more direct information about k, and old fish that provide information about the asymptotic size, L_∞. Otherwise, there may be confounding between k and L_∞. To further explore how the distribution of ages may affect bias, we computed the proportion of large fish ($Y > 0.95L_\infty$) and investigated the relationship of this proportion with bias. Let $X_{0.95L_\infty}$ be the age at which the length of a fish is 95 percent of $L_\infty = 120$. Let Pr denote the probability of the event ($X >X_{0.95L_\infty}$), which is an indicator of the proportion of old aged fish in the population. Table 2.1 demonstrates that when the skewness of the Lognormal age distributions is the same as the proportion of old aged population Pr increases the bias in the estimates of L_∞, k and a_o decreases. For example, in cases like 1-3 where the skewness is 1.6, the proportion of old aged fish increases from 0.001 to 0.14, and the bias in the estimates of L_∞ and k decreases substantially as Pr increases. Therefore, the percentage of old aged fish in the population seems to be a factor contributing to the bias in the estimates of L_∞ and k.

The results from Table 2.2 are based on two cases where the true unobserved age distributions are Lognormal with the same mean, 4. The proportion of old aged fish in the population is 0.001 for both cases, however, the level of skewness is different. Table 2.2 indicates that when the skewness increases from 1.63 to 2.67 the change in bias in the estimates is substantial. It implies that when the percentage of old aged fish in the population is small the bias in the estimates are high irrespective of the level of skewness increases.

The results from Table 2.3 are based on cases where the percentage of old aged fish is 4 percent, and skewness increases from 7.86 to 13.97. The bias in the estimates of L_∞, k, and a_o are low as skewness increases. It implies that when the percentage of old aged fish in the population is large the bias in the estimates are low irrespective of the skewness increases.

Therefore, we conclude that the estimates of L_∞, k and a_o are not only affected by a joint effect of ME in age and the nature of the misspecification of the true unobserved age distribution, but they are also affected by the percentage of old fish in the population that are near their asymptotic length.

2.3.3 Simulation Design 2: Mixture versus Gamma Distribution for True Age

Age-distributions of fish in practice is complicated and often multi-modal. For example, if reproduction was really good 5 years ago and 2 years ago, but not otherwise, then the age distribution this year will have peaks at 2 and 5. The true age distribution of a population is the result of complex temporal variability in previous reproduction and survival rates (e.g. Kitakado, 2000 [28]). In order to examine the sensitivity to Gamma model misspecification in a SEV VonB model, we investigate the following mixture distributions for true unobserved ages.

Three-Gamma Mixture Distribution: This distribution is refereed to as the "TriGamma" case which corresponds to generating X_T from a well-separated three-Gamma mixture distribution with pdf

$$f^T_{X_T}(x_T; \Theta_E) = \sum_{g=1}^{3} p_g \; f_g(x_T; \alpha_g, \beta_g), \tag{2.19}$$

where $f_g(x_T; \alpha_g, \beta_g)$ denotes the Gamma pdf which is defined in Eqn. (2.17) and $\sum_{g=1}^{3} p_g = 1$, $p_g \geq 0$ for $g = 1, 2, 3$. Let $\alpha = (\alpha_1, \alpha_2, \alpha_3)$; $\beta = (\beta_1, \beta_2, \beta_3)$; therefore, the parameter vector, $\Theta_E = (\text{p}, \alpha, \beta)$; where $\text{p} = (p_1, p_2, p_3)$. To generate X_T from the three-Gamma mixture distribution we consider $p = (0.2, 0.3, 0.5)$; $\alpha = (20, 50, 90)$; and $\beta = (0.1, 0.1, 0.1)$. The means of three independent Gamma distributions are 2, 5 and 9, respectively; and the variances are 0.2, 0.5 and 0.9. The mixture distribution is illustrated in Figure 2.4 (First Panel) and has its highest peak at age 10.

Ten-Gamma Mixture Distribution: This distribution is referred to as the "MultiGamma" case where X_T is generated from a mixture of ten Gamma distributions with pdf (defined in Eqn. 2.19), where $g = 1, 2, ..., 10$ with different means but the same CV(X_T) = 0.0025 (see Figure 2.4).

Three-Truncated Normal Mixture Distribution: This distribution is refereed to as the "TriNormal" case which corresponds to generating X_T from a mixture of three-truncated Normals with the pdf

$$f^T_{X_T}(x_T; \Theta_E) = \sum_{g=1}^{3} p_g \; f_g(x_T; \mu_g, \sigma_g), \tag{2.20}$$

where $f_g(x_T; \mu_g, \sigma_g)$ denotes the truncated Normal density, which is

$$f_g(x_T; \mu_g, \sigma_g) = \frac{\phi_g(\frac{x_T - \mu_g}{\sigma_g})}{\sigma_g(1 - \Phi_g(\frac{-\mu_g}{\sigma_g}))},$$

where $\phi(.)$ is the standard Normal density and $\Phi(.)$ is its cumulative Normal distribution function. Let $\mu = (\mu_1, \mu_2, \mu_3)$; $\sigma = (\sigma_1, \sigma_2, \sigma_3)$; furthermore, the parameter vector, Θ_E = (p,μ, σ); where p = (p_1, p_2, p_3). To generate X_T from the three-truncated Normal mixture distribution we consider $p = (0.2, 0.3, 0.5)$; $\mu = (2, 5, 9)$; and $\sigma = (0.2, 0.5, 0.9)$. Figure 2.4 shows that the mixture of three truncated Normal age distributions represents a younger age population since the distribution has its highest peak at age 2.

Results corresponding to TriGamma, MultiGamma, and TriNormal cases are presented in Figure 2.5. The estimates are very close to the true values when the σ_u is less than or equal to 0.1, but when σ_u reached 0.3 the bias was large. In these simulation designs L_∞ was underestimated while a_o and k were overestimated. This is in contrast to the situation where the Lognormal distribution of the true age was misspecified as a Gamma distribution.

2.4 Summary

Our simulation results show that

- The SEV estimates of L_∞, k and a_o are not robust when the true distribution of X_T is misspecified as the simple Gamma distribution.

- When ME in age is small, e.g. $\sigma_u \leq 0.05$; the bias of these estimates are small even when the distribution of true ages is misspecified.

- However, as σ_u increases the bias increases substantially.

In the next chapter, we propose an SEV VonB growth model that involves a mixture model component for the distribution of unobserved ages that is more robust to misspecification of this component of the model.

2.5 Figures

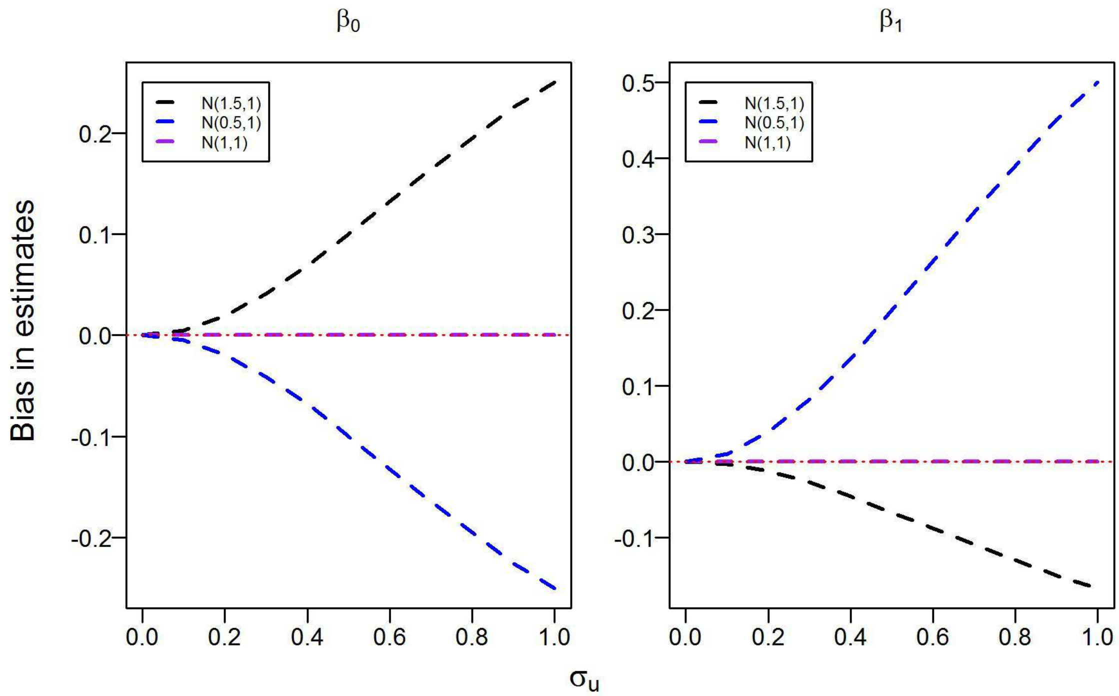

Figure 2.1: Sensitivity analysis of large sample bias in the estimates of β_0 and β_1 based on the SEV linear model where two true covariate Normal distributions, i.e., Normal(1.5,1) and Normal(0.5,1) misspecified as Normal(μ, μ^2).

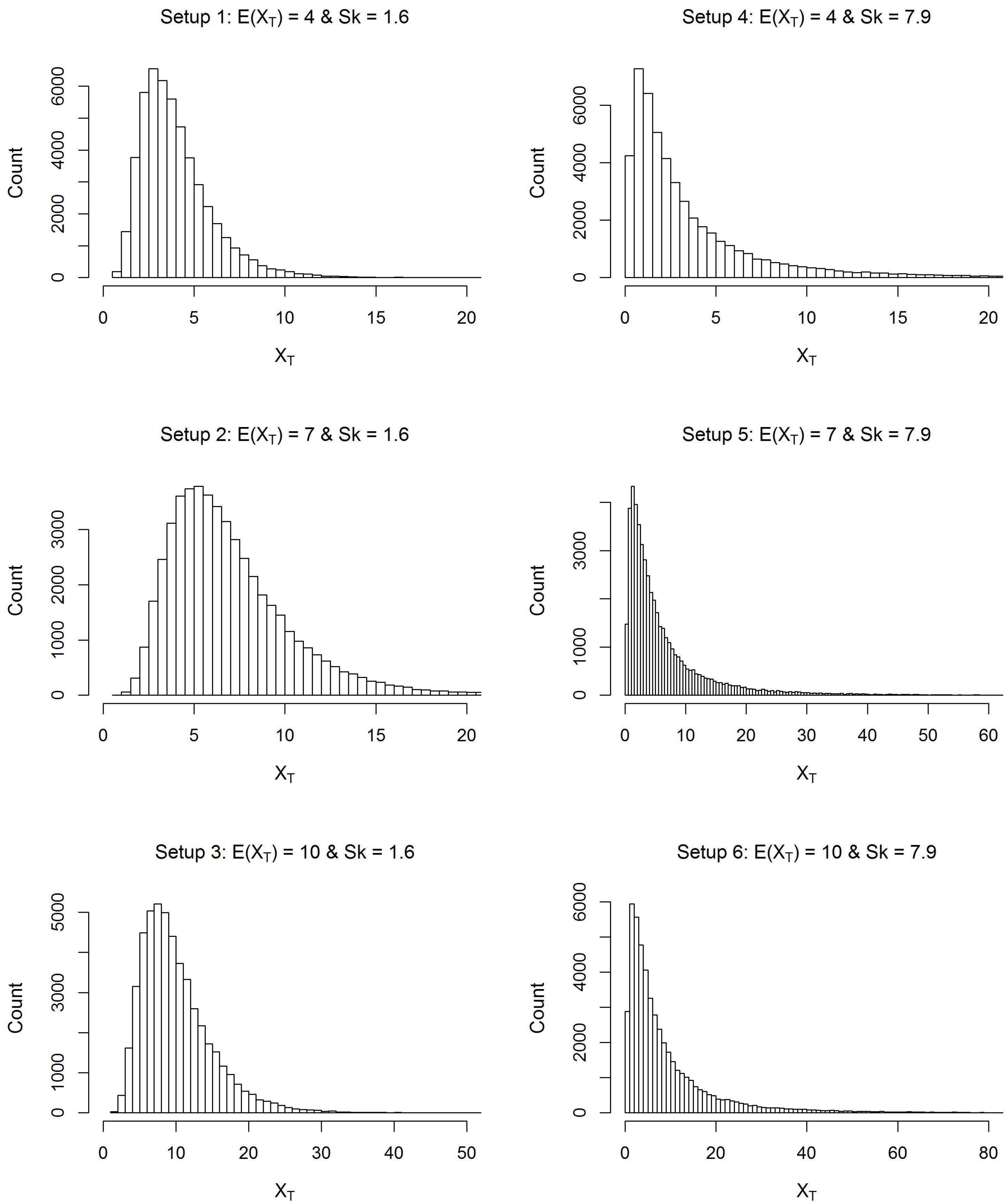

Figure 2.2: True unobserved age (X_T) distribution is Lognormal with mean, $E(X_T)$ and skewness, Sk.

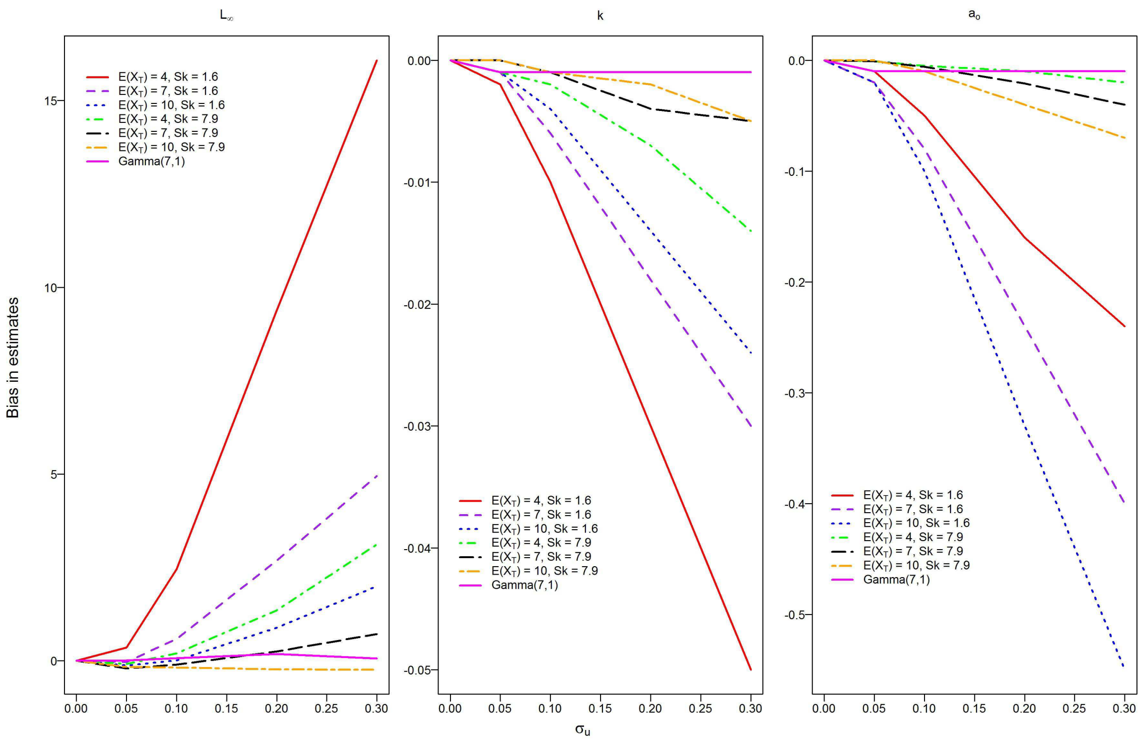

Figure 2.3: Sensitivity analysis of large sample bias in the estimates of L_∞, k, and a_o based on the SEV VonB Gamma model. Results were based on simulating data from a correctly specified Gamma distribution and six misspecified Lognormal distributions with different means, E(X_T) and skewness, Sk.

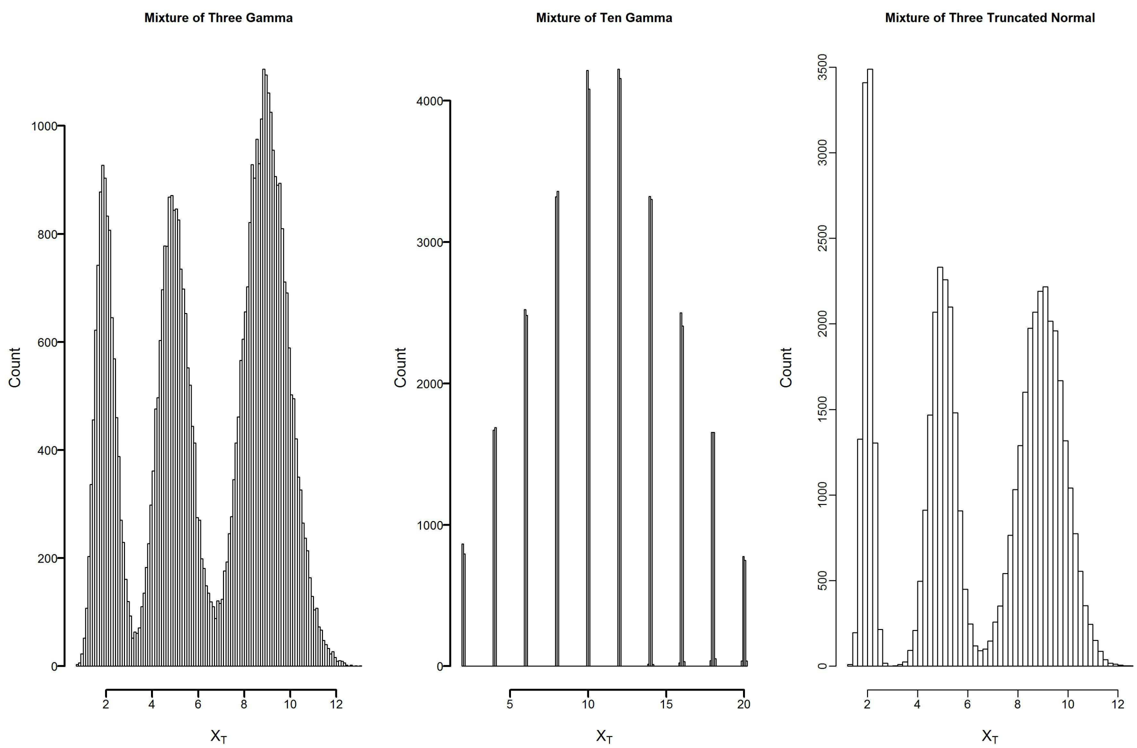

Figure 2.4: True unobserved age (X_T) mixture Gamma and truncated Normal distributions.

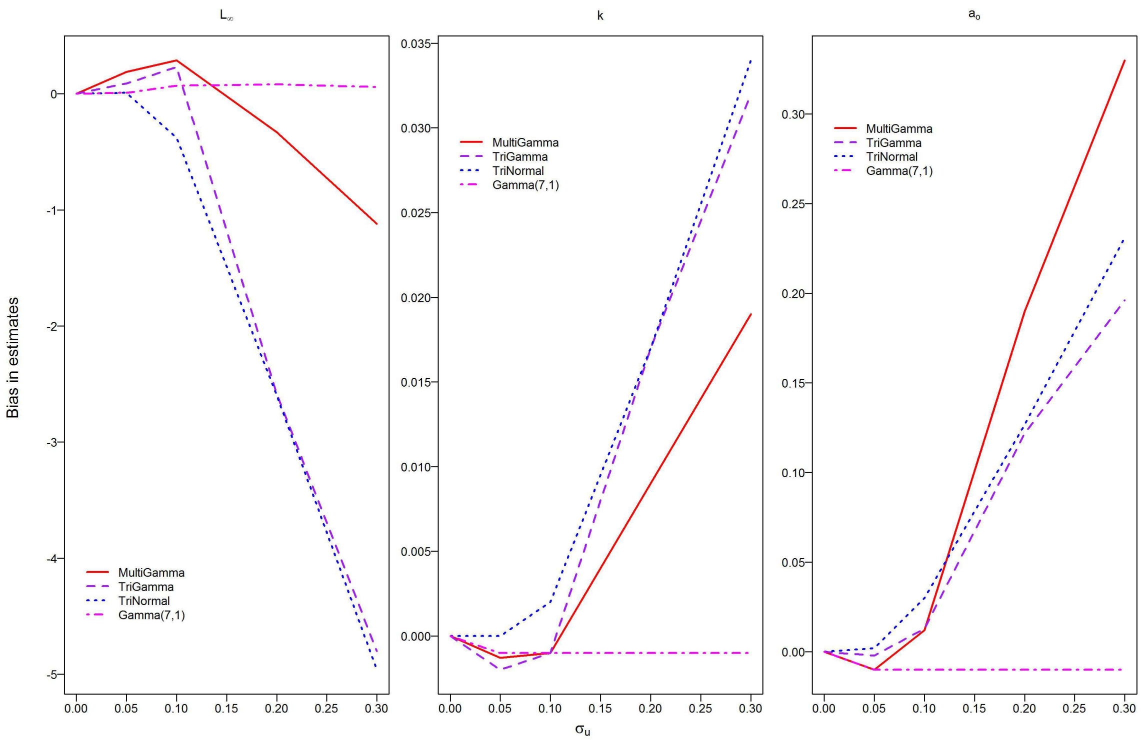

Figure 2.5: Sensitivity analysis of large sample bias in the estimates of L_∞, k, and a_o based on the SEV VonB Gamma model. Results were based on simulating data from a correctly specified Gamma distribution and three misspecified mixture distributions.

2.6 Tables

Table 2.1: Estimated values of L_∞, k, and a_o based on the SEV VonB Gamma model, versus the proportion of old aged fish in the population (Pr) and skewness of the distribution.

Simulation Setup	Skewness	Pr	$\tilde{L}_\infty$	$\tilde{k}$	$\tilde{a}_o$
1	1.6	0.001	136.08	0.154	-0.34
2	1.6	0.03	124.95	0.170	-0.49
3	1.6	0.14	122.00	0.18	-0.65
4	7.9	0.04	123.00	0.186	-0.12
5	7.9	0.10	120.72	0.195	-0.14
6	7.9	0.18	120.07	0.197	-0.17

Table 2.2: SEV VonB estimates of L_∞, k, and a_o, versus skewness of the distribution at Pr = 0.001.

Skewness	$\tilde{L}_\infty$	$\tilde{k}$	$\tilde{a}_o$
1.63	136.08	0.15	-0.30
2.67	128.34	0.17	-0.22

Table 2.3: SEV VonB estimates of L_∞, k, and a_o, versus skewness of the distribution at Pr = 0.04.

Skewness	$\tilde{L}_\infty$	$\tilde{k}$	$\tilde{a}_o$
7.86	123.0	0.186	-0.12
13.97	122.2	0.190	-0.11

Chapter 3

Robustness of SEV VonB G-Normal Mixture Model

3.1 Introduction

In Chapter 2 we showed that if the true unobserved age distribution is misspecified as the simple Gamma distribution then the bias in the estimates of VonB growth parameters increases with the magnitude of the ME in age. This is a serious drawback of using the Gamma as the assumed unobserved age distribution in a SEV VonB model for analyzing growth data. In this chapter, we investigate a more flexible mixture Normal distribution as the assumed unobserved age distribution in an SEV VonB model to get more robust estimates of the VonB growth parameters.

3.2 Finite Mixture Models

A finite mixture model is a convex combination (i.e. a weighted sum, with non-negative weights that sum to 1) of two or more pdf's. By combining the properties of the individual pdfs, mixture models are capable of approximating more complex distributions. Mixture models have been used in many applications in statistical analysis and machine learning such as modeling, clustering, classification, and latent class and survival analysis.

A mixture model is based on the assumption that the data are sampled from a population consisting of a finite collection of subpopulations, usually of the same statistical distribution type. More specifically, an independently and identically distributed (i.i.d.) random variable W arises from a finite mixture model if for all $w \subset W$,

$$f(w;\nu) = \sum_{g=1}^{G} p_g \; f_g(w;\nu_g), \tag{3.1}$$

such that $\nu = (p_1, p_2, \ldots, p_G, \nu_1, \nu_2, \ldots, \nu_G)$, $p_g \geq 0$, and $\sum_{g=1}^{G} p_g = 1$, where p_g is the g^{th} mixing proportion, ν_g is a vector of parameters, and $f_g(w;\nu_g)$ is the g^{th} component pdf.

Let $W_1, W_2, \ldots, W_G$ denote random variables from the G component distributions, and let W denote a random variable from the mixture distribution with the density in Eqn. (3.1). Then, for any function $H(.)$ for which $E(H(W_g))$ exists, and assuming that the component densities $f_g(w;\nu_g)$ exist,

$$\begin{aligned} E(H(W)) &= \int_{-\infty}^{\infty} H(W) \sum_{g=1}^{G} p_g f_g(w;\nu_g)\ dw \\ &= \sum_{g=1}^{G} p_g \int_{-\infty}^{\infty} H(W) f_g(w;\nu_g) dw \\ &= \sum_{g=1}^{G} p_g E(H(W_g)). \end{aligned} \tag{3.2}$$

Therefore, the expectation of $H(W)$ from the mixture distribution is simply a weighted average of the $E(H(W_g))$ of the G components.

3.2.1 G-Normal Mixture Distribution

In general, we define an i.i.d. random variable W drawn from G different Normal distributions with probability p_g by specifying the component pdf as

$$f(w;\nu) = \sum_{g=1}^{G} p_g\ \phi_g(w;\mu_g,\sigma_g), \tag{3.3}$$

where $\phi_g(w;\mu_g,\sigma_g)$ denotes the Normal density with mean μ_g and standard deviation σ_g which is

$$\phi_g(w;\mu_g,\sigma_g) = \frac{1}{\sigma_g\sqrt{2\pi}} e^{-(w-\mu_g)^2/2\sigma_g^2}.$$

Furthermore, $\sum_{g=1}^{G} p_g = 1$, $p_g \geq 0$ for $g = 1, 2, \ldots, G$. Let $\mu = (\mu_1, \mu_2, \ldots, \mu_G)$ with restriction $\mu_1 < \mu_2 < \ldots < \mu_G$, and $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_G)$. In this study, we call a mixture with G Normal components a $G-$Normal mixture distribution. The parameter vector includes $\nu = (p, \mu, \sigma)$ of dimension (q×1) where p, μ, and σ are (G×1) vectors. All the probabilities but the last one will be estimated. The g^{th} component probability (p_G) can be computed from the relation $p_G = 1 - \sum_{g=1}^{G-1} p_g$.

Therefore, in ν there are q-1 = G + G + G - 1 = (3G-1) parameters to estimate.

Rationale of Ordered Mean: The ordered restriction on group means, i.e., $\mu_1 < \mu_2 < \ldots < \mu_G$, is imposed to avoid non-identifiability issues with the G-Normal mixture distribution. We explain this with a two-Normal mixture distribution. The number of parameters to estimate is five, i.e., $p_1, \mu_1, \mu_2, \sigma_1, \sigma_2$. The two-Normal mixture pdf (Eqn. 3.3) with parameter values $p_1 = 0.5, \mu_1 = 1, \mu_2 = 2, \sigma_1 = 1, \sigma_2 = 1$ is

$$N_1 = 0.5 \times \phi_1(w; 1, 1) + 0.5 \times \phi_2(w; 2, 1),$$

and with another set of parameter values, i.e., $p_1 = 0.5, \mu_1 = 2, \mu_2 = 1, \sigma_1 = 1, \sigma_2 = 1$ the pdf is

$$N_2 = 0.5 \times \phi_1(w; 2, 1) + 0.5 \times \phi_2(w; 1, 1).$$

It is evident that without order restriction on group means, $N_1 = N_2$; hence, different parameter values can define the same distribution and the mixture distribution becomes non-identifiable. However, when we impose the order restriction on the means, i.e., $\mu_1 < \mu_2$, the non-identifiability problem can be avoided since there will be no switching between $\phi_1(w; \mu_1, \sigma_1)$ and $\phi_2(w; \mu_2, \sigma_2)$.

Moments: According to Eqn. (3.2) the 1st and 2nd order moments of the $G-$Normal mixture distribution are

$$E(W) = \sum_{g=1}^{G} p_g E(W_g) = \sum_{g=1}^{G} p_g \mu_g,$$

$$E(W^2) = \sum_{g=1}^{G} p_g E(W_g^2) = \sum_{g=1}^{G} p_g (\sigma_g^2 + \mu_g^2),$$

and

$$\text{Var}(W) = E(W^2) - \{E(W)\}^2 = \sum_{g=1}^{G} p_g \sigma_g^2 + \sum_{g=1}^{G} p_g \mu_g^2 - (\sum_{g=1}^{G} p_g \mu_g)^2.$$

The variance formula indicates that the variance of the mixture is the weighted variances plus a non-negative term accounting for the (weighted) dispersion of the means. These relations highlight the potential of the G-Normal mixture distribution to display non-trivial higher-order moments such as skewness and kurtosis (fat tails) and multi-modality, even in the absence of such features within the components themselves.

3.3 SEV VonB G-Normal Mixture Model

We propose an SEV VonB model with a G-Normal mixture distribution for unobserved ages (SEV VonB G-Normal mixture model). Recall that in Chapter 2 (section 2.3) we introduced an SEV VonB model which specifies Y_T as a function of X_T,

$$Y_T(X_T; L_\infty, a_o, k) = L_\infty(1 - e^{-k(X_T - a_o)}).$$

The observed length of a fish, Y, and observed age of a fish, X, including multiplicative MEs are

$$Y = Y_T(X_T; L_\infty, a_o, k)\ e^{\epsilon},$$

$$X = X_T\ e^{U},$$

where ϵ is $N(0, \sigma_e^2)$ and U is $N(0, \sigma_u^2)$. By taking logarithm transformations we get

$$\log(Y) = \log\{Y_T(X_T; L_\infty, a_o, k)\} + \epsilon, \tag{3.4}$$

$$\log(X) = \log(X_T) + U. \tag{3.5}$$

Let $\mathcal{Y} = \log(Y)$, $\mathcal{X} = \log(X)$ and $\mathcal{X}_\mathcal{T} = \log(X_T)$. Therefore, under the misspecified age model the observed likelihood is

$$L^A(\theta \mid \mathcal{Y}, \mathcal{X}) = \int f_{\mathcal{Y}|\mathcal{X}_\mathcal{T}}(\mathcal{Y} \mid \mathcal{X}_\mathcal{T}; \theta_R)\ f_{\mathcal{X}|\mathcal{X}_\mathcal{T}}(\mathcal{X} \mid \mathcal{X}_\mathcal{T}; \sigma_u)\ f^A_{\mathcal{X}_\mathcal{T}}(\mathcal{X}_\mathcal{T}; \theta_E)\ d\mathcal{X}_\mathcal{T}, \tag{3.6}$$

where

1. **Length Model:** The pdf of $\mathcal{Y}$ given $\mathcal{X}_T$ is

$$f_{\mathcal{Y}|\mathcal{X}_\mathcal{T}}(\mathcal{Y} \mid \mathcal{X}_\mathcal{T}; \theta_R) = \frac{1}{\sigma_e\sqrt{2\pi}}\ e^{-\{\mathcal{Y}-\eta(x_T;L_\infty,k,a_o)\}^2/2\sigma_e^2}, \tag{3.7}$$

where $\eta(x_T; L_\infty, k, a_o) = \log\{L_\infty(1 - e^{-k(x_T - a_o)})\}$ and $\theta_R = (L_\infty, k, a_o, \sigma_e^2)$.

2. **ME Model:** The pdf of $\mathcal{X}$ given $\mathcal{X}_T$ is

$$f_{\mathcal{X}|\mathcal{X}_\mathcal{T}}(\mathcal{X} \mid \mathcal{X}_\mathcal{T}; \sigma_u) = \frac{1}{\sigma_u\sqrt{2\pi}}\ e^{-(\mathcal{X}-\mathcal{X}_\mathcal{T})^2/2\sigma_u^2}. \tag{3.8}$$

3. **Age Model:** The assumed pdf of $\mathcal{X}_\mathcal{T}$ is

$$f^A_{\mathcal{X}_\mathcal{T}}(\mathcal{X}_\mathcal{T}; \theta_E) = \sum_{g=1}^{G} p_g\ \phi_g(\mathcal{X}_\mathcal{T}; \mu_g, \sigma_g), \tag{3.9}$$

where $\phi_g(x; \mu_g, \sigma_g)$ denotes the Normal density with mean μ_g and standard deviation σ_g, which is

$$\phi_g(x; \mu_g, \sigma_g) = \frac{1}{\sigma_g\sqrt{2\pi}} e^{-(x-\mu_g)^2/2\sigma_g^2}.$$

Furthermore, $\sum_{g=1}^{G} p_g = 1$, $p_g \geq 0$ for $g = 1, 2, \ldots, G$. The parameter vector $\theta_E = (p,\ \mu, \sigma)$ where $p = (p_1, p_2, \ldots, p_G)$; $\mu = (\mu_1, \mu_2, \ldots, \mu_G)$ with restriction $\mu_1 < \mu_2 < \ldots < \mu_G$; and $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_G)$. All the probabilities but the last one will be estimated. The G^{th} component probability (p_G) can be computed from the relation $p_G = 1 - \sum_{g=1}^{G-1} p_g$. Therefore, in θ_E there are G - 1 + G + G = (3G-1) parameters to estimate.

Modelling the Mixing Probabilities p's: The probabilities $p_1, p_2, \ldots, p_G$ for groups $1, 2, \ldots, G$ with the property that $\sum_{g=1}^{G} p_g = 1$, $p_g \geq 0$ for $g = 1, 2, \ldots, G$. There have been numerous publications focusing on how to model the probabilities $p_1, p_2, \ldots, p_G$. Recently Francis (2014) [20] studied the Logistic Normal Multinomial distribution commonly advocated for modelling probabilities. We use the continuation-ratio logit that has been used for modelling age distributions (e.g. Kvist et al., 2000 [27]; Rindford and Lewy, 2001 [36]). We transform $p_1, p_2, \ldots, p_{G-1}$ into $\lambda_1, \lambda_2, \ldots, \lambda_{G-1}$, where $\lambda_g \in (-\infty, +\infty)$ for $g = 1, 2, \ldots, G-1$. Let λ_g be the continuation-ratio logit [Eqn. (B.2) in (Appendix B)] of p_g. We will estimate continuation-ratio logits $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{G-1})$, and then use these estimated values in Eqn. (B.1) (Appendix B) in order to estimate $p_1, p_2, \ldots, p_{G-1}$. Then, the p_G will be estimated using the relationship $p_G = 1 - \sum_{g=1}^{G-1} p_g$.

Modelling the Restriction on Group Means: The ordered mean constraint assumes that $\mu_1 < \mu_2 < \ldots < \mu_G$. We will estimate the means from the following relation

$$\mu_g = \begin{cases} \mu_1 & g = 1 \\ \mu_{g-1} + e^{\mu_g^d} & g = 2, 3, \ldots, G. \end{cases}$$

where $\mu_g^d = \mu_g - \mu_{g-1}$, $\mu_g^d \in (-\infty, \infty)$ for $g = 2, 3, \ldots, G$. Therefore, $e^{\mu_g^d} \in (0, \infty)$ for $g = 2, 3, ..., G$. We will estimate the parameters in the vector $\mu^d = (\mu_1,\ \mu_2^d, \ldots,\ \mu_G^d)$. Using these we can estimate $\mu = (\mu_1, \mu_2, \ldots, \mu_G)$ from the above relation.

Determining G: It is necessary to decide how many mixtures (the value of G) to use in the pdf (Eqn. 3.9) of the G-Normal mixture distribution to model the distribution of the unobserved age X_T. Increasing the value of G will provide more flexibility. However, there is always a price to pay in terms of efficiency of the estimates for allowing extra flexibility. According to Ma and Genton (2004) [31], G= 3 can provide enough flexibility to approximate a wide variety of densities. This is especially reasonable in ME problems, because, in practice, the true covariate distribution does not usually have many well separated modes (e.g. Hossain and Gustafson, 2009 [25]). Therefore, in this study we investigate the suitability of a Normal mixture distribution with G= 2 and G= 3 as the distribution of true age in the SEV age models.

3.3.1 Estimation of Parameters

In our proposed SEV VonB G-Normal mixture model we approximate the true unobserved age distribution by assuming a G-Normal mixture distribution. For simplicity, we assumed σ_e is known. Therefore, the parameter vector includes $\theta = (L_\infty, k, a_0, \lambda, \mu^d, \sigma)$ of dimension P$\times$1, where λ is a (G-1) $\times$ 1 vector; μ^d and σ are G $\times$ 1 vectors. Therefore, in θ there are P = 3 + G + G + G - 1 = (3G+2) parameters to estimate. The log-likelihood function for θ based on $\mathcal{Y}$ and $\mathcal{X}$ under misspecification is

$$l(\theta \mid \mathcal{Y}, \mathcal{X}) = \log\{L^A(\theta \mid \mathcal{Y}, \mathcal{X})\}.$$

The estimating equation for $\theta(\sigma_u)$ under the misspecified model is

$$\sum_{i=1}^{n} S\{\theta(\sigma_u) \mid \mathcal{Y}_i, \mathcal{X}_i\} = 0.$$

To optimize the above estimating equation for $\theta(\sigma_u)$ it is difficult to follow the N-R approach discussed in Chapter 2 (section 2.2.3). The evaluation of $S\{\tilde{\theta}(\sigma_u)\}$ and

$I\{\tilde{\theta}(\sigma_u)\}$ requires a large number of derivative calculations when the number of parameters P increases. For estimating the parameters of the proposed SEV VonB G-Normal mixture model this work increases as the number of groups, G, increases. Alternatively, the Template Model Builder (TMB) frees the statistician from the task of writing, testing and maintaining derivative code.

In the next section, we discuss the TMB package and its implementation of our proposed approach to estimate parameters.

3.4 Template Model Builder

TMB (e.g. Skaugh and Fournier, 2006 [41]; Wang, 2015 [52]) is a free and open source R package (R Core Team, 2014 [34]) that is designed for estimating complex nonlinear random effects models. One needs to define the joint log-likelihood function of the data and the random effects as a C++ template function. Other operations such as integration and calculation of the marginal score function are done in R. This package evaluates and maximizes the Laplace approximation of the marginal likelihood where the random effects are automatically integrated out. This approximation, and its derivatives, are obtained using automatic differentiation of the joint likelihood.

3.4.1 Automatic Differentiation

Automatic Differentiation (AD) (e.g. Fournier et al., 2012 [18]), also known as computational differentiation or algorithmic differentiation, is a set of techniques that numerically differentiate a function, which frees us from calculating and incorporating the derivatives. Two methods, “source transformation” and “operator overloading” are commonly used to implement automatic differentiation. C++ automatic differentiation (CppAD) (e.g. Bell, 2012 [5]) implements the operator overloading approach.

This approach is easier to implement and use than "source transformation". The TMB R package uses CppAD to provide up to third order derivatives of the joint likelihood function that the user writes in the C++ template. These derivatives are required for the Laplace approximation of the marginal likelihood.

3.4.2 Laplace Approximation

The Laplace approximation (e.g. Skaugh and Fournier, 2006 [41]) is used to approximate the intractable integral in the marginal likelihood (Eqn. 3.10). Let $y = (y_1, y_2, ..., y_n)$ be the vector of response variables, $\Lambda = (\Lambda_1, \Lambda_2, ..., \Lambda_n)$ be the vector of unobserved random effects, and let $\theta = (\theta_1, \theta_2, ..., \theta_m)$ be the vector of fixed effects parameters. Moreover, $f_\theta(y \mid \Lambda)$ denotes the conditional pdf of y given Λ, and $f_\theta(\Lambda)$ denotes the marginal pdf of random effects Λ. The marginal likelihood function for θ is defined by integrating out the random effects Λ from the joint pdf of y and Λ, $f_\theta(y \mid \Lambda) f_\theta(\Lambda)$, such that

$$L(\theta) = \int f_\theta(y \mid \Lambda) f_\theta(\Lambda) d\Lambda = \int e^{h(\Lambda,\theta)} d\Lambda, \tag{3.10}$$

where $h(\Lambda, \theta) = \log\{f_\theta(y \mid \Lambda)\} \; + \; \log\{f_\theta(\Lambda)\}$ is the joint log-likelihood. The main computational challenge is in computing the integral in Eqn. (3.10), when there is no analytical solution. TMB uses the Laplace approximation for Eqn. (3.10), which yields the marginal likelihood approximation

$$L^*(\theta) = \det\{H(\theta)\}^{-1/2} e^{h(\hat{\Lambda}(\theta),\theta)}, \tag{3.11}$$

where $\hat{\Lambda}(\theta) = \arg\max_{\Lambda}\{h(\hat{\Lambda}(\theta), \theta)\}$, $H(\theta) = \frac{\partial^2}{\partial \Lambda^2} h(\Lambda, \theta)|_{\Lambda = \hat{\Lambda}(\theta)}$, and $\det\{H(\theta)\}$ denotes the determinant of $H(\theta)$. The term $e^{h(\hat{\Lambda}(\theta),\theta)}$ in Eqn. (3.11) is a profile likelihood, which treats the random effects as nuisance parameters and θ as the parameters of interest. The Hessian, H, is evaluated by CppAD. Using the AD and Laplace approximation simplifies the parameter estimation of hierarchical models. The TMB user needs only to specify the joint log-likelihood function $h(\Lambda, \theta)$. TMB uses the Cholesky decomposition of $H(\theta)$; therefore, the Laplace approximation is well-defined only if $h(\Lambda, \theta)$ is positive definite.

In an R session, we read in the data, dynamically link the C++ function template, set up the initial values for θ, specify the random effects, and optimize the objective function. TMB automatically provides a standard error report for $\hat{\theta}$, and also any differentiable function of θ, $\phi(\theta)$ that the user specifies, by using the δ-method:

$$\mathrm{Var}(\phi(\hat{\theta})) = -\left\{ \phi'(\theta)[\frac{\partial^2\{\log L^*(\theta)\}}{\partial\theta\partial\theta'}]^{-1}\phi(\theta) \right\}\Big|_{\theta=\hat{\theta}}.$$

3.4.3 SEV VonB G-Normal Mixture Model Implementation

In this section, we describe the implementation in TMB of the SEV VonB G-Normal mixture model. The observed likelihood (Eqn. 3.6) is re-written as

$$L^A(\theta \mid \mathcal{Y}, \mathcal{X}) = \int e^{h(\mathcal{X}_{\mathcal{T}}, \theta)} d\mathcal{X}_{\mathcal{T}},$$

where $h(\mathcal{X}_{\mathcal{T}}, \theta) = \log\{f_{\mathcal{Y}|\mathcal{X}_{\mathcal{T}}}(\mathcal{Y} \mid \mathcal{X}_{\mathcal{T}}; \theta_R)\} + \log\{f_{\mathcal{X}|\mathcal{X}_{\mathcal{T}}}(\mathcal{X} \mid \mathcal{X}_{\mathcal{T}}; \sigma_u)\} + \log\{f^A_{\mathcal{X}_{\mathcal{T}}}(\mathcal{X}_{\mathcal{T}}; \theta_E)\}$, which is the joint log-likelihood. For convenience we estimate the logarithm of L_∞, k, and σ, which can take values between $(-\infty, \infty)$ whereas L_∞, k, and σ vary between $(0, \infty)$. Let

$$\theta = (\log(L_\infty), \log(k), a_0, \lambda, \mu^d, \log(\sigma)).$$

The vector of random effects are $\mathcal{X}_{\mathcal{T}}$. We first specify the negative joint likelihood function $-h(\mathcal{X}_{\mathcal{T}}, \theta)$ in the C++ template (see Appendix B). The TMB code in R for an SEV VonB G-Normal mixture model then calculates the marginal likelihood function for θ using the Laplace approximation. The final step is to optimize this objective function in R.

C++ Template Code: The user template for the negative joint log-likelihood (the file named fitmix.cpp) is given in Appendix B (section B.2). The first four lines are standard and should be the same for most models. The first line includes the TMB specific macros and functions, including dependencies such as CppAD and Eigen. The following three lines are the syntax for starting a function template, where Type is a template parameter that the compiler replaces by an AD type that is used for numerical computations. Lines 5-10 declare the vector of variables to be the same as tmb.data in the R session included at Appendix B (section B.3). For example, in line 5 DATA_VECTOR(age) declares the vector age to be the same as tmb.data$age in the R session. Lines 11-17 include the parameters to be the same as parameters in the R session. For example, line 11 PARAMETER(log_Linf) declares the scalar log_Linf to be the same as parameters$log_Linf in the R session. Line 14 PARAMETER_VECTOR(lambda) declares the vector lambda to be the same as parameters$lambda in the R session. The other scalar parameters are declared in a similar manner. Note that the user template does not distinguish between the fixed parameters and random effects but rather codes them both as parameters.

Lines 33-37 include the code for estimating λ, which are the continuation ratio logits of probabilities for groups. Lines 38-42 include the pdf of a G-normal mixture distribution. Lines 43-47 include the negative joint log-likelihood. Lines 48-53 include the template syntax which reports the parameters back to R with derivatives. For example, line 48 ADREPORT(Linf) reports the Linf back to R with derivatives. Lines 54-61 include the template syntax which report the parameters back to R. The last line 62 is standard syntax, which returns the negative joint log-likelihood. Having specified the user template it can be compiled, linked, evaluated, and optimized from within R.

TMB Code in R: The TMB code in R for SEV VonB G-Normal mixture model is given in Appendix B (section B.3). In the R code the first line loads the TMB package. The second line compiles the C++ template and the third line links it. The fourth to tenth lines include the data. Lines 11-21 include the initial values for the parameters. Notice that the names of the parameters should correspond to those in the C++ template. Lines 22 to 37 include the upper and lower bounds for the regression parameter estimates, and for the nuisance parameters. Line 38 defines "obj" containing the data, parameters, and also specifies the random effects. Lines 41-45 optimize the objective function and generate a standard report.

In the next section, we conduct a simulation study to compare the performance of the SEV VonB G-Normal mixture model and SEV VonB Gamma model in estimating the VonB growth parameters.

3.5 Simulation studies

3.5.1 Simulation Settings

The response Y is generated assuming a VonB growth model (Eqn. 2.13) with the parameters fixed at $L_\infty = 120, k = 0.2, a_o = -0.1$ and $\sigma_e = 0.1$. We assume that only the age model is misspecified. For simplicity, we assumed σ_e is known. For all simulation setups we consider σ_u varies from 0 to 0.3. We use a sample of size n = 50,000 for studying the large sample bias in the estimates of $\theta_R = (L_\infty, k, a_o)$ under the computational method using TMB.

3.5.2 Analysis Methods

We focused on the bias, absolute bias, and percentage error for L_∞, k and a_o. The bias of $\hat{L}_\infty$ is

$$\text{Bias}(\hat{L}_\infty) = \tilde{L}_\infty(\sigma_u) - L^*_\infty,$$

where L^*_∞ is the true parameter value of L_∞. The absolute bias (abias) of $\hat{L}_\infty$ is

$$\text{abias}(\hat{L}_\infty) = |\text{Bias}(\hat{L}_\infty)|.$$

The percentage error (PE) in $\hat{L}_\infty$ is

$$\text{PE}(\hat{L}_\infty) = 100 \times \frac{(\tilde{L}_\infty(\sigma_u) - L^*_\infty)}{L^*_\infty}.$$

In this manner we can also find the bias, abias, PE of the estimates of k and a_o.

3.5.3 Determining the Value of G

We consider the TriGamma case (Section 2.3.3) to decide about the number of groups G in the proposed flexible G-Normal mixture distribution. In TriGamma case the true unobserved age follows a three-Gamma mixture distribution. The true unobserved age distribution is misspecified as the two-Normal mixture (where G = 2) and the three-Normal mixture (where G = 3) distribution. We perform a sensitivity analysis by seeing how the inference for the SEV VonB parameters varies as a function of G. If they remain stable as G varies, we have evidence that the inferences are insensitive to G.

Table 3.1 demonstrates that the SEV VonB estimates of L_∞, k and a_o are fairly stable for the two-Normal mixture and the three-Normal mixture age distributions. For example, the estimates of L_∞ for the SEV VonB two-Normal mixture model are

120, 119.92 and 119.4 corresponding to low (i.e. $\sigma_u = 0.05$), substantial (i.e. $\sigma_u = 0.2$) and high (i.e. $\sigma_u = 0.3$) MEs in age, respectively. In the low, substantial and high ME in age situations, the estimates of L_∞ for the SEV VonB three-Normal mixture model are similar. An identical conclusion can be drawn for k and a_o. Therefore, our results demonstrate that the large sample bias calculations are fairly insensitive to the number of groups G. Fitting a three-Normal mixture provides more flexibility, however, the efficiencies of the estimators decrease in order to allow extra flexibility. Hence, throughout this study we investigate the suitability of the two-Normal mixture as an assumed distribution of the true unobserved age in the SEV VonB G-Normal mixture model.

3.5.4 Two-Normal Mixture Versus Gamma: Sensitivity Comparison in Case of Lognormal True Age Distribution

Results presented in Table 3.2 are based on Case 1 (Section 2.3.2), where the true unobserved age distribution is Lognormal with mean 4 and skewness 1.6. Recall that in this case the SEV VonB Gamma model performed poorly in terms of large sample bias for L_∞, k and a_o (illustrated in Figure 2.3). The results in Table 3.2 demonstrate that the SEV VonB two-Normal mixture model outperforms the SEV VonB Gamma model in reducing the bias due to model misspecification and ME in age. The estimates of L_∞ and k show less percentage error when derived using the Normal-mixture age distribution compared to the Gamma age distribution. For instance, the percentage error of the estimates of L_∞ for the Normal-mixture are 0, -0.17 and 0.83 corresponding to σ_u values of 0.05, 0.1 and 0.3, respectively. However, for the Gamma age distribution these are -0.15, 1.67 and 15.52.

3.5.5 Two-Normal Mixture Versus Gamma: Sensitivity Comparison in Case of Mixture Distribution as True Unobserved Age Distribution

In Section 2.3.3 we discussed mixtures of three Gamma distributions (TriGamma), ten Gamma distributions (MultiGamma) and three truncated Normal distributions (TriNormal) as true unobserved age distributions. Recall that Figure 2.5 illustrates that if these true distributions are misspecified as a Gamma distribution then there are substantial amounts of bias in the estimates of L_∞, k, and a_0 as the ME in age increases.

Table 3.3 demonstrates that in the TriGamma case the SEV VonB two-Normal mixture model estimators outperform the SEV VonB Gamma estimators in reducing the bias. In this case the absolute biases in the estimate of L_∞ for the two-Normal mixture are 0, 0.02 and 0.6 for low (i.e. $\sigma_u = 0.05$), substantial (i.e. $\sigma_u = 0.1$) and high (i.e. $\sigma_u = 0.3$) MEs in age, respectively. However, for the simple Gamma age distribution the absolute bias in the estimates of L_∞ are 0, 0.42 and 5.6. Therefore, in this case a substantial amounts of bias reduction occurs when mis-specifying the true age distribution as a two-Normal mixture when it is a mixture of three Gamma distributions.

Table 3.4 shows that in the TriNormal case the estimates of L_∞, k and a_o for both age distributions are very close to their corresponding true values when the ME in age is low (i.e. $\sigma_u = 0.05$). However, as the ME in age increases from low to high the magnitudes of bias increases substantially for the Gamma age distribution. However, the scenario improve when the two-Normal mixture distribution is used instead of the Gamma distribution as an assumed age distribution. For example, in the case of the growth parameter k, for the Gamma age distribution the percentage of errors are 1, 17 and 35 corresponding to low (i.e. $\sigma_u = 0.05$), substantial (i.e. $\sigma_u = 0.2$) and high (i.e. $\sigma_u = 0.3$) ME in age situations, respectively. These errors under the Normal-mixture age distribution are 0, -0.5 and -1, respectively. This indicates that the SEV VonB Normal mixture model exhibits less error as compared with the SEV VonB Gamma model.

Table 3.5 demonstrates that in the MultiGamma case the percentages of error in estimates of a_o are very large for the both Normal-mixture and Gamma unobserved age distributions. However, the Normal-mixture age distribution performs relatively well in this case. For example, when ME is 0.3 the percentage error in the estimate of a_o is -115 for the two-Normal mixture age distribution while it is -450 for the Gamma age distribution. The SEV VonB estimator of a_o is underestimated for both assumed age distributions. However, in the case of L_∞ and k the percentage of errors are very low for the SEV VonB G-Normal mixture model.

3.5.6 Two-Normal Mixture Versus Lognormal: Sensitivity Comparison in Case of Lognormal True Age Distribution

In this section, we compare the bias of the estimators based on the SEV VonB G-Normal mixture model with that of the SEV VonB estimators based on the true Lognormal age distribution used to generate the true unobserved age, X_T. Hence, in the former case the age distribution was misspecified; however, in the latter case it was correctly specified. The goal is to investigate the performance in terms of bias of using the Normal mixture as an assumed age distribution instead of using the true age distribution. For this purpose we consider Case 1 (see Section 2.3.2), where the true unobserved age distribution is Lognomal with $\mu = 1.275$ and $\sigma =$

0.4723. The results presented in Table 3.6 demonstrate that the two-Normal mixture performs about the same in terms of bias in the estimates of L_∞, k and a_o compared with the true Lognormal age distribution. In the case of low ME, the two-Normal mixture performs similarly to the true Lognormal age model, particularly, in terms of absolute bias and percentage error in estimates. However, in the case of high ME the performance of the Normal mixture model deteriorates slightly as demonstrated by the absolute bias of the SEV VonB estimates for the assumed Normal mixture age distribution. Therefore, the SEV VonB G-Normal mixture model is found to provide adequate bias reduction when compared with the true age model in the simulation setting investigated.

3.6 Summary

Our simulation results show that

- The proposed G-Normal mixture distribution for unobserved ages performed well in all the situations considered as compared to the Gamma age distribution.

- The SEV VonB G-Normal mixture model provides adequate large sample bias reduction due to ME in age and model misspecification.

In the next chapter, we extend the SEV VonB model to account for between-individual variation in growth that appears because individuals achieve different asymptotic sizes.

3.7 Tables

Table 3.1: Sensitivity to model misspecification: The true unobserved age distribution is a mixture of three Gamma distributions, which is misspecified as the two-Normal mixture (G = 2) and the three-Normal mixture (G = 3) distributions. Results for estimated values, and standard error(SE) for L_∞, k and a_o.

	$\sigma_u = 0.05$		$\sigma_u = 0.1$		$\sigma_u = 0.2$		$\sigma_u = 0.3$	
Estimators	G=2	G= 3	G= 2	G= 3	G= 2	G= 3	G= 2	G= 3
$\tilde{L}_\infty$	120	119.92	120.02	120.1	119.92	119.5	119.4	118.42
SE	0.27	0.275	0.3	0.29	0.37	0.39	0.51	0.52
$\tilde{k}$	0.2	0.2	0.2	0.198	0.2	0.205	0.198	0.21
SE	0.0012	0.001	0.0013	0.008	0.001	0.002	0.002	0.002
$\tilde{a}_o$	-0.10	-0.099	-0.097	-0.104	-0.089	-0.08	-0.086	-0.09
SE	0.007	0.0077	0.008	0.0098	0.011	0.015	0.012	.015

Table 3.2: Sensitivity to model misspecification: The true unobserved age distribution is a Lognormal with $\mu = 1.275$ and $\sigma = 0.4723$, which is misspecified as the two-Normal mixture and the Gamma distributions. Results for absolute bias(abias), percentage error(PE), and standard error(SE) for L_∞, k and a_o.

	$\sigma_u = 0.05$		$\sigma_u = 0.1$		$\sigma_u = 0.2$		$\sigma_u = 0.3$	
Estimators	Mixture	Gamma	Mixture	Gamma	Mixture	Gamma	Mixture	Gamma
$\tilde{L}_\infty$	120	119.82	119.8	122.01	120.5	128.8	121	138.63
abias	0	0.18	0.2	2.01	0.5	8.8	1	18.63
PE	0	-0.15	-0.17	1.67	0.041	7.33	0.83	15.52
SE	0.42	0.43	0.45	0.49	0.41	0.77	0.86	1.15
$\tilde{k}$	0.2	0.201	0.2	0.1934	0.2	0.18	0.206	0.162
abias	0	0.001	0	0.0066	0	0.02	0.006	0.038
PE	0	-0.5	0	-3.3	0	-10	3	-19
SE	0.001	0.0015	0.0015	0.0013	0.002	0.002	0.003	0.002
$\tilde{a}_o$	-0.1	-0.1	-0.103	-0.147	-0.096	-0.25	-0.106	-0.35
abias	0	0	0.003	0.047	0.004	0.15	0.006	0.25
PE	0	0	-1	47	-5	150	3	250
SE	0.006	0.007	0.008	0.008	0.011	0.01	0.016	0.012

Table 3.3: Sensitivity to model misspecification: The true unobserved age distribution is a mixture of three Gamma distributions, which is misspecified as the two-Normal mixture and the Gamma distributions. Results for absolute bias(abias), percentage error(PE), and standard error(SE) for L_∞, k and a_o.

	$\sigma_u = 0.05$		$\sigma_u = 0.1$		$\sigma_u = 0.2$		$\sigma_u = 0.3$	
Estimators	Mixture	Gamma	Mixture	Gamma	Mixture	Gamma	Mixture	Gamma
$\tilde{L}_\infty$	120	120	120.02	119.58	119.92	117.1	119.4	114.4
abias	0	0	0.02	0.42	0.08	2.9	0.6	5.6
PE	0	0	0.016	-0.35	-0.066	-2.41	-0.5	-4.67
SE	0.27	0.27	0.30	0.28	0.37	0.27	0.51	0.25
$\tilde{k}$	0.2	0.2	0.2	0.205	0.2	0.227	0.198	0.26
abias	0	0	0	0.05	0	0.027	0.002	0.06
PE	0	0	0	2.5	0	13.5	-1	30
SE	0.0012	0.001	0.0013	0.001	0.001	0.001	0.002	0.002
$\tilde{a}_o$	-0.1	-0.085	-0.097	-0.06	-0.089	0.03	-0.086	0.105
abias	0	0.015	0.003	0.04	0.011	0.13	0.014	0.205
PE	0	-15	-1	-40	-11	130	-14	-205
SE	0.007	0.0075	0.008	0.008	0.012	0.001	0.012	.0109

Table 3.4: Sensitivity to model misspecification: The true unobserved age distribution is a mixture of three truncated Normal distributions, which is misspecified as the two-Normal mixture and the Gamma distributions. Results for absolute bias(abias), percentage error(PE), and standard error(SE) for L_∞, k and a_o.

	$\sigma_u = 0.05$		$\sigma_u = 0.1$		$\sigma_u = 0.2$		$\sigma_u = 0.3$	
Estimators	Mixture	Gamma	Mixture	Gamma	Mixture	Gamma	Mixture	Gamma
$\tilde{L}_\infty$	119.97	119.8	119.6	119.43	120.45	116.15	121.15	113.38
abias	0.03	0.2	0.4	0.57	0.45	3.85	1.15	6.62
PE	-0.03	-0.167	-0.33	-0.475	0.375	-3.2	0.95	-5.51
SE	0.32	0.31	0.33	0.31	0.46	0.28	0.59	0.25
$\tilde{k}$	0.2	0.202	0.203	0.206	0.199	0.234	0.198	0.27
abias	0	0.002	0.003	0.006	0.001	0.034	0.002	0.07
PE	0	1	1.5	3	-0.5	17	-1	35
SE	0.0015	0.0014	0.001	0.001	0.002	0.001	0.002	0.002
$\tilde{a}_o$	-0.098	-0.082	-0.09	-0.06	-0.11	0.084	-0.13	0.19
abias	0.002	0.018	0.01	0.04	0.01	0.184	0.013	0.29
PE	-2	-18	-10	-40	10	-184	30	-290
SE	0.01	0.0097	0.011	0.01	0.015	0.011	0.02	0.012

Table 3.5: Sensitivity to model misspecification: The true unobserved age distribution is a mixture of ten Gamma distributions, which is misspecified as the two-Normal mixture and Gamma distributions. Results for absolute bias(abias), percentage error(PE), and standard error(SE) for L_∞, k and a_o.

	$\sigma_u = 0.05$		$\sigma_u = 0.1$		$\sigma_u = 0.2$		$\sigma_u = 0.3$	
Estimators	Mixture	Gamma	Mixture	Gamma	Mixture	Gamma	Mixture	Gamma
$\tilde{L}_\infty$	120	120	120.09	119.83	119.7	119.35	119.3	118.03
abias	0	0	0.09	0.17	0.3	0.65	0.7	1.97
PE	0	0	0.067	-0.141	-0.208	-0.541	-0.5	-1.641
SE	0.14	0.142	0.11	0.14	0.17	0.15	0.19	0.19
$\tilde{k}$	0.2	0.2	0.199	0.206	0.2	0.22	0.21	0.25
abias	0	0	0.001	0	0	0.02	0.01	0.05
PE	0.0	0	-0.5	3	0	10	5	25
SE	0.001	0.001	0.001	0.0011	0.002	0.0015	0.0019	0.002
$\tilde{a}_o$	-0.1	-0.085	-0.102	-0.093	-0.027	0.115	0.015	0.35
abias	0	0.015	0.002	0.007	0.073	0.215	0.115	0.45
PE	0	-15	2	-7	-73	-215	-115	-450
SE	0.012	0.013	0.014	0.014	0.019	0.0178	0.025	0.021

Table 3.6: Comparison of the two-mixture Normal age distributions with the true age Lognormal distribution. Results for absolute bias(abias), percentage error(PE), and standard error(SE) for L_∞, k and a_o.

	$\sigma_u = 0.05$		$\sigma_u = 0.1$		$\sigma_u = 0.2$		$\sigma_u = 0.3$	
Estimators	Mixture	Lognormal	Mixture	Lognormal	Mixture	Lognormal	Mixture	Lognormal
$\tilde{L}_\infty$	120	119.99	119.8	120.02	120.5	119.9	121	120.4
abias	0	0.01	0.2	0.02	0.5	0.1	1	0.4
PE	0	0	-0.167	0.0167	0.416	-0.083	0.83	0.34
SE	0.42	0.41	0.45	0.46	0.41	0.54	0.86	0.63
$\tilde{k}$	0.2	0.2	0.2	0.202	0.2	0.2	0.206	0.205
abias	0	0	0	0.002	0	0	0.006	0.005
PE	0	0	0	1	0	0	3	2.5
SE	0.001	0.0014	0.0015	0.0013	0.002	0.002	0.003	0.002
$\tilde{a}_o$	-0.1	-0.11	-0.103	-0.097	-0.096	-0.097	-0.106	-0.1
abias	0	0.01	0.003	0.003	0.004	0.003	0.006	0
PE	0	10	3	-3	-4	-3	6	0
SE	0.006	0.007	0.008	0.007	0.011	0.022	0.016	0.011

Chapter 4

Robustness of SEV VonB Model Including Between-Individual Variation in Growth

4.1 Introduction

In Chapter 3 we demonstrated that if the true unobserved age distribution is misspecified as the proposed flexible G-Normal mixture distributions then the estimates of the SEV VonB growth parameters are fairly close to their corresponding true values even as the magnitude (i.e. variance) of the ME in age increases. The proposed estimators based on the SEV VonB G-Normal mixture model outperform the SEV VonB Gamma model estimators in terms of reducing large sample bias due to ME in age and true age distribution misspecification. In this chapter, we compare the finite sample bias of the parameter estimators based on the SEV VonB G-Normal mixture model and on the SEV VonB Gamma model. We consider robustness to mean lack of finite sample bias in the estimators.

The VonB model (Eqn. 1.22) assumes that all individuals in the population have the same VonB growth parameters (L_∞ and k) and does not account for the variability among individuals. Different studies have been carried out to extend this model to account for between-individual (BI) variation in growth. Shelton and Mangel (2012) [40] extended the VonB model to account for BI variation in growth by varying the VonB parameters (L_∞ and k) among individuals. Schafer (2002) [39] argue that a better way to formulate the VonB model is not to use the growth parameter L_∞ but rather to use a different constant parameter B such that $L_\infty = \mathrm{B}/k$. They variy only k to account for BI variation in growth. In this approach the growth parameters k and L_∞ are negatively correlated. Therefore, a fast grower will have a lower L_∞ and vice versa. However, in practice, this may not be the case. In this chapter, we investigate the finite sample robustness of the SEV VonB estimators based on the flexible G-Normal mixture age distribution, when the VonB model includes BI variation in growth.

4.2 SEV VonB BI Model

We will extend the SEV VonB model (Section 2.3) to include BI variation in growth. We assume BI variation in growth appears because individuals achieve different asymptotic sizes. Let $L_{\infty i}$ be the asymptotic size of the i$^{\text{th}}$ fish. The length of the i$^{\text{th}}$ fish with multiplicative ME is

$$Y_i = L_{\infty i}(1 - e^{-k(X_{T_i} - a_o)})\ e^{\epsilon_i},$$

where

$$L_{\infty i} = L_\infty\ e^{\delta_i},$$

where δ_i is $N(0, \sigma_\delta^2)$ and ϵ_i is $N(0, \sigma_e^2)$. We assume δ_i and ϵ_i are independent. Therefore, by combining the above equations we have

$$Y_i = L_\infty(1 - e^{-k(X_{T_i} - a_o)})\ e^{c_i},$$

where $c_i = (\delta_i + \epsilon_i)$ is Normally distributed with mean $E(\delta_i + \epsilon_i) = 0$ and variance $\sigma_c^2 = \sigma_\delta^2 + \sigma_e^2$. Let σ_c^2 be the ME variance in length confounded with BI variation in growth. The age of the $\mathrm{i^{th}}$ fish with multiplicative ME is

$$X_i = X_{T_i}\ e^{\mathrm{U}_i},$$

where U_i is $N(0, \sigma_u^2)$. By taking logarithm transformations we get

$$\log(Y_i) = \log\{L_\infty(1 - e^{-k(X_{T_i} - a_o)})\} + c_i, \tag{4.1}$$

$$\log(X_i) = \log(X_{T_i}) + U_i. \tag{4.2}$$

SEV VonB G-Normal Mixture BI Model: The SEV VonB G-Normal mixture BI model is an extension of the SEV VonB G-Normal mixture model (Section 3.3) to include BI variation in growth. This model is actually the same as in Section 3.3, except that the interpretation of σ_c is different from σ_e. So far we have considered σ_e to be known. However, we will now treat σ_c as a parameter of interest. Therefore, the parameter of interest is $\theta_R = (L_\infty, k, a_0, \sigma_c)$. In order to estimate $\theta = (L_\infty, k, a_0, \sigma_c, \lambda, \mu^d, \sigma)$ we optimize the likelihood function described by Eqn. (3.6).

SEV VonB Gamma BI Model: Similarly, we can extend the SEV VonB Gamma model (Section 2.3) to account for BI in growth. We called it the SEV VonB Gamma BI model. In this model, we estimate $\theta = (L_\infty, k, a_0, \sigma_c, \alpha, \beta)$ by optimizing the the likelihood function described by Eqn. (2.3). We use TMB to estimate the parameters of the SEV VonB Gamma model.

4.2.1 Finite Sample Bias

Bias can be calculated by first simulating a large number of data sets (say S) of finite size n, and then estimating the parameters of the misspecified model for each of the data sets. Then the results are averaged over the S data sets to approximate the bias in the estimated parameters. This is called the repeated finite sample method of approximating bias due to model misspecification and ME in age. Recall that the parameter of interest is θ_R and for the misspecified model the SEV MLE of θ_R is $\hat{\theta}_R$. The bias in the SEV MLE of $\hat{\theta}_R$ is

$$\text{Bias}(\hat{\theta}_R) = E(\hat{\theta}_R) - \theta_R^*,$$

where θ_R^* is the true parameter value of θ_R. The finite sample method generates S samples of finite size n, to approximate $\text{E}(\hat{\theta}_R)$. Therefore, the repeated finite sample method provides the approximate bias at a given sample size. The $E(\hat{\theta}_R)$ can be approximated by the average estimated value $\bar{\hat{\theta}}_R$ which is

$$\bar{\hat{\theta}}_R = \frac{\sum_{j=1}^{S} \hat{\theta}_{Rj}}{S},$$

where $\hat{\theta}_{Rj}$ is the estimated value of θ_R for the $\text{j}^{\text{th}}(\text{j} = 1, 2, ..., S)$ data set. Therefore the finite sample bias of $\hat{\theta}_R$ is

$$\text{Bias}(\hat{\theta}_R) \text{ approximately equal to } (\bar{\hat{\theta}}_R - \theta_R^*).$$

Therefore, the SEV MLE for θ_R, i.e. $\hat{\theta}_R$, is robust if

$$\bar{\hat{\theta}}_R(\sigma_u) \text{ approximately equal to } \theta_R^* \text{ for } \sigma_u \ \geq 0.$$

4.3 Simulation Studies

We conduct a simulation study to compare the performance of the SEV VonB G-Normal mixture BI model and the SEV VonB Gamma BI model for estimating the VonB growth parameters $\theta_R = (L_\infty, k, a_0, \sigma_c)$. The response Y is generated assuming a VonB growth model (Eqn. 4.1) with the parameters fixed at $L_\infty = 120, k = 0.2, a_o = -0.1$ and $\sigma_c = 0.1$. For all simulations we consider σ_u to vary from 0 to 0.25. The performance of the model estimators will be measured using bias and root mean squared error (RMSE) based on $S = 500$ simulated data sets consisting of different sample sizes using the following two examples. We consider sample sizes $n = 300$ and $n = 400$.

Example 1: We consider Case 1 (Section 2.3.2), where the true unobserved age distribution is Lognormal with mean 4 and skewness 1.6, which represents a light tail distribution.

Example 2: We consider the "TriGamma" case (described in Section 2.3.3), where the true unobserved age distribution is a mixture of three independent Gamma distributions.

4.3.1 Analysis Methods

In the case of repeated samples, the results are averaged over $S = 500$ data sets, each of size 200 and 400. Therefore, the average estimated value of L_∞, $\bar{\hat{L}}_\infty$, is

$$\bar{\hat{L}}_\infty = \frac{\sum_{j=1}^{S} \hat{L}_{\infty j}}{S},$$

where $\hat{L}_{\infty j}$ is the SEV MLE of L_∞ for the j^{th} simulated dataset. We focus on the bias and the RMSE for $\bar{\hat{L}}_\infty$, $\bar{\hat{k}}$, $\bar{\hat{a}}_o$ and $\bar{\hat{\sigma}}_c$. The bias of $\bar{\hat{L}}_\infty$ is

$$\mathrm{Bias}(\bar{\hat{L}}_\infty) = \bar{\hat{L}}_\infty - L^*_\infty,$$

where L^*_∞ is the true value of L_∞. The RMSE of $\bar{\hat{L}}_\infty$ is

$$\mathrm{RMSE}(\bar{\hat{L}}_\infty) = \sqrt{\frac{\sum_{j=1}^{S} (\hat{L}_{\infty j} - L^*_\infty)^2}{S}}.$$

In this manner we can find the bias and RMSE of the estimates $\bar{\hat{k}}$, $\bar{\hat{a}}_o$ and $\bar{\hat{\sigma}}_c$.

4.3.2 Repeated Sampling

Results of Example 1:

The simulated bias in the estimated parameters and RMSE of the estimates are shown in Figures 4.1 and 4.2. The average estimated values of the growth parameters are given in Appendix C (Table C.1). The frequency distributions of the 500 estimated parameters based on the SEV VonB two-Normal mixture BI model and the SEV VonB Gamma BI model are presented in Figures C.1-C.4 (see Appendix C). Overall, the bias as in the estimates of the VonB growth parameters based on the SEV VonB two-normal mixture BI model are fairly close to zero when compared to the bias for the SEV VonB Gamma BI model. The SEV VonB Gamma BI model performs poorly in terms of bias. For instance, when the sample size is 200 the results in Figure 4.1 (Top panel) illustrate that the bias in $\bar{\hat{L}}_\infty$ for the SEV VonB two-Normal mixture BI model are 0.07, 0.4 and 3.53 corresponding to σ_u values of $0.05, 0.1$ and 0.25, respectively. However, for the SEV VonB Gamma BI model these biases are 1.6, 3.9 and 19.45, respectively. Furthermore, when the sample size increased from 200 to 400, the bias in the estimates of the VonB growth parameters also decreases irrespective of the models. In addition, the SEV VonB two-Normal mixture BI model performs relatively well compared to the SEV VonB Gamma BI model in terms of estimating σ_c. It implies that the SEV VonB two-Normal mixture BI model is capable of correctly estimating the BI variation in growth. Therefore, in Example 1, a substantial amount of finite sample bias reduction occurs when using the proposed SEV VonB two-Normal mixture BI model.

The performance of the proposed SEV VonB two-Normal mixture BI model is superior to that of the SEV VonB Gamma BI model in terms of bias-variance trade off which is reflected in the RMSE values. For the SEV VonB two-Normal mixture BI model, the RMSE of the estimator of L_∞ varies from 7.1 to 16.7, while it varies from 7.5 to 27.6 for the SEV VonB Gamma BI model. An identical scenario is evident for the RMSE of the estimators for k and a_o. However, the RMSE of the estimator of σ_c is almost the same for both models. Therefore, in terms of bias-variance trade-off the SEV VonB Gamma BI model performs worse when compared with the proposed SEV VonB two-Normal mixture BI model.

Results of Example 2:

The simulated bias in the estimated parameters and RMSE of the estimates are shown in Figures 4.3 and 4.4. The average estimated values of the growth parameters are given in Table 2 of Appendix C (Table C.2). The frequency distribution of the

500 estimated parameters based on the SEV VonB two-Normal mixture BI model and the SEV VonB Gamma BI model are presented in Figures C.5-C.8 (see Appendix C). The SEV VonB two-Normal mixture BI model provides average estimates of the VonB growth parameters that are generally closer to their true population values when compared to those of the SEV VonB Gamma BI model. Therefore, the SEV VonB two-Normal mixture BI model performs better than the SEV VonB Gamma BI model in terms of reducing the bias. The SEV VonB Gamma BI model estimates have higher RMSE than the proposed SEV VonB two-Normal mixture BI model estimates (i.e., SEV VonB Gamma BI model estimates are less accurate).

Finally, we conclude that the proposed SEV VonB G-Normal mixture BI model estimators of the growth parameters have less bias and provide more accurate estimates than that of the SEV VonB Gamma BI model we investigated. Therefore, the proposed SEV VonB G-Normal mixture BI model is the estimating method we pursue in the remainder of this thesis.

4.4 Robustness of the SEV VonB G-Normal Mixture BI Model under both ME and Age Models Misspecifications

Recall that in our proposed model we have assumed that the length and ME models are correctly specified; and only the age model is misspecified. The simulation results suggest that the SEV MLE based on the proposed flexible G-Normal mixture distribution for unobserved ages are robust to misspecification of the true unobserved age distribution. While estimating the parameters in the SEV VonB BI model we assume that the ME variance in age σ_u is known. However, it is not possible to estimate the

ME in age until we have validation or replication data. In fisheries science sometimes these data are not readily available. Hence, in this situation we need to get information about the ME variance in age either from previous study results or from the researchers own experience. Therefore, there is always a chance of misspecifying the ME variance in age in the SEV VonB BI model. In this situation, misspecification not only occurs in the age models, but also in the ME model. This joint misspecification in both ME and age models may hamper the robustness of the SEV MLE based on the SEV VonB G-Normal mixture BI model.

4.4.1 Model Framework

Let σ_u be the true but unknown ME variance in age and define σ_u^a as the assumed ME variance in age. In the ME model, let $f_{\mathcal{X}|\mathcal{X}_T}(\mathcal{X} \mid \mathcal{X}_T; \sigma_u)$ be the true pdf of $\mathcal{X}$ given $\mathcal{X}_T$ and $f^A_{\mathcal{X}|\mathcal{X}_T}(\mathcal{X} \mid \mathcal{X}_T; \sigma_u^a)$ be the assumed pdf of $\mathcal{X}$ given $\mathcal{X}_T$. In our proposed SEV VonB G-Normal mixture BI model we assume that the ME model is correctly specified, i.e., $f_{\mathcal{X}|\mathcal{X}_T}(\mathcal{X} \mid \mathcal{X}_T; \sigma_u) = f^A_{\mathcal{X}|\mathcal{X}_T}(\mathcal{X} \mid \mathcal{X}_T; \sigma_u^a)$. Now, the observed likelihood (Eqn. 3.6) for $(\mathcal{Y}, \mathcal{X})$ under the misspecified age model is

$$\int f_{\mathcal{Y}|\mathcal{X}_T}(\mathcal{Y} \mid \mathcal{X}_T; \theta_R)\ f_{\mathcal{X}|\mathcal{X}_T}(\mathcal{X} \mid \mathcal{X}_T; \sigma_u)\ f^A_{\mathcal{X}_T}(\mathcal{X}_T; \theta_E)\ d\mathcal{X}_T.$$

The simulation results indicate that the SEV VonB G-Normal mixture BI model estimators are robust to misspecification of the true unobserved age distribution. However, when the ME variance in age is incorrectly assumed or estimated (i.e., $\sigma_u \neq \sigma_u^a$; consequently, $f_{\mathcal{X}|\mathcal{X}_T}(\mathcal{X} \mid \mathcal{X}_T; \sigma_u) \neq f_{\mathcal{X}|\mathcal{X}_T}(\mathcal{X} \mid \mathcal{X}_T; \sigma_u^a)$) then misspecification in the ME model occurs along with age model misspecification. Therefore, under misspecification in both ME and age models the observed likelihood is

$$\int f_{\mathcal{Y}|\mathcal{X}_T}(\mathcal{Y} \mid \mathcal{X}_T; \theta_R)\ f^A_{\mathcal{X}|\mathcal{X}_T}(\mathcal{X} \mid \mathcal{X}_T; \sigma^a_u)\ f^A_{\mathcal{X}_T}(\mathcal{X}_T; \theta_E)\ d\mathcal{X}_T, \tag{4.3}$$

where

1. **Length Model:** The pdf of $\mathcal{Y}$ given $\mathcal{X}_T$ is defined in Eqn. (3.7) with the parameter vector $\theta_R = (L_\infty, k, a_o, \sigma_c)$.

2. **ME Model:** The pdf of $\mathcal{X}$ given $\mathcal{X}_T$ is

$$f^A_{\mathcal{X}|\mathcal{X}_T}(\mathcal{X} \mid \mathcal{X}_T; \sigma^a_u) = \frac{1}{\sigma^a_u\sqrt{2\pi}} e^{-(\mathcal{X}-\mathcal{X}_T)^2/2\sigma^{a\,2}_u},$$

3. **Age Model:** The assumed pdf of $\mathcal{X}_T$ is defined in Eqn. (3.9) with the parameter vector θ_E = (p,μ, σ); where p $= (p_1, p_2, \ldots, p_G)$ with $\sum_{g=1}^G p_g = 1$, $p_g \geq 0$ for $g = 1, 2, \ldots, G$; $\mu = (\mu_1, \mu_2, \ldots, \mu_G)$ with restriction $\mu_1 < \mu_2 < \ldots < \mu_G$; and $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_G)$.

In the next section, we investigate the impact of mis-specifying both the true ME and the age models on the SEV VonB estimators.

4.4.2 Simulation Results

We randomly generate $S = 500$ independent datasets. We consider sample sizes $n = 200$ and $n = 500$. For each data set we perform the following steps:

- **Step 1:** We generate n true ages X_T from a Lognormal distribution (defined in Eqn. 2.18) with $\mu = 1.275$ and $\sigma = 0.4723$.

- **Step 2:** We generate n true lengths Y_T assuming a VonB growth model (Eqn. 1.22) using $L_\infty = 120, k = 0.2$, $a_o = -0.1$ and the true age X_T generated in

Step 1.

- **Step 3:** We use Eqn. (4.1) to generate n observed lengths $\mathcal{Y}$ with $\sigma_c = 0.1$, and Eqn. (4.2) to generate n observed ages $\mathcal{X}$ with true ME variance in age (σ_u) taken to be values of 0.05 and 0.15.

- **Step 4:** We find the MLEs of the parameters in θ by maximizing the observed likelihood Eqn. (4.3) with a range of σ_u^a values varying from 0 and 0.25.

We calculate the average estimated value of the parameter of interest θ_R. This average is taken over $S = 500$ estimated parameters. In addition, we calculate the 1st and 3rd quartiles of the 500 estimated values of θ_R.

Figure 4.5 illustrates the joint impact of mis-specifying the true age distribution as a two-Normal mixture when it is actually something else, like a Lognormal distribution and wrongly assuming the true ME variance in age when it is 0.05. When $\sigma_u = 0.05$ misspecified by a range of σ_u^a values vary from 0 to 0.25 then the bias in the estimates of parameters based on the SEV VonB two-Normal mixture BI model increase substantially. When σ_u^a is close to $\sigma_u = 0.05$ all of the average estimated values are about the same as the true values of the corresponding parameters. For example, when we take $\sigma_u^a = 0.05$ the average estimated value of L_∞ is about 120. This is expected since the estimators of the SEV VonB growth parameters are robust when we correctly assume the true ME variance in age; only the true unobserved Lognormal age distribution is misspecified by the proposed flexible G-Normal mixture distribution. When σ_u^a is in the vicinity of $\sigma_u = 0.05$ the average estimated values of L_∞ are still about the same as the true value of L_∞, which is 120. Figure 4.5 illustrates that when σ_u^a varies from 0.03 to 0.08, there is little bias in the estimates of L_∞. However, when the $\sigma_u = 0.05$ is wrongly assumed by σ_u^a (say 0.25) then there is a substantial amount of bias in the estimate of L_∞. Virtually identical results can be seen in the case of k and a_o. The average estimates of σ_c are very close to the true values when the σ_u^a is less than or equal to 0.08, but when σ_u^a reaches 0.15 the bias is large. Moreover, the average estimated values are zero for σ_u^a values 0.2 and 0.25. This implies that when σ_u^a is away from σ_u the average estimated values of σ_c reach zero. The results do not change when we use the sample size of 500 instead of the sample size of 300.

The results in Figure 4.6 illustrate the joint impact of mis-specifying the true age distribution as a two-Normal mixture when it is actually something else, like a Lognormal distribution and wrongly assuming the true ME variance in age, when it is 0.15. The estimates are very close to the true values when the σ_u^* value varies between 0.12 to 0.17, but when σ_u^a reaches 0.25 the bias is large. Furthermore, when σ_u^a is less than or equal to 0.1, the bias is large. In addition, the average estimates of σ_c are fairly close to its true value when σ_u^a is close to 0.15.

4.5 Summary

Results show that

- The proposed SEV VonB G-Normal mixture BI model performs relatively well in all the scenarios compared to the SEV VonB Gamma BI model in the case of finite sample.

- The SEV VonB G-Normal mixture BI model provides adequate finite sample bias reduction due to ME in age and age model misspecification.

- It also provides reasonable bias-variance trade-off, which is reflected by the RMSE values.

- When both ME and age models misspecifications occur, the estimators based on the SEV VonB G-Normal mixture BI model are not robust.

In the next chapter, we study the growth of Greenland Hailbut in the Northwest Atlantic using the SEV VonB G-Normal mixture BI model.

4.6 Figures

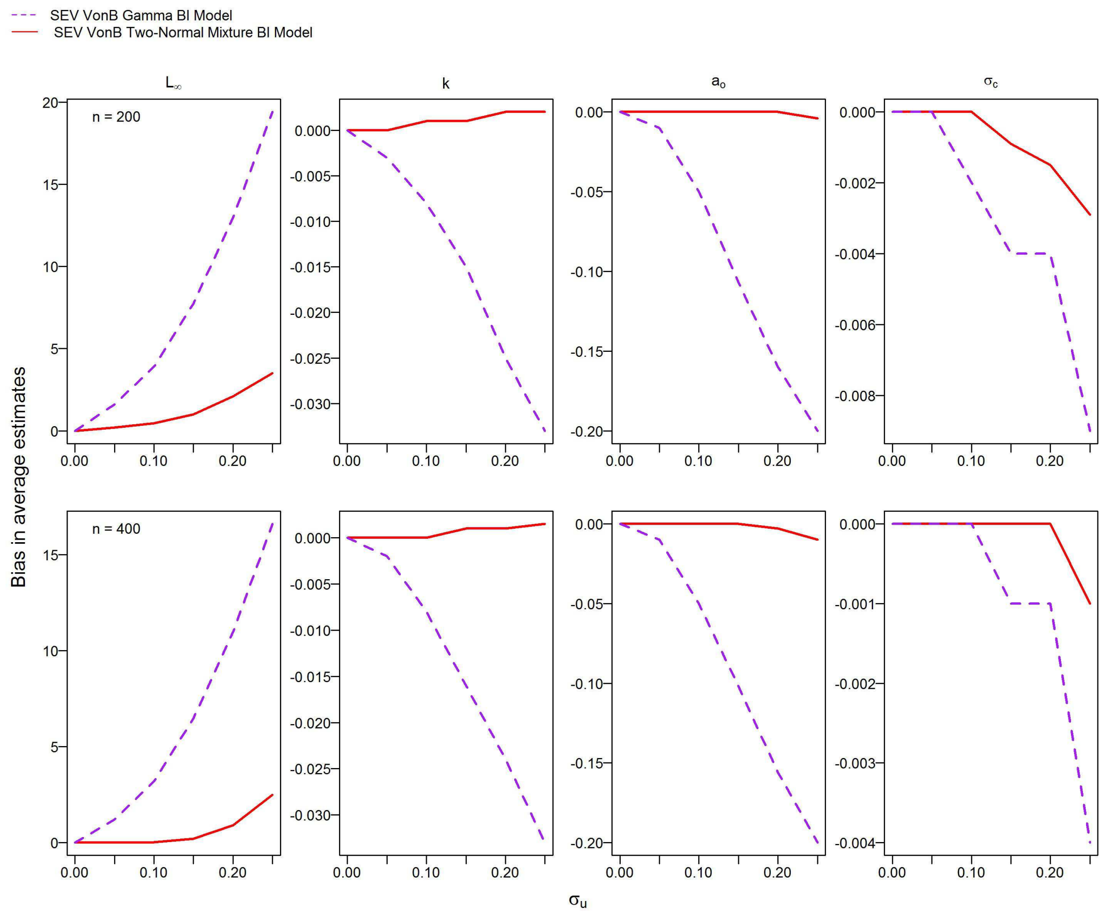

Figure 4.1: Sensitivity analysis of bias in the average estimates of L_∞, k, a_o and σ_c based on the SEV VonB BI model. The true unobserved age distribution is a Log-normal distribution with $\mu = 1.275$ and $\sigma = 0.4723$ which is misspecified as the two-Normal mixture and Gamma distributions.

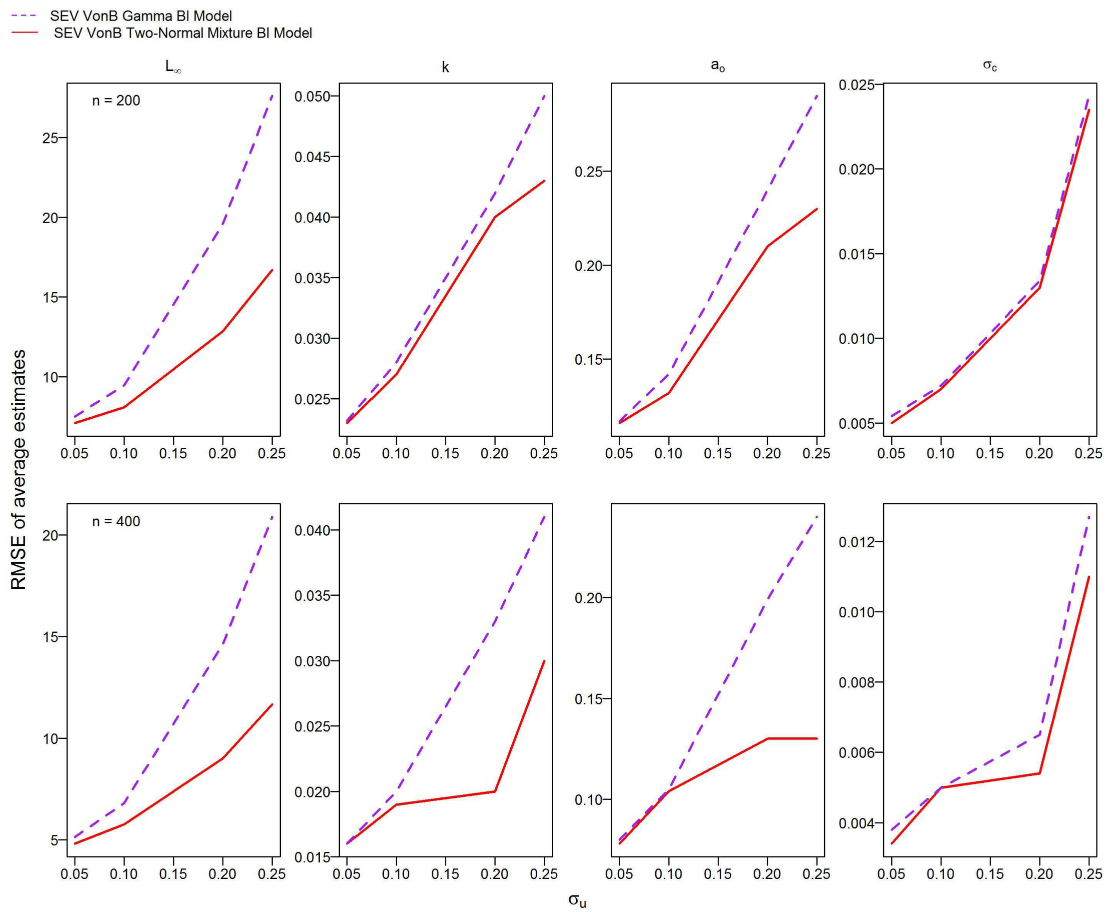

Figure 4.2: Sensitivity analysis of root mean squared error (RMSE) in the average estimates of L_∞, k, a_o and σ_c based on the SEV VonB BI model. The true unobserved age distribution is a Lognormal distribution with $\mu = 1.275$ and $\sigma = 0.4723$ which is misspecified as the two-Normal mixture and Gamma distributions.

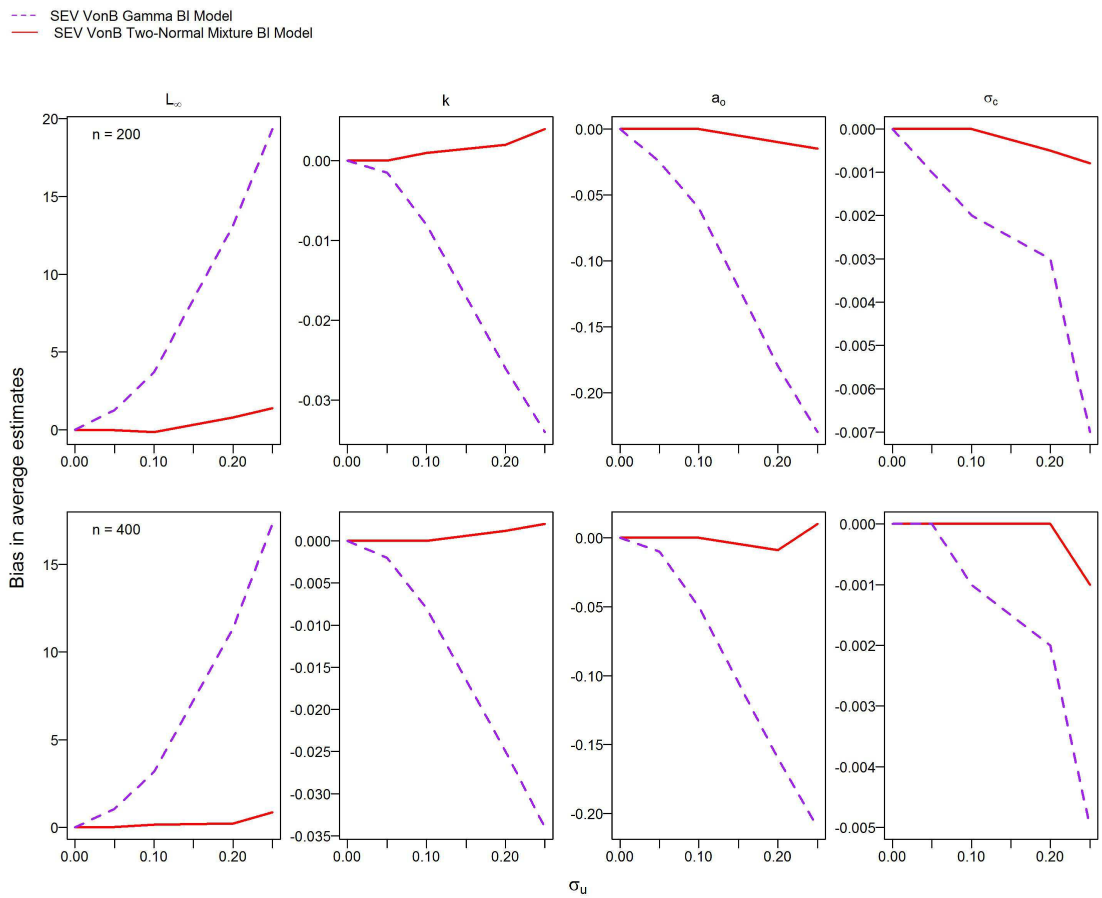

Figure 4.3: Sensitivity analysis of bias in the average estimates of L_∞, k, a_o and σ_c based on the SEV VonB BI model. The true unobserved age distribution is a mixture of three Gamma distributions which is misspecified as the two-Normal mixture and Gamma age distributions.

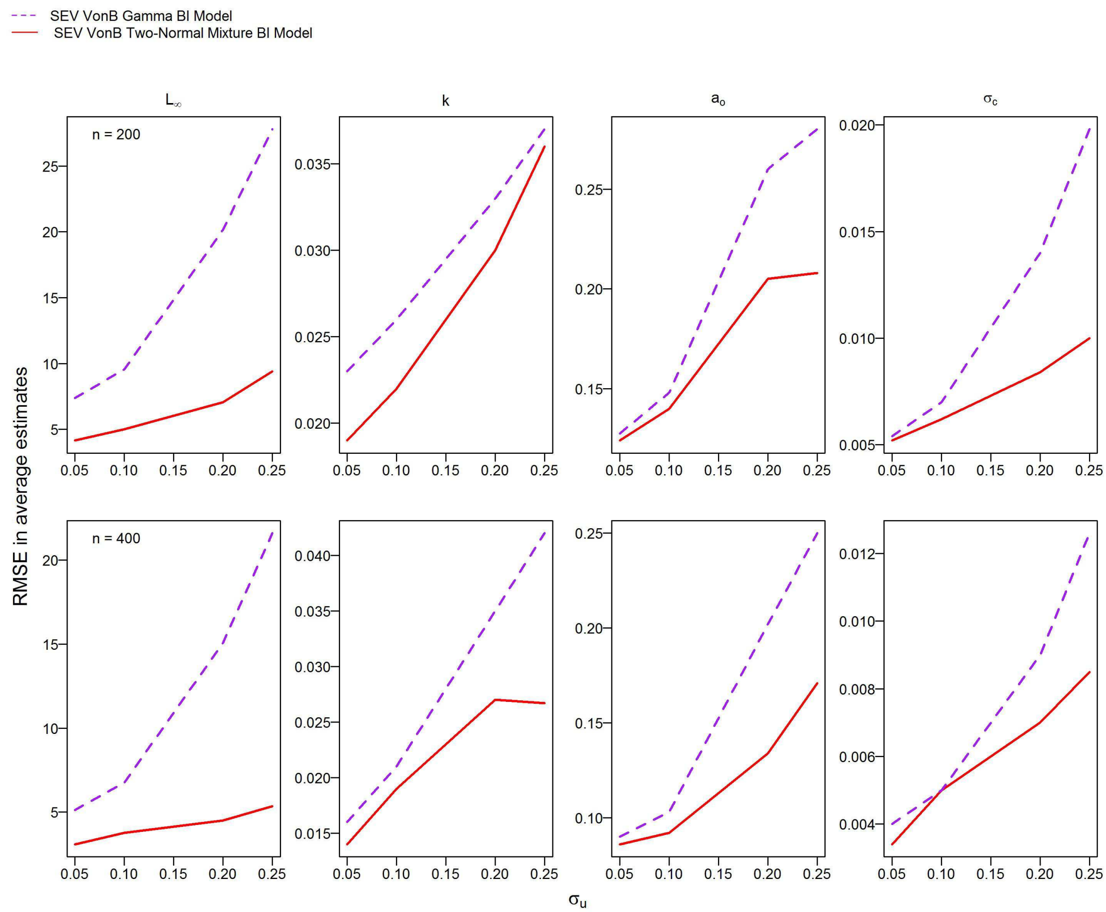

Figure 4.4: Sensitivity analysis of root mean squared error (RMSE) in the average estimates of L_∞, k, a_o and σ_c based on the SEV VonB BI model. The true unobserved age distribution is a mixture of three Gamma distributions which is misspecified as the two-Normal mixture and Gamma age distributions.

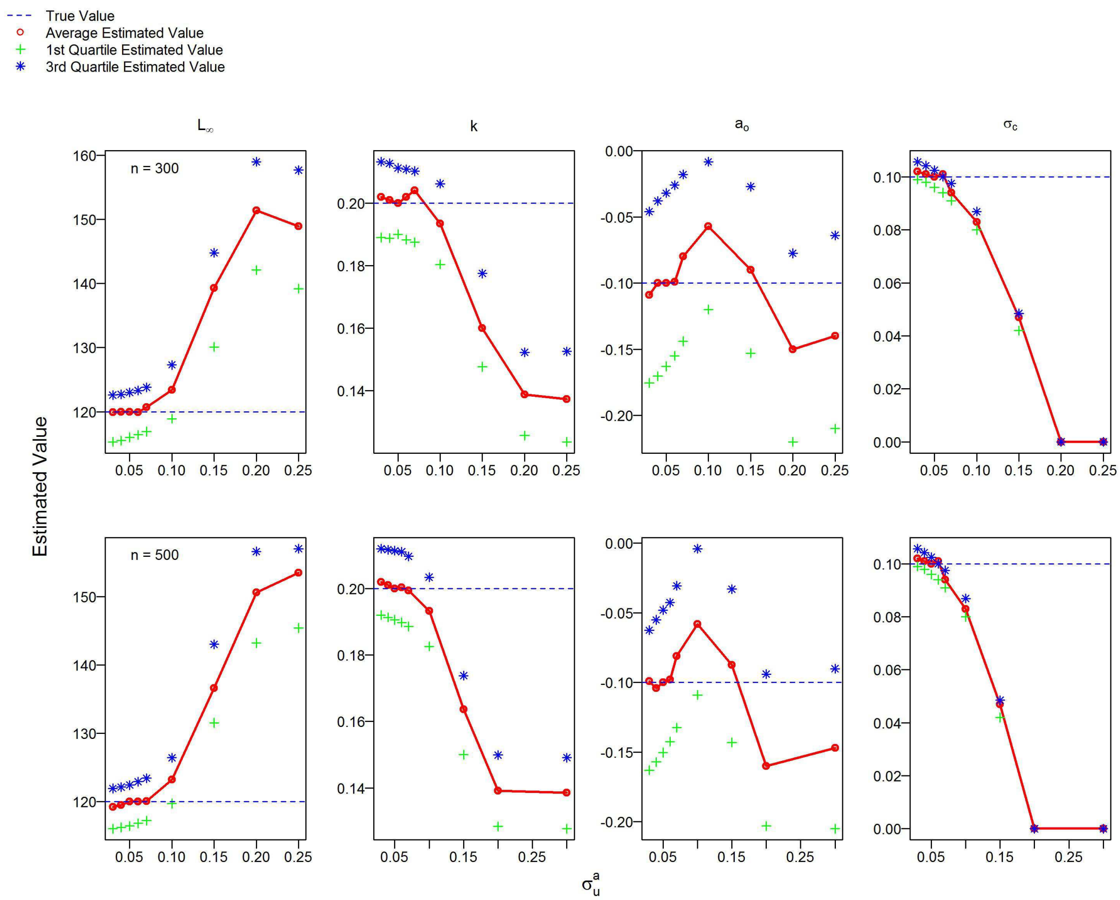

Figure 4.5: The estimated values based on the SEV VonB G-Normal mixture BI model estimators of L_∞, k, a_o and σ_c, when the true measurement error variance in age ($\sigma_u = 0.05$) is wrongly assumed by σ_u^a. The true unobserved age distribution is a Lognormal with $\mu = 1.275$ and $\sigma = 0.4723$, which is misspecified as the two-Normal mixture distribution. The first and third quartiles of the estimated parameters are presented.

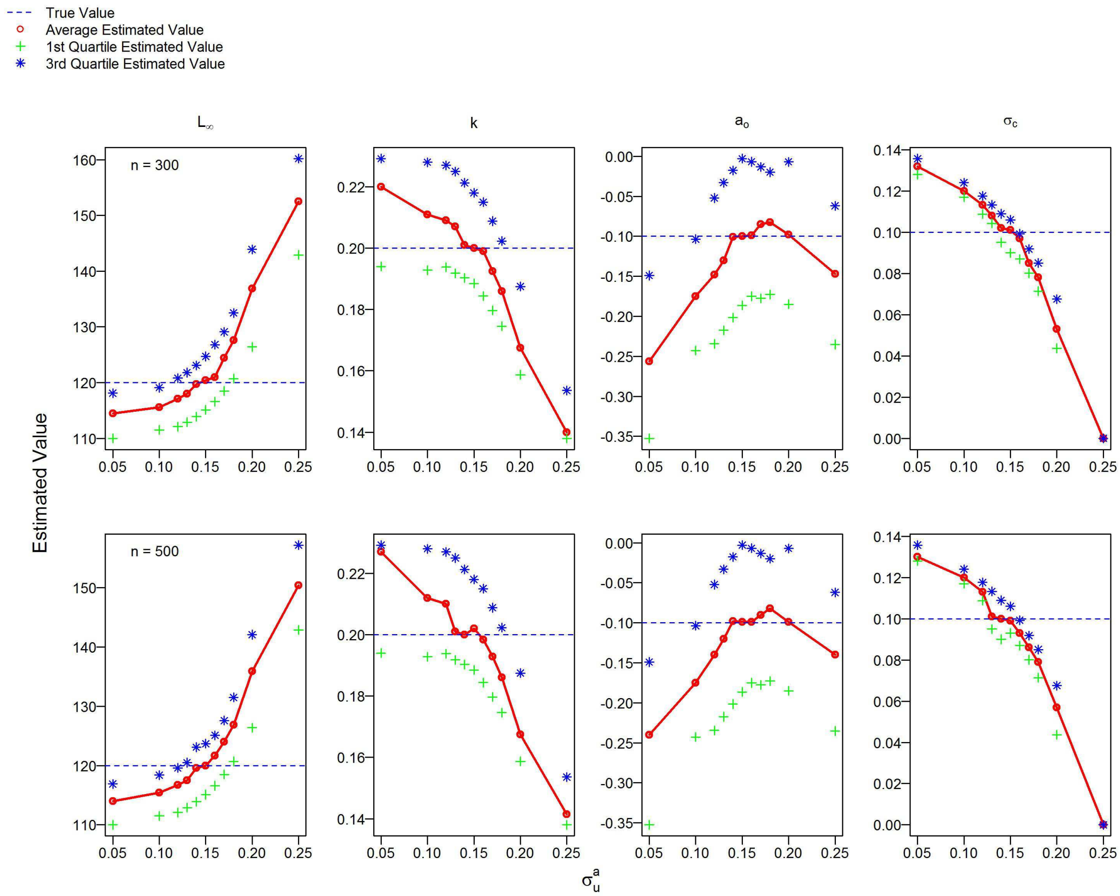

Figure 4.6: The estimated values based on the SEV VonB G-Normal mixture BI model estimators of L_∞, k, a_o and σ_c, when the true measurement error variance in age ($\sigma_u = 0.15$) is wrongly assumed by σ_u^a. The true unobserved age distribution is a Lognormal with $\mu = 1.275$ and $\sigma = 0.4723$, which is misspecified as the two-Normal mixture distribution. The first and third quartiles of the estimated parameters are presented.

Chapter 5

Application

5.1 Background

Greenland Halibut (Hippoglossoides reinhardtius) are a relatively large, deep water flatfish, and right-eyed flounder from the family Pleuronectidae. They inhabit the continental shelf and slope down to depths of 2200 meters, and are among the top predators in the Northwest Atlantic. They are found in the northern Atlantic and northern Pacific oceans. The Northwest Atlantic Fisheries Organization (NAFO) manages the stock off the East Coast of Canada in the Northwest Atlantic. Greenland Halibut was added to the seafood red list by Greenpeace International [22]. The seafood red list is a list of fish that are commonly sold in supermarkets around the world, and which have a very high risk of being sourced from unsustainable fisheries. This makes understanding their growth patterns a highly important subject, especially in the context of sustainable fishery management.

In this chapter, we study the growth of Greenland Halibut (Reinharditus hippoglossoides) in the Northwest Atlantic. We apply the SEV VonB G-Normal mixture BI model (see Section 4.2) to the age-to-length Greenland Hailbut data collected in the NAFO management unit Subarea 2 + Divisions 3KLMNO.

5.1.1 Sampling Scheme

NAFO conducted annual autumn surveys in NAFO management unit Subarea 2 + Divisions 3kLMNO during 1976 and 1977. The age, gender and length of fish records are mainly from these surveys. In addition, some smaller fish ($\leq$ 21 cm) collected in 2007 by NAFO are considered in this analysis. The sample consists of 271 fish, of which 104 are male and 167 are female. The relationship between fish length and age is curvilinear as can be seen from Figure 5.1. The maximum age is 33 years and maximum length is 110 cm. The maximum length is 70 cm for males and 110 cm for females.

Table 5.1 demonstrates that the average age and length of fish are about 8.5 years and 48 cm, respectively. Around fifty percent of fish are aged less than 6 years; for female and male fish it is 8 years and 5 years. Therefore, we have a young age fish population irrespective of sex. The variability in age of female fish (CV = 72 percent) is higher than that of male fish (CV = 57 percent). In addition, the median length of fish was 43 cm. On average the length of female fish (54 cm) is higher than that of male fish (37 cm). Overall, the high variability in the length of fish is evident from the CV of 54 percent. The variability in the length of female fish (CV = 52 percent) is higher than that of the male fish (CV = 44 percent).

5.2 Fitting of the SEV VonB Two-Normal Mixture BI Model with Greenland Hailbut Data

We fit the proposed SEV VonB two-Normal mixture BI model to estimate the VonB growth parameters L_∞, k, a_o and σ_c for the Greenland Hailbut data collected by

NAFO. To fit the model we need ME variance in age, σ_u, to be known. Since there are no gold-standards or replicated measurements of age by which to formally estimate the magnitude of the ME in age, sensitivity analyses are performed under different values for the magnitude of ME in age.

5.2.1 Fitting of the SEV VonB Two-Normal Mixture BI Model to the Full Data

The assumed value of σ_u is denoted by σ_u^a. We consider $\sigma_u^a = 0.05$, 0.12, 0.15, 0.2, and 0.25, which represents a range from no ME to substantial ME in age. For example, $\sigma_u^a = 0.25$ can be regarded as 25 percent ME in age.

Table 5.2 displays that the estimated values of the VonB growth parameters, i.e. L_∞, k, a_o and σ_c for the different values of σ_u^a. Overall, the estimated growth parameters increase with the value of σ_u^a. If there is no ME in age, i.e. $\sigma_u^a = 0$, the estimated asymptotic length (L_∞) for fish is 123.82 cm and the estimated growth parameter k, is 0.06. For small changes in the ME in age the estimated asymptotic length remains quite stable. For instance, the estimated asymptotic length is 124.5 cm when $\sigma_u^a = 0.05$, even for $\sigma_u^a = 0.15$ the estimated length is 126 cm. Therefore, the increase in the estimate of L_∞ for $\sigma_u^a = 0.15$ relative to $\sigma_u^a = 0$ is quite small. However, when the assumed ME in age increases substantially to ($\sigma_u^a = 0.25$) the estimated asymptotic length increases to 155 cm. This implies that if we ignore the impact of ME in age, the estimated asymptotic length will be much larger. Furthermore, the estimated values of k and a_o are insensitive to the magnitude of the assumed ME, σ_u^a. For example, when there is no ME in age the estimated value of k is 0.06, while it is 0.052 when $\sigma_u^a = 0.2$. Therefore, the increase in the estimate of k for $\sigma_u^a = 0.2$ relative to $\sigma_u^a = 0$ is quite small. The estimated variance of length confounded with BI variation in growth σ_c decreases with the increase in σ_u^a. The estimates of σ_c are fairly stable when σ_u^a is less than or equal to 0.15, but when σ_u^a reaches 0.25 the estimated value is 0 which is the lower bound of σ_c.

In addition, the standard error of the estimators for the growth parameters increases with the increase in σ_u^a, which reflects the larger information loss associated with higher ME in age. For example, the standard error of $\hat{L}_\infty$ is 7.73 when $\sigma_u^a = 0.05$, however, it reaches to 13.93 as σ_u^a increases to 0.25.

5.2.2 Fitting of the SEV VonB Two-Normal Mixture BI Model to the Female Data

We fit the SEV VonB two-normal Mixture Model to Greenland Halibut data, for females. Table 5.3 gives the VonB growth parameter estimates of L_∞, k, a_o and σ_c for different values of σ_u. We observe that the estimates of all the parameters increase with the assumed value of σ_u. When there is no ME in age, i.e. $\sigma_u^a = 0$, the estimated asymptotic length for female fish is 122.32 cm and the estimated growth parameter k is 0.062. These estimates are not comparable with the results of Dwyer et al. (2016)[17] since our estimates are based on the VonB model with multiplicative error while they used the VonB model with additive error. Moreover, our sample size is larger. The increase in the estimate of L_∞ for $\sigma_u^a = 0.15$ relative to $\sigma_u^a = 0$ is small. However, when σ_u^a increases substantially, the asymptotic length goes to 160 cm. Furthermore, the standard errors of the estimators of L_∞ increase with σ_u^a. The estimates of k and a_o are insensitive to the assumed value of σ_u.

5.2.3 Fitting of the SEV VonB Two-Normal Mixture BI Model to the Male Data

Table 5.4 provides estimated values of the VonB growth parameters for all assumed σ_u values in the case of male data. Overall, the estimated growth parameters increase with increasing σ_u^a. When there is no ME in the age of male fish (i.e. $\sigma_u^a = 0$) the estimated asymptotic length is 111.68 cm. The increase in the estimates of L_∞ for $\sigma_u^a = 0.15$ relative to $\sigma_u^a = 0$ is high. The standard errors of these estimates increase substantially with increasing σ_u^a. Moreover, the estimates of k and a_o are sensitive to the magnitude of ME in age.

5.3 Summary

In this chapter, we studied the impact of ME in age on the growth of Greenland Halibut in the Northwest Atlantic. We applied the SEV VonB two-Normal BI mixture model to the data. ME variance is often estimated in practice using multiple readers for the same fish, however, unfortunately, there is no such information available for Greenland Halibut. In addition, the analysis of age reader variation data can also be complicated and is beyond the scope of this thesis. As the SEV estimates of the growth parameters depend on σ_u, more precise knowledge of σ_u leads to more accurate estimates of the growth parameters (see section 4.4.2). Therefore, we performed a sensitivity analysis with different ME scenarios. We considered plausible values of ME in age σ_u^a varies from 0 to 0.25. The estimated asymptotic length of male fish is lower than that of female fish when $\sigma_u^a = 0$. The difference between sexes (e.g. Dwyer et al. (2016)[17]) is thought to be due to manner in which males and females direct excess energy into growth and reproduction. Therefore, we compare the growth rate curves by sex of Greenland Halibut over the different ME scenarios. The estimated values of L_∞ increases substantially when the ME in age increases, irrespective of sex. Therefore, the estimates of L_∞ is sensitive to the magnitude of σ_u^a. However, the estimates of k and a_o are insensitive to the value of σ_u^a for both the full and female data. The estimate of the variance of length confounded with BI variation in growth σ_c decreases when the σ_u^a increases.

5.4 Figure

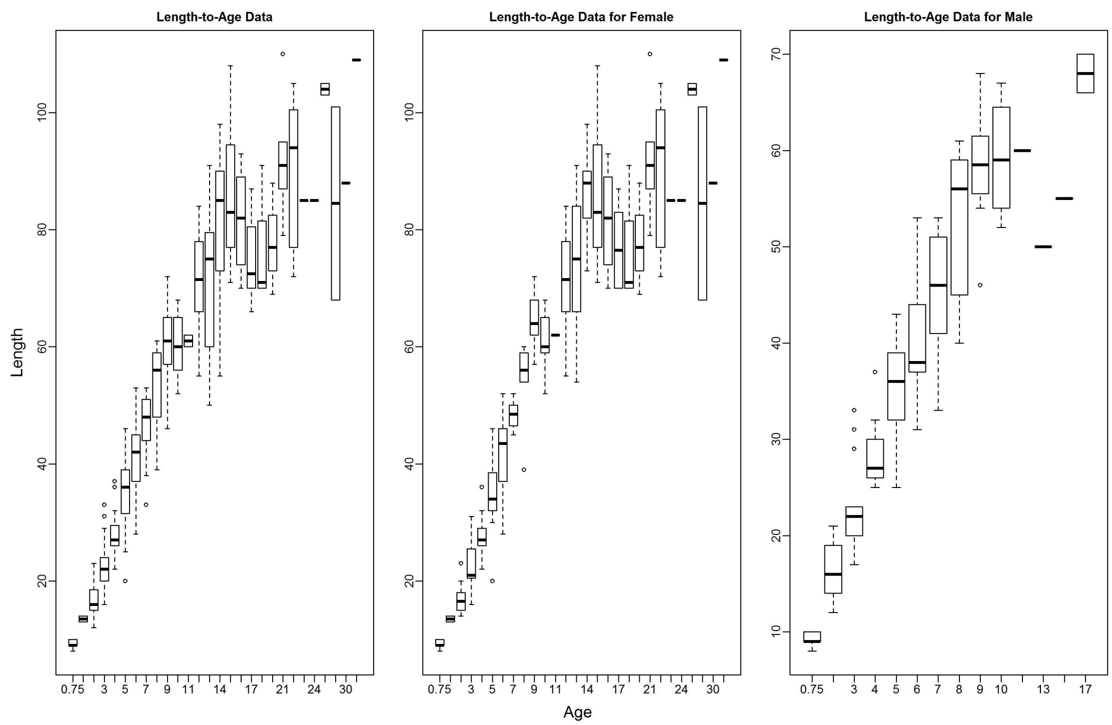

Figure 5.1: Length versus age plot of the Greenland Hailbut fish by their sex. The data collected by NAFO management unit Subarea 2 + Divisions 3kLMNO.

5.5 Tables

Table 5.1: Summary table of length and age of Greenland Halibut. The results for mean, median and coefficient of variation (CV) for length and age of Greenland Halibut by their sex.

Data	Estimators	Length	age
Full	Mean	47.5	8.46
	Median	43	6
	CV	54%	76.5%
Female	Mean	54.35	10.24
	Median	54	8
	CV	51.5%	71.5%
Male	Mean	36.5	5.61
	Median	37	5
	CV	43.75%	56.8%

Table 5.2: Parameter estimation results of Greenland Halibut based on the SEV VonB G-Normal mixture BI model for different assumed values of ME in age (σ_u^a). The results for the estimated values and its corresponding standard errors (SE) of the parameters based on the full data.

	σ_u^a					
Estimators	0	0.05	0.12	0.15	0.2	0.25
$\hat{L}_\infty$	123.82	124.53	125.2	126	138.65	155.53
SE($\hat{L}_\infty$)	7.8	7.64	8.24	8.53	10.8	13.93
$\hat{k}$	0.06	0.059	0.056	0.057	0.052	0.044
SE($\hat{k}$)	0.0056	0.0054	0.0054	0.0056	0.0058	0.0054
$\hat{a}_o$	-0.5	-0.497	-0.56	-0.44	-0.454	-0.48
SE($\hat{a}_o$)	0.18	0.0078	0.070	0.082	0.087	0.088
$\hat{\sigma}_c$	0.158	0.153	0.130	0.117	0.082	0
SE($\hat{\sigma}_c$)	0.006	0.0069	0.0079	0.0088	0.012	0.018

Table 5.3: Parameter estimation results of Greenland Halibut female fish based on the SEV VonB G-Normal mixture BI model for different assumed values of ME in age (σ_u^a). The results for the estimated values and its corresponding standard errors (SE) of the parameters based on the female data.

	σ_u^a					
Estimators	0	0.05	0.12	0.15	0.2	0.25
$\hat{L}_\infty$	122.32	122.72	125.51	125.7	136.9	160.38
SE($\hat{L}_\infty$)	8.35	8.07	8.87	9.78	12.92	21.59
$\hat{k}$	0.062	0.061	0.06	0.058	0.053	0.042
SE($\hat{k}$)	0.007	0.0064	0.0067	0.007	0.0077	0.008
$\hat{a}_o$	-0.508	-0.503	-0.487	-0.488	-0.50	-0.57
SE($\hat{a}_o$)	0.11	0.10	0.11	0.113	0.124	0.121
$\hat{\sigma}_c$	0.159	0.152	0.132	0.118	0.092	0.021
SE($\hat{\sigma}_c$)	0.007	0.0087	0.0097	0.01	0.013	0.089

Table 5.4: Parameter estimation results of Greenland Halibut male fish based on the SEV VonB G-Normal mixture BI model for different assumed values of ME in age (σ_u^a). The results for the estimated values and its corresponding standard errors (SE) of the parameters based on the male data.

	σ_u^a			
Estimators	0	0.05	0.12	0.15
$\hat{L}_\infty$	111.68	114.17	136.88	172.67
SE($\hat{L}_\infty$)	22.13	21.98	39.083	76.56
$\hat{k}$	0.068	0.065	0.052	0.039
SE($\hat{k}$)	0.018	0.015	0.019	0.021
$\hat{a}_o$	-0.443	-0.446	-0.476	-0.52
SE($\hat{a}_o$)	0.13	0.15	0.14	0.15
$\hat{\sigma}_c$	0.157	0.151	0.127	0.105
SE($\hat{\sigma}_c$)	0.01	0.011	0.013	0.016

Chapter 6

Conclusion

In this thesis, we proposed a flexible SEV VonB G-Normal mixture model when adjusting for the bias from the estimated regression growth parameters, i.e. L_∞, k and a_o. Our goal was to achieve robustness of the estimators under unobserved age model misspecification, while at the same time retaining the efficiency of the parametric inferences. By robustness we mean lack of bias in estimators for the parameters, regardless of the magnitude of the ME.

In Chapter 2, we discussed the SEV VonB model where the unobserved age is random; however, its distribution is not known in reality. Any distributional assumption on it, may be subject to misspecification. Cope and Punt (2007) [15] suggested an SEV model with a Gamma distribution for the unobserved true ages, X_T, and their results showed that this approach provided more precise estimates of the growth parameters compared to the nonlinear least squares method. Our simulation results show that when ME is low, the large sample bias in estimators of L_∞, k and a_o are small. However, as ME in age increases, the bias increases substantially. Therefore, the SEV VonB Gamma model estimators are not reliable to adjust the bias from the estimated growth parameters.

We compared the estimators based on the SEV VonB G-Normal mixture model and SEV VonB Gamma model in terms of large sample bias in Chapter 3. Extensive simulation studies demonstrate that the proposed SEV VonB G-Normal mixture model performs well in all the situations considered and outperforms the SEV VonB Gamma model. We advocate for using the G-Normal mixture distribution as the age model because of its flexibility in accommodating all possible features (e.g. skewness, heavy tailedness, and multimodality behaviour) that a true unobserved age distribution may have. The competing Gamma age model lacks the ability to capture one or more of these possible features.

Parameter estimation for our proposed SEV VonB G-Normal mixture model is complicated since the misspecified joint likelihood function involves integration; consequently, no closed-form solution is available. However, TMB reduces the complicity of integration by assuming Laplace approximation of the likelihood. The user only needs to specify the joint log-likelihood function; TMB then provides the marginal likelihood and its gradient function automatically. The analytical gradient greatly improves the speed and accuracy of marginal MLE's using a gradient-based optimization method. Therefore, the parameter estimation for our proposed SEV VonB G-normal mixture model is fairly simple using TMB.

Furthermore, in Chapter 4, we extended the proposed SEV VonB model to take into account for the BI variation in growth of the fish. We compared the SEV estimators of VonB growth parameters under the G-Normal mixture and Gamma age distributions in terms of finite sample bias. The proposed SEV VonB G-Normal mixture BI model is found to demonstrate adequate bias reduction performance and provide very reasonable bias-variance trade-off across different values of ME.

Moreover, we took into account for ME model misspecification along with age model misspecification. ME model misspecification occurs if there is no replication or validation data of age by which to formally estimate the magnitude of the ME σ_u. Therefore, σ_u to be the true but unknown ME variance in age and instead of it we use σ_u^a as an assumed ME variance in age. The simulation results indicate that the SEV VonB estimators based on the proposed flexible G-Normal mixture distribution for unobserved ages are no longer robust when we wrongly assume the true ME variance in age. However, the estimates of the VonB growth parameters are fairly close to their true values if σ_u^a is close to σ_u.

In Chapter 5, we applied the SEV VonB G-Normal mixture BI model to model the growth of the Greenland Halibut collected in the NAFO management unit Subarea 2 + Divisions 3kLMNO. Unfortunately, with this data there is no mechanism by which to rigorously estimate the nature or magnitude of the ME in age. We consider plausible values of ME in age σ_u^a vary from 0 to 0.25, which spans a range from no ME to a substantial ME in age. The estimated asymptotic length of male fish is smaller than that of female fish when $\sigma_u^a = 0$. The estimated values of L_∞ increases substantially when the ME in age increases irrespective of sex. Therefore, the estimates of L_∞ is sensitive to the assumed value of σ_u. However, the estimates of k and a_o are insensitive to the assumed value of σ_u for the full and female data. The estimate of the variance of length confounded with BI variation in growth σ_c decreases when the σ_u^a increases. Finally, we hope this analysis will serve as a resource for further modelling of 3kLMNO Greenland Halibut growth to predict fishery trends; this will help maintain a healthy and sustainable fish population.

Appendix A

Some Details for the SEV VonB Gamma Model

Recall that the observed likelihood for the SEV VonB Gamma model (Eqn. 2.3) defined as

$$L^A(\theta \mid Y, X) = \int f_{Y|X_T}(y \mid x_T; \theta_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f^A_{X_T}(x_T; \theta_E)\ dx_T,$$

where $\theta = (\theta_R, \theta_E)$, $\theta_R = (L_\infty, k, a_o)$, and $\theta_E = (\alpha, \beta)$. Therefore, $\theta = (L_\infty, k, a_o, \alpha, \beta)$ be a (5×1) vector. Recall that the $f_{Y|X_T}(y \mid x_T; \theta_R)$, $f_{X|X_T}(x \mid x_T; \sigma_u)$, and $f^A_{X_T}(x_T; \theta_E)$ are defined in Eqns. (2.15)-(2.17).

To estimate the parameters using the estimation method (section 2.2.3) we need to find the $S(\theta \mid Y, X)$ and $I(\theta \mid Y, X)$.

A.1 Score Vector for θ

The score function for θ under the misspecified model defined as

$$S(\theta \mid Y, X) = \frac{\frac{\partial}{\partial \theta} L^A(\theta \mid Y, X)}{L^A(\theta \mid Y, X)}.$$

where $\theta = (L_\infty, k, a_o, \alpha, \beta)$. The derivatives of $L^A(\theta \mid Y, X)$ with respect to θ are given below. We use the following formulas

$$\text{digamma}(\alpha) = \frac{\frac{\partial}{\partial \alpha}\Gamma(\alpha)}{\Gamma(\alpha)};$$

$$\text{trigamma}(\alpha) = \frac{\partial^2}{\partial \alpha^2}\log(\Gamma(\alpha)).$$

A.1.1 Derivation for $\frac{\partial}{\partial\theta} L^A(\theta \mid Y, X)$

$$\frac{\partial L^A(\theta \mid Y,X)}{\partial L_\infty} = \int f_{Y|X_T}(y \mid x_T; \theta_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f^A_{X_T}(x_T; \theta_E)\{\frac{\log(Y) - \log(Y_T)}{Y_T}\}$$
$$\{\frac{1 - e^{-k(X_T - a_o)}}{\sigma_e^2}\}\ dx_T.$$

$$\frac{\partial L^A(\theta \mid Y,X)}{\partial k} = \int f_{Y|X_T}(y \mid x_T; \theta_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f^A_{X_T}(x_T; \theta_E)\{\frac{\log(Y) - \log(Y_T)}{Y_T}\}$$
$$\{\frac{L_\infty\ (X_T - a_o)\ e^{-k(X_T - a_o)}}{\sigma_e^2}\}\ dx_T.$$

$$\frac{\partial L^A(\theta \mid Y,X)}{\partial a_o} = -\int f_{Y|X_T}(y \mid x_T; \theta_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f^A_{X_T}(x_T; \theta_E)\{\frac{\log(Y) - \log(Y_T)}{Y_T}\}$$
$$\{\frac{L_\infty\ k\ e^{-k(X_T - a_o)}}{\sigma_e^2}\}\ dx_T.$$

$$\frac{\partial L^A(\theta \mid Y,X)}{\partial \alpha} = \int f_{Y|X_T}(y \mid x_T; \theta_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f^A_{X_T}(x_T; \theta_E)$$
$$[\log(X_T) - \log(\beta) - \text{digamma}(\alpha)]\ dx_T.$$

$$\frac{\partial L^A(\theta \mid Y,X)}{\partial \beta} = \int f_{Y|X_T}(y \mid x_T; \theta_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f^A_{X_T}(x_T; \theta_E)(\frac{X_T}{\beta^2} - \frac{\alpha}{\beta})\ dx_T.$$

Using the results of section A.1.1, we can derive the $S(\theta \mid Y, X)$ for θ.

A.2 Hessian Matrix for θ

The Hessian matrix for θ is

$$I(\theta \mid Y, X) = \frac{\partial^2 \log(L^A(\theta \mid Y, X))}{\partial\theta\ \partial\theta'}.$$

This is a $(j \times k)$ matrix; $j, k = 1, 2, ..., 5$. The $(j,k)^{\text{th}}$ element of $I(\theta\,|Y, X)$, $I_{j,k}(\theta\,|Y, X)$, is

$$I_{j,k}(\theta\,|Y,X) = \frac{L^A(\theta \mid Y,X)\frac{\partial^2}{\partial\theta_k\partial\theta_j}L^A(\theta \mid Y,X) - \frac{\partial}{\partial\theta_k}L^A(\theta \mid Y,X)\frac{\partial}{\partial\theta_j}L^A(\theta \mid Y,X)}{[L^A(\theta \mid Y,X)]^2}. \tag{A.1}$$

A.2.1 Derivation for $\frac{\partial^2}{\partial\theta\partial\theta'}L^A(\theta \mid Y, X)$

$$\begin{aligned}\frac{\partial^2 L^A(\theta \mid Y,X)}{\partial L_\infty^2} = & \int\int f_{Y|X_T}(y \mid x_T;\theta_R)\ f_{X|X_T}(x \mid x_T;\sigma_u)\ f^A_{X_T}(x_T;\theta_E)\ \{\frac{1-e^{-k(X_T-a_o)}}{\sigma_e}\}^2 \\ & [\{\frac{\log(Y)-\log(Y_T)}{Y_T\ \sigma_e}\}^2 - \{\frac{1+\log(Y)-\log(Y_T)}{Y_T^2}\}]\ dx_T.\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 L^A(\theta \mid Y,X)}{\partial k^2} = & \int f_{Y|X_T}(y \mid x_T;\theta_R)\ f_{X|X_T}(x \mid x_T;\sigma_u)\ f^A_{X_T}(x_T;\theta_E)\ \{\frac{L_\infty\ (X_T-a_o)^2 e^{-k(X_T-a_o)}}{\sigma_e^2}\} \\ & (L_\infty\ e^{-k(X_T-a_o)}[\{\frac{\log(Y)-\log(Y_T)}{Y_T\ \sigma_e}\}^2 - \{\frac{1+\log(Y)-\log(Y_T)}{Y_T^2}\}] \\ & -\{\frac{\log(Y)-\log(Y_T)}{Y_T}\})dx_T.\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 L^A(\theta \mid Y,X)}{\partial a_o^2} = & \int f_{Y|X_T}(y \mid x_T;\theta_R)\ f_{X|X_T}(x \mid x_T;\sigma_u)\ f^A_{X_T}(x_T;\theta_E)\{\frac{L_\infty\ k^2\ e^{-k(X_T-a_o)}}{\sigma_e^2}\} \\ & (L_\infty\ e^{-k(X_T-a_o)}\ [\{\frac{\log(Y)-\log(Y_T)}{Y_T\ \sigma_e}\}^2 - \{\frac{1+\log(Y)-\log(Y_T)}{Y_T^2}\}] \\ & -\{\frac{\log(Y)-\log(Y_T)}{Y_T}\})\ dx_T.\end{aligned}$$

$$\begin{aligned}\frac{\partial L^A(\theta \mid Y,X)}{\partial \alpha^2} = & \int f_{Y|X_T}(y \mid x_T;\theta_R)\ f_{X|X_T}(x \mid x_T;\sigma_u)\ f^A_{X_T}(x_T;\theta_E) \\ & [\{\log(X_T)-\log(\beta)-\text{digamma}(\alpha)\}^2 - \text{trigamma}(\alpha)]\ dx_T.\end{aligned}$$

$$\frac{\partial L^A(\theta \mid Y, X)}{\partial \beta^2} = \int f_{Y|X_T}(y \mid x_T; \theta_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f^A_{X_T}(x_T; \theta_E)$$
$$[(\frac{X_T}{\beta^2} - \frac{\alpha}{\beta})^2 + (\frac{\alpha}{\beta^2} - \frac{2X_T}{\beta^3})]\ dx_T.$$

$$\frac{\partial^2 L^A(\theta \mid Y, X)}{\partial L_\infty\ \partial k} = \int f_{Y|X_T}(y \mid x_T; \theta_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f^A_{X_T}(x_T; \theta_E)\ \{\frac{(X_T - a_o)\ e^{-k(X_T - a_o)}}{\sigma_e^2}\}$$
$$([\{\frac{\log(Y) - \log(Y_T)}{Y_T\ \sigma_e}\}^2 - \{\frac{1 + \log(Y) - \log(Y_T)}{Y_T^2}\}]\ Y_T\ + \{\frac{\log(Y) - \log(Y_T)}{Y_T}\})dx_T.$$

$$\frac{\partial^2 L^A(\theta \mid Y, X)}{\partial L_\infty\ \partial a_o} = \int f_{Y|X_T}(y \mid x_T; \theta_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f^A_{X_T}(x_T; \theta_E)\ \{\frac{k\ e^{-k(X_T - a_o)}\ Y_T}{\sigma_e^2}\}$$
$$([\{\frac{1 + \log(Y) - \log(Y_T)}{Y_T^2}\} - \{\frac{\log(Y) - \log(Y_T)}{Y_T\ \sigma_e}\}^2]\ Y_T$$
$$-\ \{\frac{\log(Y) - \log(Y_T)}{Y_T}\})\ dx_T.$$

$$\frac{\partial^2 L^A(\theta \mid Y, X)}{\partial L_\infty\ \partial \alpha} = \int f_{Y|X_T}(y \mid x_T; \theta_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f^A_{X_T}(x_T; \theta_E)\ \{\frac{\log(Y) - \log(Y_T)}{Y_T}\}$$
$$\{\frac{1 - e^{-k(X_T - a_o)}}{\sigma_e^2}\}\ [\log(X_T) - \log(\beta) - \text{digamma}(\alpha)]\ dx_T.$$

$$\frac{\partial^2 L^A(\theta \mid Y, X)}{\partial L_\infty\ \partial \beta} = \int f_{Y|X_T}(y \mid x_T; \theta_R)\ f_{X|X_T}(x \mid x_T; \sigma_u)\ f^A_{X_T}(x_T; \theta_E)\ \{\frac{\log(Y) - \log(Y_T)}{Y_T}\}$$
$$\{\frac{1 - e^{-k(X_T - a_o)}}{\sigma_e^2}\}\ (\frac{X_T}{\beta^2} - \frac{\alpha}{\beta})\ dx_T.$$

$$\frac{\partial^2 L^A(\theta \mid Y, X)}{\partial k \ \partial a_o} = \int f_{Y|X_T}(y \mid x_T; \theta_R) \ f_{X|X_T}(x \mid x_T; \sigma_u) \ f^A_{X_T}(x_T; \theta_E) \{ \frac{(L_\infty - Y_T)}{\sigma_e^2} \} [k(X_T - a_o)\{ \{ \frac{\log(Y) - \log(Y_T)}{Y_T} \} + (L_\infty - Y_T)[\{ \frac{1 + \log(Y) - \log(Y_T)}{Y_T^2} \} - \{ \frac{\log(Y) - \log(Y_T)}{Y_T \ \sigma_e} \}^2] \} - \{ \frac{\log(Y) - \log(Y_T)}{Y_T} \}] dx_T.$$

$$\frac{\partial^2 L^A(\theta \mid Y, X)}{\partial k \ \partial \alpha} = \int f_{Y|X_T}(y \mid x_T; \theta_R) \ f_{X|X_T}(x \mid x_T; \sigma_u) \ f^A_{X_T}(x_T; \theta_E) \ L_\infty \ (X_T - a_o) \ e^{-k(X_T - a_o)} (\{ \frac{\log(Y) - \log(Y_T)}{Y_T \ \sigma_e} \}^2) \{ \log(X_T) - \log(\beta) - \text{digamma}(\alpha) \} \ dx_T.$$

$$\frac{\partial^2 L^A(\theta \mid Y, X)}{\partial k \ \partial \beta} = \int f_{Y|X_T}(y \mid x_T; \theta_R) \ f_{X|X_T}(x \mid x_T; \sigma_u) \ f^A_{X_T}(x_T; \theta_E) \ L_\infty \ (X_T - a_o) \ e^{-k(X_T - a_o)} (\{ \frac{\log(Y) - \log(Y_T)}{Y_T \ \sigma_e} \}^2)(\frac{X_T}{\beta^2} - \frac{\alpha}{\beta}) \ dx_T.$$

$$\frac{\partial^2 L^A(\theta \mid Y, X)}{\partial a_o \ \partial \alpha} = -\int f_{Y|X_T}(y \mid x_T; \theta_R) \ f_{X|X_T}(x \mid x_T; \sigma_u) \ f^A_{X_T}(x_T; \theta_E) \ k \ L_\infty \ e^{-k(X_T - a_o)} (\{ \frac{\log(Y) - \log(Y_T)}{Y_T \ \sigma_e} \}^2) \{ \log(X_T) - \log(\beta) - \text{digamma}(\alpha) \} \ dx_T.$$

$$\frac{\partial^2 L^A(\theta \mid Y, X)}{\partial a_o \ \partial \beta} = -\int f_{Y|X_T}(y \mid x_T; \theta_R) \ f_{X|X_T}(x \mid x_T; \sigma_u) \ f^A_{X_T}(x_T; \theta_E) \ k \ L_\infty \ e^{-k(X_T - a_o)} (\{ \frac{\log(Y) - \log(Y_T)}{Y_T \ \sigma_e} \}^2)(\frac{X_T}{\beta^2} - \frac{\alpha}{\beta}) \ dx_T.$$

$$\frac{\partial^2 L^A(\theta \mid Y, X)}{\partial \alpha \ \partial \beta} = \int f_{Y|X_T}(y \mid x_T; \theta_R) \ f_{X|X_T}(x \mid x_T; \sigma_u) \ f^A_{X_T}(x_T; \theta_E) [\{ \log(X_T) - \log(\beta) - \text{digamma}(\alpha) \} \ (\frac{X_T}{\beta^2} - \frac{\alpha}{\beta}) \ - \frac{1}{\beta}] \ dx_T.$$

We can obtain the Hessian matrix of θ using the results of sections A.2.1 and A.1.1 in Eqn. (A.1).

Appendix B

TMB Code for SEV VonB G-Normal Mixture Model

B.1 Continuation Ratio Logit

Consider the probabilities $p_1, p_2, ..., p_g$ for groups $1, 2, ..., G$ with the property that $\sum_{g=1}^{G} p_g = 1$, $p_g \geq 0$ for $g = 1, 2, ..., G$. Aitchison (2003) [2] argued that for ages and lengths the multiplicative Logistic transformation is more appropriate.

$$p_g = \begin{cases} \frac{exp(\lambda_g)}{\prod_{i=1}^{g}(1+exp(\lambda_i))} & g = 1, 2, ..., (G-1) \\ \frac{1}{\prod_{i=1}^{g}(1+exp(\lambda_i))} & g = G. \end{cases} \quad \text{(B.1)}$$

The inverse of this transformation is

$$\lambda_g = \log(\frac{p_g}{1 - p_1 - p_2 -, ..., -p_g}) = \log(\frac{p_g}{p_{g+1}+, ..., +p_G}), \; g = 1, 2, ..., (G-1). \quad \text{(B.2)}$$

This was the approach used by Stewart and Field (2011) [44] in diet composition analyses of seabirds and seals. Eqn. (B.2) is also the continuation-ratio logit as defined by Agersti (1996) [1] and has been used for modelling length and age distributions by Kvist et al. (2000) [27], Rindorf and Lewy (2001) [36], Cadigan (2016) [9], DFO (2011) [16], and possibly others.

B.2 C++ Template Code

The C++ template in the file named “fitmix.cpp” for the SEV VonB G-Normal Mixture model is

```
# include <TMB.hpp >
# include <iostream >
template <class Type >
Type objective_function <Type >::operator() ()
{
```

```
DATA_VECTOR(age);
DATA_VECTOR(length);
DATA_VECTOR(log_age);
DATA_VECTOR(log_length);
DATA_SCALAR(sig_log_len_me);
DATA_SCALAR(sig_log_age_me);
PARAMETER(log_Linf);
PARAMETER(log_k);
PARAMETER(ao);
PARAMETER_VECTOR(lambda);
PARAMETER_VECTOR(log_true_age);
PARAMETER_VECTOR(mix_mu_parm);
PARAMETER_VECTOR(log_mix_sigma);
int n = age.size();
int nmix = lambda.size();
Type zero = 0.0;
Type one = 1.0;
Type half = 0.5;
Type Linf = exp(log_Linf);
Type k = exp(log_k);
vector <Type >mix_sigma = exp(log_mix_sigma);
vector <Type >true_age = exp(log_true_age);
vector <Type >mu = Linf*(one - exp(-k*(true_age-ao)));
vector <Type >log_mu = log(mu);
vector <Type >p(nmix+1);
vector <Type >mix_mu(nmix+1);
```

```
[31] vector <Type >log_p(nmix+1);
[32] Type lterm = zero;
[33] for(int i = 0;i <nmix;++i){
[34] lterm += log(1 + exp(lambda(i)));
[35] log_p(i) = lambda(i) - lterm;
}
[36] log_p(nmix) = -lterm;
[37] p = exp(log_p);
[38] mix_mu(0) = mix_mu_parm(0);
[39] vector <Type>mix_norm = p(0)*dnorm(log_true_age,mix_mu(0),mix_sigma(0));
[40] for(int i = 1;i <= nmix;++i){
[41] mix_mu(i) = mix_mu(i-1) + exp(mix_mu_parm(i));
[42] mix_norm += p(i)*dnorm(log_true_age,mix_mu(i),mix_sigma(i));
}
[43] Type nll = zero;
[44] vector <Type>len_resid_std = (log_length-log_mu)/sig_log_len_me;
[45] nll -= dnorm(log_length,log_mu,sig_log_len_me,true).sum();
[46] nll -= dnorm(log_age,log_true_age,sig_log_age_me,true).sum();
[47] nll -= log(mix_norm).sum();
[48] ADREPORT(Linf)
[49] ADREPORT(k)
[50] ADREPORT(ao)
[51] ADREPORT(lambda)
[52] ADREPORT(mix_mu_parm)
[53] ADREPORT(mix_sigma)
[54] REPORT(true_age);
```

```
[55] REPORT(mu);
[56] REPORT(len_resid_std);
[57] REPORT(p);
[58] REPORT(mix_mu);
[59] REPORT(mix_sigma);
[60] REPORT(mix_norm);
[61] REPORT(nll)
[62] return nll;
}
```

B.3 TMB Code in R

Below is the operations we use in an R session:

```
[1] library(TMB)
[2] compile(“fitmix.cpp”)
[3] dyn.load(“fitmix”)
[4] tmb.data= list(
[5] age= age_obs,
[6] length= len_obs,
[7] log_age= log(age_obs),
[8] log_length= log(len_obs),
[9] sig_log_len_me= sig_log_len_me,
[10] sig_log_age_me= sig_log_age_me )
[11] g=2 # Determines the Number of Groups, G
[12] mix_mu= quantile(log(tmb.data$age),probs=((0:(g+1))/(g+1))[2:(g+1)])
[13] mix_mu_parm= c(mix_mu[1],diff(mix_mu))
```

```
[14] parameters= list(
[15] log_Linf= log(120),
[16] log_k= log(0.2),
[17] ao= -0.1,
[18] lambda= rep(0,g-1),
[19] log_true_age= log(true.age),
[20] mix_mu_parm= mix_mu_parm,
[21] log_mix_sigma= rep(log(0.5),g)
)
[22] parameters.U= list(
[23] log_Linf= log(200),
[24] log_k= log(0.5),
[25] ao= 10,
[26] lambda= rep(Inf,g-1),
[27] mix_mu_parm= rep(Inf,g),
[28] log_mix_sigma= rep(Inf,g)
)
[29] parameters.L= list(
[30] log_Linf= log(50),
[31] log_k= log(0.02),
[32] ao= -10,
[33] lambda= rep(-Inf,g-1),
[34] mix_mu_parm= rep(-Inf,g),
[35] log_mix_sigma= rep(-Inf,g)
)
[36] lower= unlist(parameters.L);
```

[37] upper= unlist(parameters.U);

[38] obj= MakeADFun(tmb.data,parameters,random=c("log_true_age"),DLL="fit-mix", random.start = expression(last.par[random]), inner.control=list(maxit=5000,trace=F));

[39] obj$fn(obj$par)

[40] obj$gr(obj$par)

[41] opt= nlminb(objpar,objfn,obj$gr,upper=upper,lower=lower, control = list(trace=10, eval.max=1e6,iter.max=1000))

[42] obj$gr(opt$par)

[43] rep= obj$report()

[44] sd.rep=sdreport(obj)

[45] summary(sd.rep,"report")

Appendix C

Repeated Sampling Results

C.1 Tables and Figures

Table C.1: Sensitivity to model misspecification: The true unobserved age distribution is Lognormal with $\mu = 1.275$ and $\sigma = 0.4723$, which is misspecified as the two-Normal mixture and the Gamma distributions. Results for average estimated values, and root mean squared errors(RMSE) of L_∞, k, a_o and σ_c were based on repeated Sampling.

		$\sigma_u = 0.05$		$\sigma_u = 0.1$		$\sigma_u = 0.2$		$\sigma_u = 0.25$	
Sample Size	Estimators	Mixture	Gamma	Mixture	Gamma	Mixture	Gamma	Mixture	Gamma
200	$\hat{L}_\infty$	120.07	121.58	120.4	123.9	122.17	132.9	123.53	139.45
	RMSE($\hat{L}_\infty$)	7.1	7.5	8.1	9.48	12.85	19.6	16.7	27.6
	$\hat{k}$	0.2	0.197	0.201	0.192	0.202	0.175	0.202	0.167
	RMSE($\hat{k}$)	0.023	0.0232	0.027	0.028	0.04	0.042	0.043	0.05
	$\hat{a}_o$	-0.1	-0.11	-0.1	-0.15	-0.1	-0.26	-0.103	-0.31
	RMSE($\hat{a}_o$)	0.116	0.117	0.132	0.142	0.21	0.24	0.23	.288
	$\hat{\sigma}_c$	0.1	0.1	0.099	0.098	0.098	0.096	0.097	0.092
	RMSE($\hat{\sigma}_c$)	0.005	0.0054	0.007	0.0072	0.013	0.0134	0.0235	.0243
400	$\hat{L}_\infty$	120.01	121.22	120.03	123.18	120.93	130.98	122.5	136.6
	RMSE($\hat{L}_\infty$)	4.8	5.14	5.76	6.8	9	14.62	11.67	20.9
	$\hat{k}$	0.2	0.198	0.201	0.192	0.201	0.176	0.201	0.168
	RMSE($\hat{k}$)	0.016	0.016	0.019	0.02	0.02	0.033	0.03	0.041
	$\hat{a}_o$	-0.1	-0.115	-0.1	-0.15	-0.103	-0.256	-0.11	-0.3
	RMSE($\hat{a}_o$)	0.078	0.08	0.104	0.105	0.13	0.199	0.13	.24
	$\hat{\sigma}_c$	0.1	0.1	0.1	0.099	0.1	0.099	0.099	0.096
	RMSE($\hat{\sigma}_c$)	0.0034	0.0038	0.005	0.0049	0.0054	0.0065	0.011	.0127

Table C.2: Sensitivity to model misspecification: The true unobserved age distribution is a mixture of three Gamma distributions, which is misspecified as the two-Normal mixture and the Gamma distributions. Results for average estimated values, and root mean squared errors(RMSE) of L_∞, k, a_o and σ_c were based on repeated Sampling.

		$\sigma_u = 0.05$		$\sigma_u = 0.1$		$\sigma_u = 0.2$		$\sigma_u = 0.25$	
Sample Size	Estimators	Mixture	Gamma	Mixture	Gamma	Mixture	Gamma	Mixture	Gamma
200	$\hat{L}_\infty$	119.98	121.26	119.84	123.7	120.88	133.14	121.46	139.34
	RMSE($\hat{L}_\infty$)	4.145	7.38	5.01	9.56	7.07	20.16	9.41	27.8
	$\hat{k}$	0.2	0.1985	0.201	0.192	0.202	0.174	0.205	0.166
	RMSE($\hat{k}$)	0.019	0.023	0.022	0.026	0.03	0.033	0.036	0.037
	$\hat{a}_o$	-0.1	-0.125	-0.099	-0.16	-0.11	-0.28	-0.115	-0.33
	RMSE($\hat{a}_o$)	0.124	0.128	0.14	0.148	0.205	0.26	0.208	0.28
	$\hat{\sigma}_c$	0.1	0.099	0.1	0.098	0.1	0.097	0.099	0.093
	RMSE($\hat{\sigma}_c$)	0.0052	0.0054	0.0062	0.007	0.0084	0.014	0.01	.0198
400	$\hat{L}_\infty$	120.03	121.05	120.15	123.17	120.21	131.4	120.85	137.27
	RMSE($\hat{L}_\infty$)	3.08	5.12	3.78	6.75	4.5	15.065	5.36	21.61
	$\hat{k}$	0.201	0.198	0.2	0.192	0.202	0.175	0.202	0.166
	RMSE($\hat{k}$)	0.014	0.016	0.019	0.021	0.027	0.035	0.0267	0.042
	$\hat{a}_o$	-0.1	-0.11	-0.1	-0.15	-0.109	-0.26	-0.09	-0.31
	RMSE($\hat{a}_o$)	0.086	0.09	0.092	0.103	0.134	0.202	0.171	0.25
	$\hat{\sigma}_c$	0.1	0.099	0.1	0.099	0.1	0.098	0.098	0.095
	RMSE($\hat{\sigma}_c$)	0.0034	0.004	0.005	0.005	0.007	0.009	0.0085	0.0126

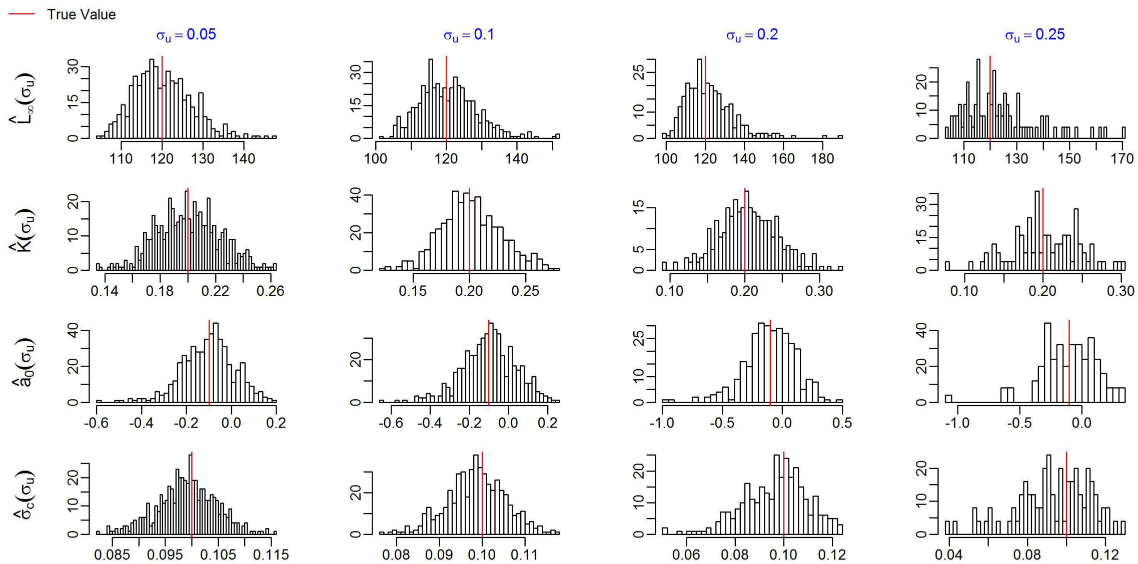

Figure C.1: Frequency distribution of $\hat{L}_\infty(\sigma_u), \hat{k}(\sigma_u)$, $\hat{a}_o(\sigma_u)$ & $\hat{\sigma}_c(\sigma_u)$ based on the SEV VonB two-Normal mixture BI model. The true unobserved age distribution is a Lognormal with $\mu = 1.275$ and $\sigma = 0.4723$ which is misspecified as the two-Normal mixture distribution. We consider the sample size 200.

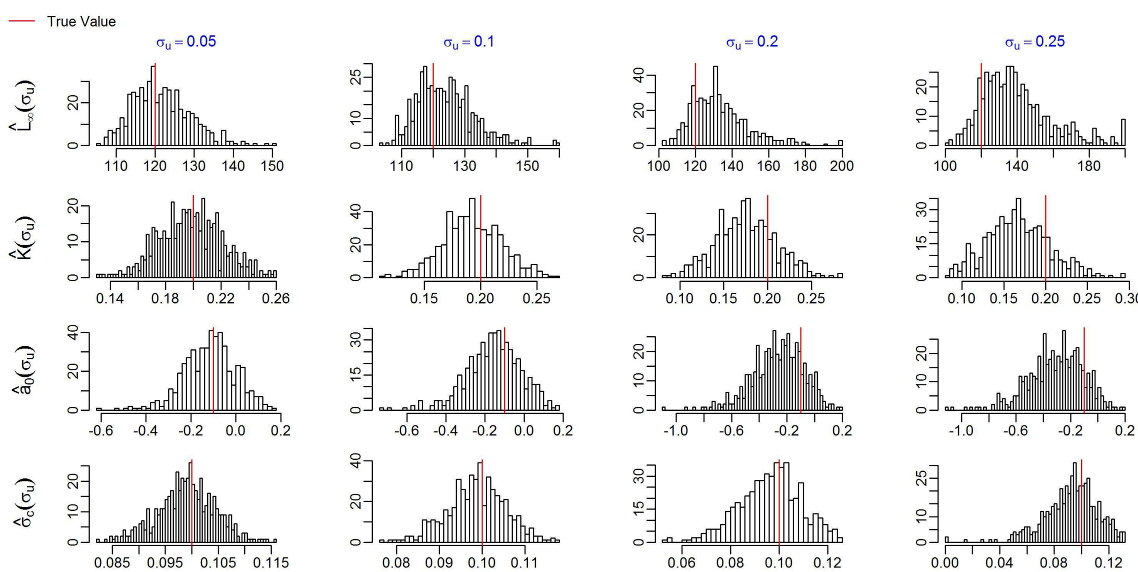

Figure C.2: Frequency distribution of $\hat{L}_\infty(\sigma_u), \hat{k}(\sigma_u)$, $\hat{a}_o(\sigma_u)$ & $\hat{\sigma}_c(\sigma_u)$ based on the SEV VonB Gamma BI model. The true unobserved age distribution is a Lognormal with $\mu = 1.275$ and $\sigma = 0.4723$ which is misspecified as the Gamma distribution. We consider the sample size 200.

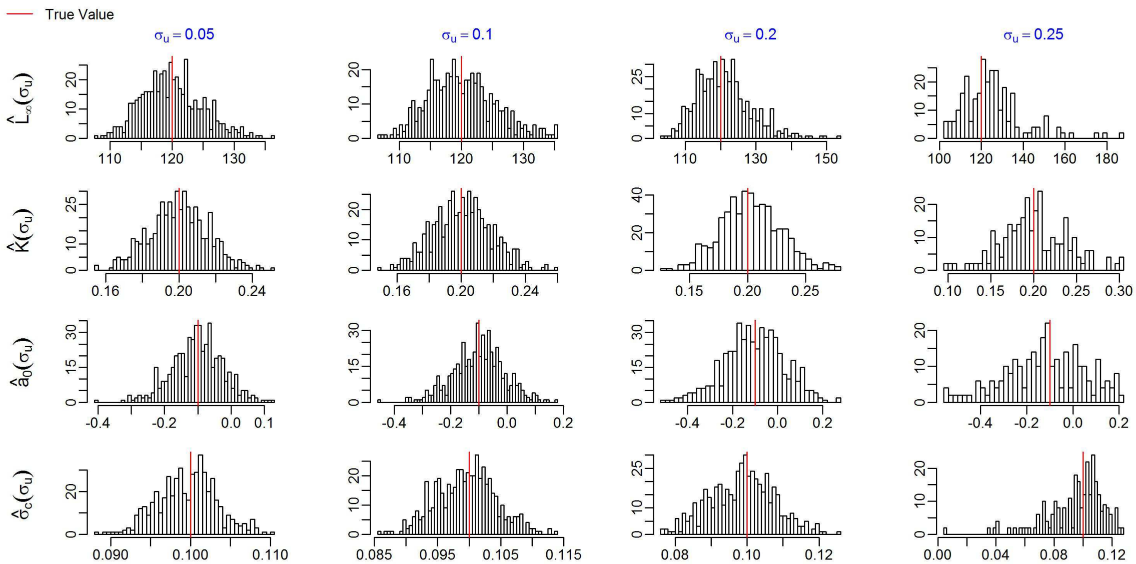

Figure C.3: Frequency distribution of the estimates $\hat{L}_\infty(\sigma_u), \hat{k}(\sigma_u)$, $\hat{a}_o(\sigma_u)$ & $\hat{\sigma}_c(\sigma_u)$ based on the SEV VonB two-Normal mixture BI model. The true unobserved age distribution is a Lognormal with $\mu = 1.275$ and $\sigma = 0.4723$ which is misspecified as the two-Normal mixture distribution. We consider the sample size 400.

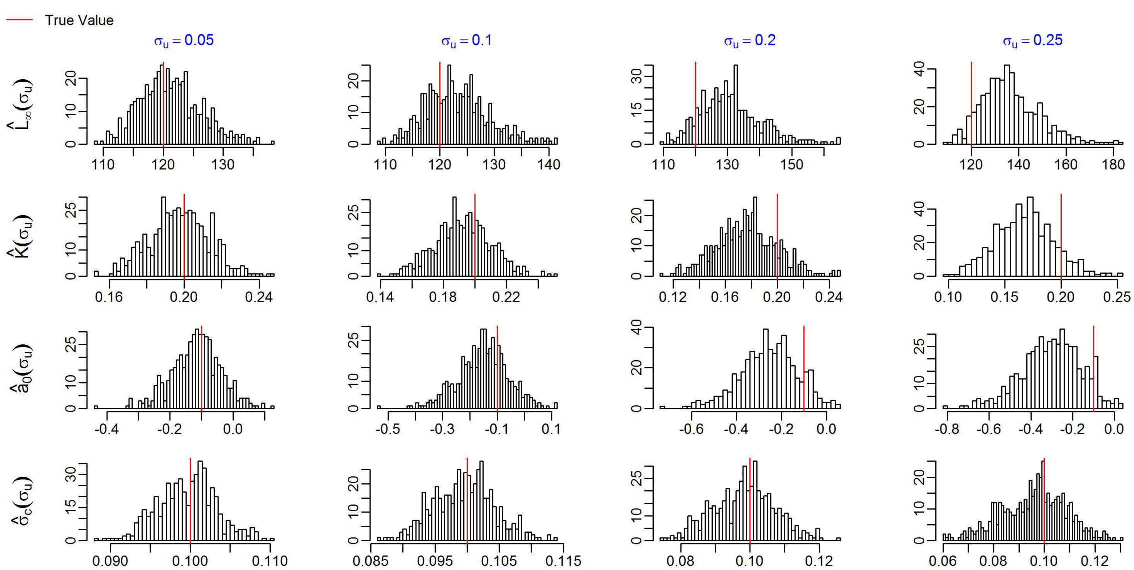

Figure C.4: Frequency distribution of the estimates $\hat{L}_\infty(\sigma_u), \hat{k}(\sigma_u)$, $\hat{a}_o(\sigma_u)$ & $\hat{\sigma}_c(\sigma_u)$ based on the SEV VonB Gamma BI model. The true unobserved age distribution is a Lognormal with $\mu = 1.275$ and $\sigma = 0.4723$ which is misspecified as the Gamma distribution. We consider the sample size 400.

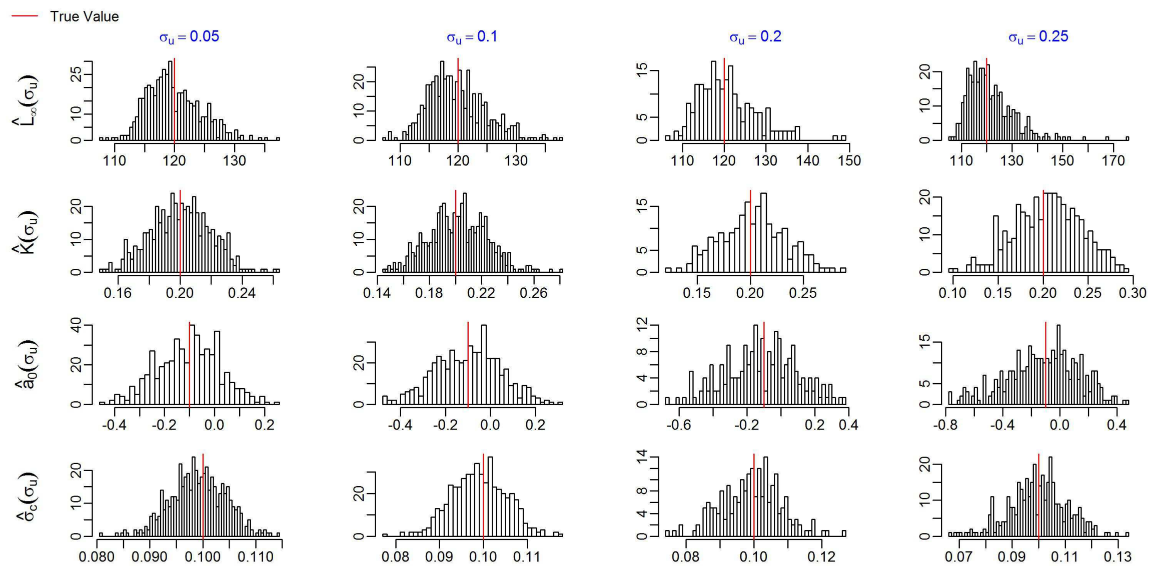

Figure C.5: Frequency distribution of the estimates $\hat{L}_\infty(\sigma_u), \hat{k}(\sigma_u)$, $\hat{a}_o(\sigma_u)$ & $\hat{\sigma}_c(\sigma_u)$ based on the SEV VonB two-Normal mixture BI model. The true unobserved age distribution is a mixture of three Gamma distributions which is misspecified as the two-Normal mixture distribution. We consider the sample size 200.

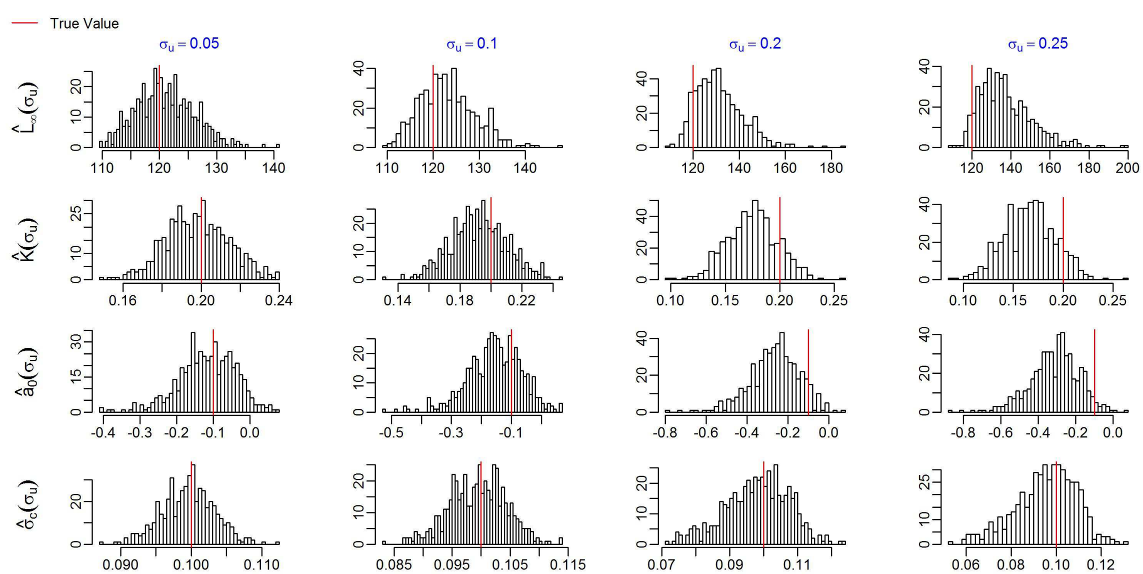

Figure C.6: Frequency distribution of the estimates $\hat{L}_\infty(\sigma_u), \hat{k}(\sigma_u)$, $\hat{a}_o(\sigma_u)$ & $\hat{\sigma}_c(\sigma_u)$ based on the SEV VonB Gamma BI model. The true unobserved age distribution is a mixture of three Gamma distributions which is misspecified as the Gamma distribution. We consider the sample size 200.

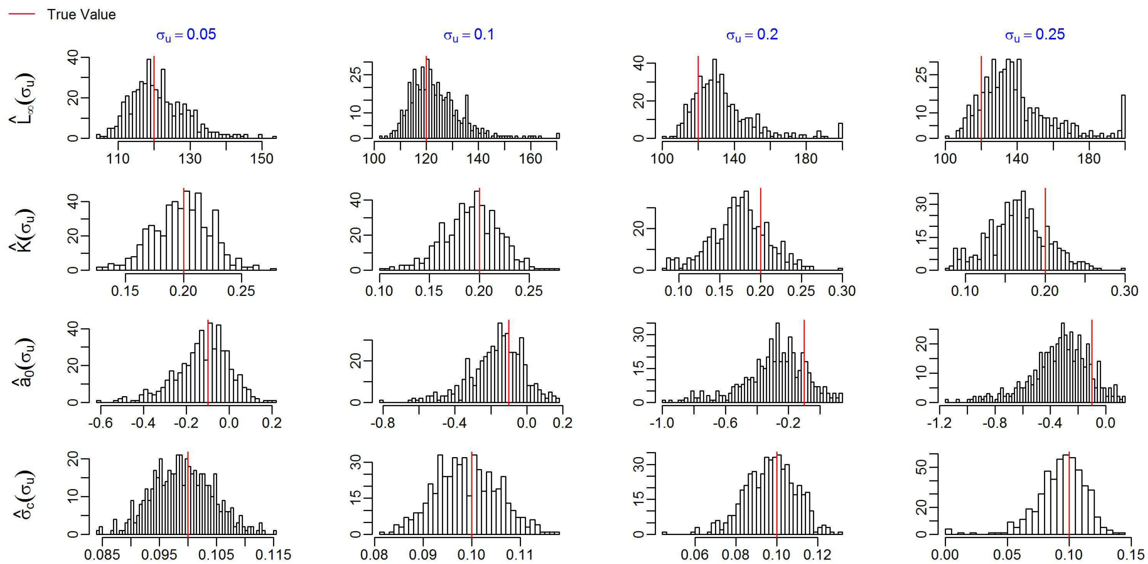

Figure C.7: Frequency distribution of the estimates $\hat{L}_\infty(\sigma_u), \hat{k}(\sigma_u)$, $\hat{a}_o(\sigma_u)$ & $\hat{\sigma}_c(\sigma_u)$ based on the SEV VonB two-Normal mixture BI model. The true unobserved age distribution is a mixture of three Gamma distributions which is misspecified as the two-Normal mixture distribution. We consider the sample size 400.

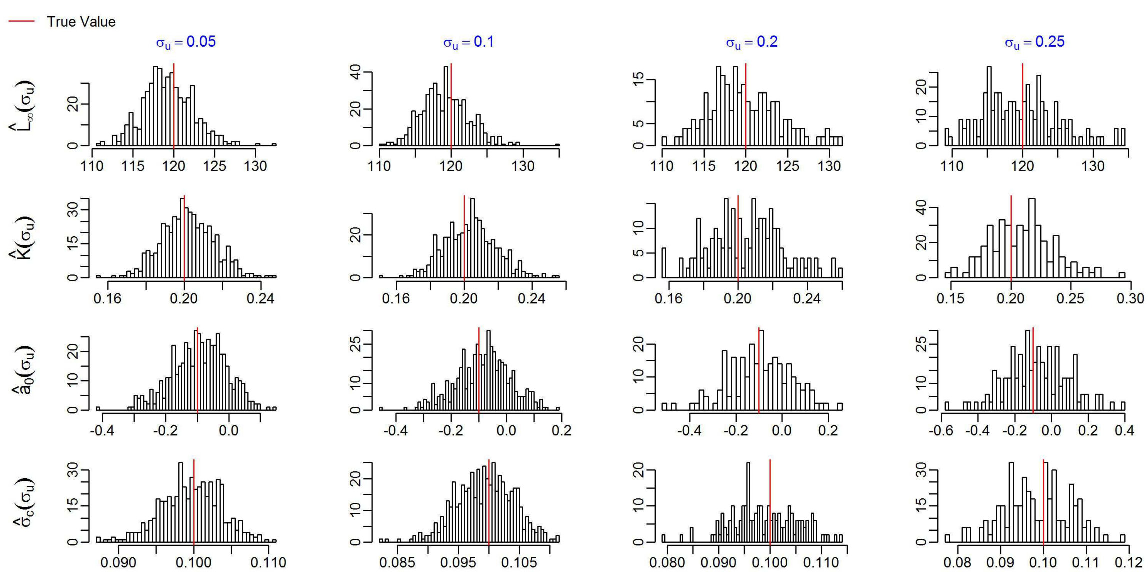

Figure C.8: Frequency distribution of the estimates $\hat{L}_\infty(\sigma_u), \hat{k}(\sigma_u)$, $\hat{a}_o(\sigma_u)$ & $\hat{\sigma}_c(\sigma_u)$ based on the SEV VonB Gamma BI model. The true unobserved age distribution is a mixture of three Gamma distributions which is misspecified as the Gamma distribution. We consider the sample size 400.

Bibliography

[1] Agresti, A. (1996). An introduction to categorical data analysis. Wiley, New York.

[2] Aitchison, J. (2003). The statistical analysis of compositional data. Reprint edition. The Blackburn Press, Caldwell, New Jersey.

[3] Barndorff-Nielsen, O. E., and Cox, D. R. (1994). Inference and Asymptotics. London: Chapman and Hall.

[4] Berger, J. O., Liseo, B., and Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. Statistical Science, 14(1): 1-28.

[5] Bell, B. M. (2012). CppAD: a package for C++ algorithmic differentiation. Computational Infrastructure for Operations Research.

[6] Bradford, M. J. (1991). Effects of ageing errors on recruitment time series estimated from sequential population analysis. Canadian Journal of Fisheries and Aquatic Sciences, 48: 555-558.

[7] Buzas, J. S., Tosteson, T. D., and Stefanski, L. A. (2003). Measurement error. Institute of Statistics Mimeo Series No. 2544, 4:192.

[8] Cadigan, N. G., and Campana, S. E. (2016). Hierarchical model-based estimation of population growth curves for redfish (Sebastes mentella and Sebastes fasciatus) off the Eastern coast of Canada. ICES Journal of Marine Science: Journal du Conseil, Oxford University Press, fsw195.

[9] Cadigan, N. G. (2016). A state-space stock assessment model for northern cod, including under-reported catches and variable natural mortality rates. Canadian Journal of Fisheries and Aquatic Sciences, 73(2): 296-308.

[10] Carroll, R. J., and Stefanski, L. A. (1990). Approximate quasilikelihood estimation in models with surrogate predictors. Journal of the American Statistical Association, 85:652663.

[11] Carroll, R. J., Roeder, K., and Wasserman, L. (1999). Flexible parametric measurement error models. Biometrics, 55: 44-54.

[12] Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). Measurement error in nonlinear models: a modern perspective. CRC press.

[13] Chen, Y., and Mello, L. G. S. (1999).Growth and maturation of cod (Gadus morhua) of different year classes in the Northwest Atlantic, NAFO subdivision 3Ps. Fisheries research, 42: 87-101.

[14] Cook, J.R., and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. Journal of the American Statistical Association, 89:1314-1328.

[15] Cope, J. M., and Punt, A. E. (2007). Admitting ageing error when fitting growth curves: an example using the von Bertalanffy growth function with random effects, Canadian Journal of Fisheries and Aquatic Sciences, 64: 205-218.

[16] DFO. (2011). Stock Assessment of Northern (2J3KL) cod in 2011. DFO Canadian Science Advisory Secretariat Science Advisory Report 2011/041.

[17] Dwyer, K. S., Treble, M. A. and Campana, S. E. (2016). Age and growth of Greenland Halibut (Reinhardtius hippoglossoides) inthe Northwest Atlantic: A changing perception based on bombradiocarbon analyses. Fisheries Research, 179: 342:250.

[18] Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., ... and Sibert, J. (2012). AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models.Optimization Methods and Software 27(2), 233-249.

[19] Francis, R. I. C. C. (1988). Are growth parameters estimated from tagging and agelength data comparable?. Canadian Journal of Fisheries and Aquatic Sciences, 45: 936-2342.

[20] Francis, R. I. C. (2014). Replacing the multinomial in stock assessment models: a first step. Fisheries Research, 151: 7084. doi:10.1016/j.fishres.2013.12.015.

[21] Fuller, W. (2009). Measurement error models. New York: John Willey & Sons.

[22] Greenpeace International. Red list: Seafood to Avoid at the Grocery Store. http://www.greenpeace.org/international/en/campaigns/oceans/seafood/red-list-of-species/.

[23] Heagerty, P. J., and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalized linear mixed models. Biometrika, 88: 973-985.

[24] Hossain, S., and Gustafson, P. (2011). Assessing large sample bias in misspecified model scenarios with reference to exposure model misspecification in error-in-variable regression: A new computational approach. Journal of Statistical Planning and Inference, 141: 1161-1169.

[25] Hossain, S., and Gustafson, P. (2009). Bayesian adjustment for covariate measurement errors: A flexible parametric approach. Statistics in Medicine, 28: 1580-1600.

[26] Huang, X., Stefanski, L. A., and Davidian, M. (2006). Latent-model robustness in structural measurement error models. Biometrika, 93(1), pp.53-64.

[27] Kvist, T., Gislason, H., and Thyregod, P. (2000). Using continuation-ratio logits to analyze the variation of the age composition of fish catches. J. Appl. Stat. 27: 303319. doi:10.1080/02664760021628.

[28] Kitakado, T. (2000). Application of the Conditional Score Method for Estimating a Fish Growth Curve. Japanese Journal of Biometrics, 21: 29-39.

[29] Kimura, D. K. (1980). Likelihood methods for the Von Bertalanffy growth curve. U.S. Natl. Mar. Fish. Serv. Fish. Bull, 77(4):765-776.

[30] Kuchenhoff, H., and Carroll, R. J. (1997). Segmented regression with errors in predictors: semi-parametric and parametric methods. Statistics in Medicine, 16:169-188.

[31] Ma, Y, Genton, MG. (2004). Flexible class of skew-symmetric distributions. Scandinavian Journal of Statistics, 31: 459-468.

[32] Mohammed, Z. (2015). Modelling the Growth of Northern Cod. MAS thesis report. Memorial University of Newfoundland.

[33] Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. Biometrika, 69:331342.

[34] R Core Team. (2014). R: A Language and Environment for Statistical Computing., R Foundation for Statistical Computing, Vienna, Austria.

[35] Reeves, S. A. (2003). A simulation study of the implications of age reading errors for stock assessment and management advice. ICES J. Mar. Sci. 60(2):314-328.

[36] Rindorf, A., and Lewy, P. (2001). Analyses of length and age distributions using continuation-ratio logits. Can. J. Fish. Aquat. Sci. 58(6): 11411152. doi:10.1139/f01-062.

[37] Sainsbury J. (1980). Effect of individual variability on the von Bertalanffy growth equation. Can. J. Fish. Aquat. Sci., 37:241-247.

[38] Schnute, J. T., Mulligan, T. J. and Kuhn, B. R. (1990). An errors-in-variables bias model with an application to salmon hatchery data. Canadian Journal of Fisheries and Aquatic Sciences, 47: 1453-1467.

[39] Schafer D. (2002). Likelihood analysis and flexible structural modeling for measurement error model regression. Journal of Computational Statistics and Data analysis, 72: 33-45.

[40] Shelton A. O., and Mangel M. (2012). Estimating von Bertalanffy parameters with individual and environmental variations in growth, Journal of biological dynamics, 6(2):3-30.

[41] Skaug, H. J., and Fournier, D. A. (2006). Automatic approximation of the marginal likelihood in non-gaussian hierarchical models. Computational Statistics & Data Analysis 51(2), 699-709.

[42] Stefanski, L. A., and Cook, J. R. (1995). Simulation extrapolation: the measurement error in Jackknife. Journal of the American Statistical Association, 90: 1247-1256.

[43] Stefanski, L. (2000). Measurement error models. Journal of the American Statistical Association, Taylor& Francis, 95, 1353-1358.

[44] Stewart, C., and Field, C. 2011. Managing the essential zeros in quantitative fatty acid signature analysis. J. Agric. Biol. Environ. Stat., 16: 4569. doi:10.1007/s13253-010-0040-8.

[45] Suh, E. Y., and Schafer, D. W. (2002). Semiparametric maximum likelihood for nonlinear regression with measurement errors. Biometrics, 58: 448-453.

[46] Tomilnson, P. K., and Abramson, N. J. (1961). Fitting a Von Bertalanffy curve by least squares. Calif. Dep. Fish Game Bull. 116-169.

[47] Walters, C. J., and Ludwig, D. (1981). Effects of measurement errors on the assessment of stock recruitment relationships. Canadian Journal of Fisheries and Aquatic Sciences, 38: 704-710.

[48] Uhler, R. S. (1980). Least squares regression estimates of the Schaefer production model: some Monte Carlo simulation results. Canadian Journal of Fisheries and Aquatic Sciences, 37: 1284-1294.

[49] Vincenzi, S., Mangel, M., Crivelli, A. J., Munch, S. and Skaug, H. J. (2014). Determining individual variation in growth and its implication for life-history and population processes using the Empirical Bayes method. PLoS Computational Biology, 10(9), p.e1003828.

[50] Von Bertalanffy, L. (1960). Principles and theory of growth. In: Nowinski, W.W. (Ed.), Fundamental Aspects of Normal and Malignant Growth. Elsevier, Amsterdam, pp. 137259.

[51] Quinn, T. J., and Deriso, R. B. (1999). Quantitative fish dynamics. Oxford University Press.

[52] Wang, Shijia. (2015). Estimation and inference with complex count data from fisheries surveys, including over-dispersion, many nuisance parameters, and correlation. MSc thesis report. Memorial University of Newfoundland.

[53] Huang, X. (2006). Robustness in latent variable models. PhD Dissertation. North Carolina State University.