Use of Difference-Based Methods to Explore Statistical and Mathematical Model Discrepancy in Inverse Problems

H.T. Banks, Jared Catenacci and Shuhua Hu
Center for Research in Scientific Computation
North Carolina State University
Raleigh, NC 27695-8212 USA
September 15, 2015

Abstract

Normalized differences of several adjacent observations, referred to as pseudo measurement errors in this paper, are used in so-called difference-based estimation methods as building blocks for the variance estimate of measurement errors. Numerical results demonstrate that pseudo measurement errors can be used to serve the role of measurement errors. Based on this information, we propose the use of pseudo measurement errors to determine an appropriate statistical model and then to subsequently investigate whether there is a mathematical model misspecification or error. In the presence of model misspecification, we also propose to use the information provided by pseudo measurement errors to quantify uncertainty in parameter estimation by bootstrapping methods. A number of numerical examples are given to illustrate the effectiveness of these proposed methods.

Key Words: pseudo measurement errors, statistical model, inverse problems, model misspecification, bootstrapping.

Mathematics Subject Classification: 34A55, 62H12, 62F40.

1 Introduction

A number of difference-based methods have been proposed in the literature to estimate the variance of measurement errors in a nonparametric regression where the mean function of observations is unknown and estimated using some nonparametric methods (e.g., see [15, 19, 21, 22] and the reference therein). These methods involve differencing the data and do not require estimation of the mean function. Specifically, the estimated variance is defined as the weighted average of the squared normalized difference of $\nu + 1$ observations, where ν is an integer. These normalized differences of $\nu + 1$ observations are called *pseudo measurement errors* in this paper. The purpose of this paper is to illustrate how these pseudo measurement errors can be used as a possible way of detecting **statistical model misspecification** or discrepancy as well as **mathematical model misspecification** or discrepancy within the context of a least squares inverse problem. That is, given a set of observed data, a mathematical model describing the observed process is fitted to the data via a least squares formulation by estimating a set of unknown parameters. As has become conventional in inverse problems, we would also like to quantify the uncertainty present in the estimation of the mathematical model parameters using confidence intervals. To do this, one must have a correctly specified statistical model which describes the data collection process along with a mathematical model which is assumed to describe the process under observation. Typically, one uses residual plots as illustrated in [8, 12] to ensure that the assumptions made in specifying both the statistical and mathematical model are not violated. However, the residuals can only be computed **after** the inverse problem has been completed. Furthermore, if the residual plots do not illustrate the desired random patterns, it can be difficult, if not impossible to determine if this is caused by mathematical model discrepancy or statistical model discrepancy, or both. We will show how differencing techniques can be used directly on the data to deduce if the assumptions of the statistical model have been violated prior to running the inverse problem. Then residual plots can be used afterwards to verify that the mathematical model is sufficiently accurate.

We consider inverse or parameter estimation problems in the context of a parameterized (with vector parameter $\mathbf{q} \in \Omega^{\kappa_q} \subset \mathbb{R}^{\kappa_q}$) n -dimensional vector dynamical system (for a physical or biological process) or **mathematical model** given by

$$\frac{d\mathbf{x}}{dt}(t) = \mathbf{h}(t, \mathbf{x}(t), \mathbf{q}), \quad (1.1)$$

$$\mathbf{x}(t_s) = \mathbf{x}_0, \quad (1.2)$$

with **observation process**

$$\mathbf{f}(t; \boldsymbol{\theta}) = \mathcal{C}\mathbf{x}(t; \boldsymbol{\theta}), \quad (1.3)$$

where $\boldsymbol{\theta} = (\mathbf{q}^T, \tilde{\mathbf{x}}_0^T)^T \in \Omega \subset \mathbb{R}^{\kappa_q + \tilde{n}} = \mathbb{R}^{\kappa_\theta}$, $\tilde{n} \leq n$, κ_q is the number of unknown dynamic parameters, $\tilde{n}$ is the number of unknown initial conditions $\tilde{\mathbf{x}}_0$, and the observation operator $\mathcal{C}$ maps $\mathbb{R}^n$ to $\mathbb{R}^m$. The sets Ω^{κ_q} and Ω are assumed known restraint sets for the parameters.

We make some standard assumptions (see [8, 12]) underlying our inverse problem formulations.

- Assume that there exists a *true or nominal set* of parameters $\boldsymbol{\theta}_0 \in \Omega$.

- Ω is a compact subset of Euclidian space of $\mathbb{R}^{\kappa_\theta}$ and $\mathbf{f}(t; \boldsymbol{\theta})$ is continuous on $[0, T] \times \Omega$.

Denote $\hat{\boldsymbol{\theta}}$ as the estimated parameter for $\boldsymbol{\theta}_0 \in \Omega$. The inverse problem is based on statistical assumptions on the observation error in the data. We consider a general **statistical model** of the form

$$\mathbf{Y}_j = \mathbf{f}(t_j; \boldsymbol{\theta}_0) + \mathbf{g}(t_j; \boldsymbol{\theta}_0, \boldsymbol{\mathcal{E}}_j), \quad j = 1, 2, \dots, N, \quad (1.4)$$

where $\mathbf{Y}_j = (Y_{1j}, Y_{2j}, \dots, Y_{mj})^T$, and $\mathbf{f}(t_j; \boldsymbol{\theta}_0) = (f_1(t_j; \boldsymbol{\theta}_0), f_2(t_j; \boldsymbol{\theta}_0), \dots, f_m(t_j; \boldsymbol{\theta}_0))^T$ denotes observations of the mathematical model describing the underlying physical or biological process with the nominal parameters $\boldsymbol{\theta}_0$ at the measurement point t_j . The so called measurement or observation error at t_j is represented by $\mathbf{g}(t_j; \boldsymbol{\theta}_0, \boldsymbol{\mathcal{E}}_j)$ where $\boldsymbol{\mathcal{E}}_j = (\mathcal{E}_{1j}, \mathcal{E}_{2j}, \dots, \mathcal{E}_{mj})^T$ is a $m \times 1$ random vector, and N is the total number of observations.

The simplest form of (1.4) is when $\mathbf{g}(t_j; \boldsymbol{\theta}_0, \boldsymbol{\mathcal{E}}_j) = \boldsymbol{\mathcal{E}}_j$, this is known as an **absolute error model**. Another popular choice for a statistical model is given by $\mathbf{g}(t_j; \boldsymbol{\theta}_0, \boldsymbol{\mathcal{E}}_j) = \mathbf{f}(t_j; \boldsymbol{\theta}_0) \circ \boldsymbol{\mathcal{E}}_j$ which results in a **relative error model**, where $\circ$ denotes component-wise multiplication of the vectors. In this work we will only concern ourselves with the special case of $\mathbf{g}(t_j; \boldsymbol{\theta}_0, \boldsymbol{\mathcal{E}}_j) = \mathbf{f}^\gamma(t_j; \boldsymbol{\theta}_0) \circ \boldsymbol{\mathcal{E}}_j$ so that

$$\mathbf{Y}_j = \mathbf{f}(t_j; \boldsymbol{\theta}_0) + \mathbf{f}^\gamma(t_j; \boldsymbol{\theta}_0) \circ \boldsymbol{\mathcal{E}}_j, \quad j = 1, 2, \dots, N, \quad (1.5)$$

where $\mathbf{f}^\gamma(t_j; \boldsymbol{\theta}_0) = (f_1^{\gamma_1}(t_j; \boldsymbol{\theta}_0), f_2^{\gamma_2}(t_j; \boldsymbol{\theta}_0), \dots, f_m^{\gamma_m}(t_j; \boldsymbol{\theta}_0))^T$ and $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_m)$. Here $\mathbf{f}^\gamma \circ$ denotes the component-wise exponentiation by γ of the vector function $\mathbf{f}$ followed by component-wise multiplication of the vectors $\mathbf{f}^\gamma(t_j; \boldsymbol{\theta}_0)$ and $\boldsymbol{\mathcal{E}}_j$.

There are numerous examples in which such statistical models have been found appropriate if a proper value of γ is chosen. These examples include modeling of HIV viral infections [5] for data consisting of CD4⁺ T cell counts with $\gamma_1 = 0$ and viral RNA counts with $\gamma_2 = 1.2$; prion aggregation kinetic models [7] with $\gamma_i = \gamma = 0.6$; insect populations undergoing pesticide treatments [3, 4] with $\gamma \approx .8$ or $.85$; and cell proliferation studies modeling flow cytometry data with $\gamma = 0.5$ [8, p. 87], [24] for a dividing population of lymphocytes labeled with the intracellular dye CFSE. In this latter example, as cells divide, the highly fluorescent intracellular CFSE is partitioned evenly between two daughter cells. A flow cytometer measures the CFSE fluorescence intensity (FI) of labeled cells as a surrogate for the mass of CFSE within a cell, thus providing an indication of the number of times a cell has divided. In these applications we note that γ is a to-be-determined tuning parameter for the statistical model. Furthermore, we assume that for any fixed j , $\mathcal{E}_{ij}$, $i = 1, 2, \dots, m$, are independent with mean zero and $\text{Var}(\mathcal{E}_{ij}) = \sigma_{0,i}^2$.

We remark that for the case where the measurement errors are heteroscedastic, the difference-based methods were specially designed for the general case (1.4). For these more general formulations the weights for the squared pseudo measurement errors (used in the estimation of the variance) involve kernel functions, and the bandwidth (a free parameter) of the kernel function has a strong influence on the resulting estimate. To our knowledge, a proper choice of bandwidth is still a current topic of research. In addition, variance estimation is not the focus of this paper.

To represent a collected data set, we let $\mathbf{y}_j = (y_{1j}, y_{2j}, \dots, y_{mj})^T$ be a realization of $\mathbf{Y}_j$,

$$\mathbf{y}_j = \mathbf{f}(t_j; \boldsymbol{\theta}_0) + \mathbf{f}^\gamma(t_j; \boldsymbol{\theta}_0) \circ \boldsymbol{\varepsilon}_j, \quad j = 1, 2, \dots, N,$$

where $\boldsymbol{\varepsilon}_j = (\varepsilon_{1j}, \varepsilon_{2j}, \dots, \varepsilon_{mj})^T$ is a realization of $\boldsymbol{\mathcal{E}}_j$, $j = 1, 2, \dots, N$. We remark that difference-based methods for calculation of pseudo measurement errors can also be used in the case where the components of $\mathbf{Y}_j$ may be observed at different measurement points (that is, $Y_{ij} = f_i(t_{ij}; \boldsymbol{\theta}_0) + f_i^{\gamma_i}(t_{ij}; \boldsymbol{\theta}_0)\boldsymbol{\varepsilon}_{ij}$, $j = 1, 2, \dots, N$, $i = 1, 2, \dots, m$). But for notational convenience, we only consider the case where all the components of $\mathbf{Y}_j$ are observed at the same measurement time points.

Of course, our model $\mathbf{f}(t_j; \boldsymbol{\theta}_0)$ typically does not perfectly describe the underlying process in question. This results in what we will refer to as mathematical model misspecification or discrepancy. Additionally, even in our assumption that the measurement errors can be described by (1.5), the value of γ is not known a priori. We will refer to this as statistical model misspecification or discrepancy.

For the sake of clarity, we present a basic example to make our motives clear. Consider a population under observation which is believed to follow the well known logistic model given by

$$\dot{x}(t) = bx(t) \left(1 - \frac{x(t)}{\kappa} \right), \quad x(t_s) = x_0. \quad (1.6)$$

Here x denotes the number of individuals, r is the intrinsic growth rate, and κ represents the carrying capacity and we assume that we can observe the number of individuals at times t_j , $j = 1, 2, \dots, N$. In this case we would have the statistical model

$$Y_j = f(t_j; \boldsymbol{\theta}_0) + f(t_j; \boldsymbol{\theta})^\gamma \boldsymbol{\varepsilon}_j, \quad j = 1, 2, \dots, N, \quad (1.7)$$

where $f(t_j; \boldsymbol{\theta}) = x(t_j; \boldsymbol{\theta})$ denotes the solution to (1.6), (i.e. $\mathcal{C} = I$), and $\boldsymbol{\theta} = (b, \kappa, x_0)^T$ are the unknown mathematical model parameters. Recall that if $\gamma = 0$ we have an absolute error model, which is interpreted as meaning that the observation errors are independent of the size of the population itself, and if $\gamma = 1$ we have a relative error model, which indicates that the observation errors are a multiple of the population size itself. In general, for any $\gamma > 0$ we are making the assumption that the observation errors are dependent on the size of the population itself. Notice that $\mathbb{E}(Y_j) = f(t_j; \boldsymbol{\theta}_0)$, and $\text{Var}(Y_j) = f(t_j; \boldsymbol{\theta}_0)^{2\gamma} \sigma_0^2$, thus the variance is non-constant provided $\gamma \neq 0$.

In practice, one often assumes a specific statistical model (i.e., assumes a specific value for γ_i) and then chooses an appropriate method to carry out parameter estimation (for example, if $\gamma_i = 0$, $i = 1, 2, \dots, m$, then an ordinary least squares method is appropriate). One can then use residual plots to determine whether or not the assumed statistical model is appropriate (e.g., see [8, Chapter 3]). If one assumes a statistical model with $\gamma_i = 0$, $i = 1, 2, \dots, m$, and the resulting residual plots exhibit a non random pattern, such as a fan shaped pattern, then the assumed statistical model is not reasonable. In the case where the assumed statistical model is inappropriate, one tries another set of values for γ_i 's and carries out another inverse problem. This may be done iteratively until one finds values $\hat{\gamma}_i$'s such that the plot of the modified residuals $r_{ij}/|y_{ij} - f_i(t_j; \hat{\boldsymbol{\theta}})|^{\hat{\gamma}_i}$ versus t_j forms a horizontal band around the horizontal axis (where $r_{ij} = y_{ij} - f_i(t_j; \hat{\boldsymbol{\theta}})$ is the residual and $\hat{\boldsymbol{\theta}}$ denotes the estimate for $\boldsymbol{\theta}_0$). We thus see that if the involved inverse problem is computationally expensive, then this procedure can become prohibitive to implement. In addition, as we see later in this paper, residual plots may provide incorrect information for the variance of measurement errors in cases where mathematical model misspecification is present. In this

paper, we propose to use the information provided by the pseudo measurement errors to directly determine appropriate values for the γ_i 's. We note that difference-based methods used to calculate pseudo measurement errors *do not involve any inverse problem calculations* and are *independent* of the chosen mathematical model. Hence, if our proposed method is successful, it should provide a much more efficient and accurate way to determine appropriate values for the γ_i 's.

After determining appropriate values for the γ_i 's (i.e., an appropriate statistical model), one may then use pseudo measurement errors to determine whether there is a mathematical model error. Specifically, one first uses the appropriate statistical model (chosen based on the information provided by pseudo measurement errors) to carry out parameter estimation. Then one compares the residual plot and the plot of $\hat{\varepsilon}_{ij}$ versus t_j to determine whether there is a mathematical model misspecification. For example, if there is a discernible divergence between these two plots, then there is a mathematical model error and the degree of this error is determined by the degree of the divergence. Specifically, the difference between residuals and $\hat{\varepsilon}_{ij}$'s provides some prior information on this error. However, we remark that no discernible divergence between these two plots *does not imply* that there is no mathematical model error. As we shall see, there are some cases where two different mathematical models could give the same solution in the given sampling time region.

In the presence of mathematical model misspecification, we also propose to use the information provided by pseudo measurement errors to preform bootstrapping to quantify the uncertainty in parameter estimates. Bootstrapping is a popular tool for construction of confidence intervals for parameter estimators (e.g., see [8, Chapter 3] and [17]). It involves constructing a family of samples or simulated data sets based on random sampling with replacement. One uses each of these samples to solve the inverse problem to obtain a new estimate, and then constructs the confidence intervals based on this family of estimates for the parameter estimators. We remark that there are two common ways to construct a family of samples or simulated data sets. One involves resampling the original data set, and it is based on the assumption that the original data are independent and identically distributed (i.i.d.). However, this method does not work well for cases where models are used to describe dynamic systems as observations are often not identically distributed even in the case where measurement errors are independent and identically distributed. The other method involves resampling residuals from an initial estimation to the original data set, and this is based on the assumptions that the regression function correctly specifies the observed part of the system and the first two moments of measurement errors are correctly specified (that is, the given statistical model is correctly specified). We thus see that this method does not work for the case where there is a mathematical modeling misspecification, which is often the situation in describing a real system. To alleviate this difficulty, we propose to use difference-based methods to obtain the pseudo measurement errors and then create bootstrapping samples using random sampling with replacement from these pseudo measurement errors.

The remainder of this paper is organized as follows. In Section 2, we give an introduction to difference-based methods for variance calculations. We then apply these methods in Section 3 to simulated data sets to verify the accuracy of these methods and then use the obtained pseudo measurement errors to determine proper statistical models for several experimental data sets. In Section 4, we use the information provided by these pseudo measurement errors to determine whether there is possible mathematical model error and

illustrate how to carry out bootstrapping to quantify the uncertainty of corresponding parameter estimators in the presence of model misspecification. Finally, in Section 5 we conclude the paper with some summary remarks and future research efforts.

2 Difference-Based Methods

All of the difference-based methods are based on the assumption that the true or nominal regression function $\mathbf{f}$ is sufficiently smooth and the maximum of the length of sampling time intervals (i.e., $\max\{t_{j+1}-t_j, j = 1, 2, \dots, N-1\}$) is sufficiently small. The estimated variance is defined as the weighted average of the squared normalized differences of $\nu+1$ observations. Specifically, the normalized differences of $\nu+1$ observations, *pseudo measurement errors*, are defined as being either symmetric around y_{ij} as in

$$\hat{\varepsilon}_{ij} = \sum_{k=0}^{\nu} w_k y_{i,j-\left[\frac{\nu+1}{2}\right]+k+1}, \quad j = \left[\frac{\nu}{2}\right], \left[\frac{\nu}{2}\right] + 1, \dots, N - \left[\frac{\nu}{2}\right], \quad i = 1, 2, \dots, m, \quad (2.1)$$

or asymmetric about y_{ij} as in

$$\hat{\varepsilon}_{ij} = \sum_{k=0}^{\nu} w_k y_{i,j+k}, \quad j = 1, 2, \dots, N - \nu, \quad i = 1, 2, \dots, m. \quad (2.2)$$

Here $[a]$ denotes the smallest integer that is greater than or equal to a , and the w_k 's are some real numbers which satisfy the conditions

$$\sum_{k=0}^{\nu} w_k = 0, \quad \sum_{k=0}^{\nu} w_k^2 = 1.$$

We remark that the above conditions are necessary to obtain an asymptotically unbiased estimator for the variance (e.g., see [21]) and so that the choice of the form for the pseudo measurement errors (i.e., choosing either (2.1) or (2.2)) does not affect the asymptotic result (e.g., see [19]).

Based on the choice for the values of ν and w_k 's, we introduce here three of these difference-based methods. One of them involves first-order backward differencing data (i.e., $\nu = 1$) and the associated pseudo measurement errors are calculated as follows:

$$\hat{\varepsilon}_{ij}^{1st} = \frac{1}{\sqrt{2}}(y_{i,j-1} - y_{i,j}), \quad j = 2, 3, \dots, N, \quad i = 1, 2, \dots, m. \quad (2.3)$$

For the case $\gamma_i = 0$ (i.e., constant variance error), the estimate for $\sigma_{0,i}$ is then given by

$$\hat{\sigma}_{0,i}^{1st} = \sqrt{\frac{1}{N-1} \sum_{j=2}^N (\hat{\varepsilon}_{ij}^{1st})^2}, \quad i = 1, 2, \dots, m. \quad (2.4)$$

Another method involves second-order differencing data; that is $\nu = 2$. The pseudo measurement errors are given by

$$\hat{\varepsilon}_{ij}^{2nd} = \frac{1}{\sqrt{6}}(y_{i,j-1} - 2y_{ij} + y_{i,j+1}), \quad j = 2, 3, \dots, N-1, \quad i = 1, 2, \dots, m. \quad (2.5)$$

For the case $\gamma_i = 0$, the associated estimate for $\sigma_{0,i}$ is then given by

$$\hat{\sigma}_{0,i}^{2\text{nd}} = \sqrt{\frac{1}{N-2} \sum_{j=2}^{N-1} (\hat{\varepsilon}_{ij}^{2\text{nd}})^2}, \quad i = 1, 2, \dots, m. \quad (2.6)$$

The last method we consider involves applying a first-order differencing operator l times (i.e. a l th order difference), where l is some integer. Let Δ^l denote the first order differencing operator applied l times with $\Delta y_{ij} = y_{i,j+1} - y_{ij}$, $j = 1, 2, \dots, N-1$, $i = 1, 2, \dots, m$. Then pseudo measurement errors are calculated as follows:

$$\hat{\varepsilon}_{ij}^{1\text{st}-l} = \sqrt{\left(\frac{2l}{l}\right)^{-1}} \Delta^l y_j, \quad j = 1, 2, 3, \dots, N-l, \quad i = 1, 2, \dots, m, \quad (2.7)$$

where $\left(\frac{2l}{l}\right) = \frac{(2l)!}{(l!)^2}$ with $l!$ be the factorial of l . For the case $\gamma_i = 0$, the estimate for $\sigma_{0,i}$ is then given by

$$\hat{\sigma}_{0,i}^{1\text{st}-l} = \sqrt{\frac{1}{N-l} \sum_{j=1}^{N-l} (\hat{\varepsilon}_{ij}^{1\text{st}-l})^2}, \quad i = 1, 2, \dots, m. \quad (2.8)$$

It was suggested in [23] that the choice of $l = 3$ works well in practice. This is also the value that we use for l in all the numerical results shown in this paper.

We remark that for the case where one suspects that observation coordinates of $\mathbf{Y}_j$ may have different constant variance (i.e., $\sigma_{0,i}$ may not be equal to $\sigma_{0,k}$), one usually uses an iterative process to carry out parameter estimation for mathematical model parameters and $\sigma_{0,i}$'s (e.g., see [8, Section 3.2.2] for details). However, by using the above methods one is able to obtain the estimates for $\sigma_{0,i}$'s and hence one does not need to use such an iterative inverse problem procedure. This can significantly speed up the desired inverse problem methodology.

3 Application on Determining an Appropriate Statistical Model to Use

In this section, we first apply the three methods introduced in Section 2 to some simulated data sets to demonstrate the accuracy of these methods as well as the accuracy of the proposed method in using the information provided by pseudo measurement errors to determine an appropriate statistical model. We then apply these methods to some experimental data sets to determine the appropriate statistical models.

3.1 Numerical Results for Simulated Data Sets

We first consider applying the three methods introduced in Section 2 to two simulated data sets: one is generated using the logistic growth model and the other is generated using a classical SIR model. These examples are used to support the accuracy of the proposed method for using the information provided by pseudo measurement errors to determine appropriate values for the γ_i 's in (1.5).

3.1.1 Example 1: Logistic Growth Model

We consider the standard logistic growth model

$$\dot{x} = bx \left(1 - \frac{x}{\kappa}\right), \quad x(t_s) = x_0 \quad (3.1)$$

as described in Section 1.

We assume that we can observe the number of individuals; that is, $f = x$. To generate the simulated data, we simulate (3.1) with parameter values and initial values chosen as $b = 3, \kappa = 100, x_0 = 10$. We then impose a normal distribution on $\mathcal{E}_j$ with zero mean and standard deviation $\sigma_0 = 0.05$, where the measurement time points are $t_j = t_s + (j-1)\frac{t_f - t_s}{N-1}$ with $t_s = 0, t_f = 2.5, j = 1, 2, \dots, N$ and $N = 201$. In other words, the simulated data set $\{y_j\}_{j=1}^N$ is generated as follows:

$$y_j = x(t_j) + x^\gamma(t_j)\varepsilon_j, \quad j = 1, 2, \dots, N, \quad (3.2)$$

where ε_j is a realization of $\mathcal{E}_j$. We first will confirm that the pseudo measurement errors provide a reasonable approximation of the true measurement errors ε_j . Then we show that the pseudo measurement errors can be used to determine the unknown value of γ .

Table 1 summarizes the estimates for σ_0 found using the three methods introduced in Section 2 for the case where $\gamma = 0$ in (3.2). From this table, we see that the last two methods give reasonable estimates for σ_0 while the first method considerably overestimates the value of σ_0 . But we do see from other numerical results (see the SIR example discussed in the next section) that this simple method does give a good estimate for some examples. But in summary, this suggests that this first-order differencing method does not consistently perform well. This is consistent with the observation made in [15], wherein it is suggested that the first-order differencing method should not be used since it does not always behave well.

σ_0	$\hat{\sigma}_0^{1st}$	$\hat{\sigma}_0^{2nd}$	$\hat{\sigma}_0^{1st-3}$
5.000e-02	3.930e-01	5.210e-02	5.216e-02

Table 1: Results for the logistic example in the case where $\gamma = 0$: the true value of σ_0 as well as its estimates obtained using the three mentioned methods ($\hat{\sigma}_0^{1st}$ is obtained using the first-order differencing method, $\hat{\sigma}_0^{2nd}$ is obtained using the second-order differencing method, and $\hat{\sigma}_0^{1st-3}$ is obtained by applying the first-order differencing operator for 3 times).

Figure 1 presents the plot of pseudo measurement errors $\hat{\varepsilon}_j^{1st}$ (obtained using the first-order differencing method and denoted as “estimates” in the legend of this figure) versus t_j and the plot of ε_j (simulated measurement errors and denoted as “true values” in the legend) versus t_j for the cases where $\gamma = 0$ (left column) and $\gamma = 1$ (right column) in (3.2). From this figure, we see that for the case where $\gamma = 0$ the plot of $\hat{\varepsilon}_j^{1st}$ versus t_j diverges from the plot of ε_j versus t_j except at the very end. This is consistent with considerable difference of the estimate $\hat{\sigma}_0^{1st}$ from its true value. However, for the case where $\gamma = 1$ the time plot for the pseudo measurement errors exhibits a pattern similar to that for the true values.

Figure 1: Comparison of the plot of $\hat{\varepsilon}_j^{1st}$ (denoted as “estimates” in the legend) versus t_j and the plot of ε_j (denoted as “true value”) versus t_j : (left panel) results obtained for $\gamma = 0$; (right panel) results obtained for $\gamma = 1$, where $\hat{\varepsilon}_j^{1st}$ ’s are obtained by the first-order differencing method.

This suggests that for this case this method works well. Figure 2 illustrates the results for the pseudo measurement errors $\hat{\varepsilon}_j^{2nd}$ obtained using second-order differencing method while Figure 3 depicts the results for the pseudo measurement errors $\hat{\varepsilon}_j^{1st-3}$ obtained by applying the first-order differencing operator for 3 times. From these two figures, we see that the time

Figure 2: Comparison of the plot of the $\hat{\varepsilon}_j^{2nd}$ (denoted as “estimates” in the legend) versus t_j and the plot of ε_j (denoted as “true value”) versus t_j for the case $\gamma = 0$ (left panel) and the case $\gamma = 1$ (right panel), where the $\hat{\varepsilon}_j^{2nd}$ are obtained using the second-order differencing method.

plot for the pseudo measurement errors exhibits the same pattern as that for the true ones. This suggests that these two methods work well at approximating the true measurement

Figure 3: Comparison of the plot of the $\hat{\varepsilon}_j^{1st-3}$ (denoted as “estimates” in the legend) versus t_j and the plot of ε_j (denoted as “true value”) versus t_j for the case $\gamma = 0$ (left panel) and the case $\gamma = 1$ (right panel), where the $\hat{\varepsilon}_j^{1st-3}$ are obtained by applying the first-order differencing operator for 3 times.

errors for both the constant variance error case and the relative error case.

Of course, in practice one does not have a value of the measurement errors themselves, rather only the data measurement $\{y_j\}_{j=1}^N$ is known. We claim that if the pseudo measurement errors provide a reasonable approximation of the true (but unknown) measurement errors, then we can use the pseudo measurement errors to determine an appropriate value for γ . For example, if the plot of pseudo measurement errors $\hat{\varepsilon}_j$ versus t_j seems to form a horizontal band around the horizontal axis, then $\gamma = 0$ may be appropriate. However, if one finds that the plot of $\hat{\varepsilon}_j$ versus t_j does not appear to be identically distributed, then $\gamma \neq 0$ and one needs to find a proper nonzero value for γ . To do this, we try different values for γ until one finds a value $\hat{\gamma}$ such that the plot of $\hat{\varepsilon}_j/|y_j - \hat{\varepsilon}_j|^{\hat{\gamma}}$ versus t_j forms a horizontal band around the horizontal axis. To verify whether or not this works, we take this logistic model with simulated data generated using $\gamma = 1$ as an example. We note that for this data set the plot of $\hat{\varepsilon}_j$ versus t_j (e.g., see the right panel of Figure 3) exhibits a fan shaped pattern. To begin with, we choose $\tilde{\gamma} = 2$. The resulting plot for $\eta_j^{\tilde{\gamma}} = \hat{\varepsilon}_j^{1st-3}/|y_j - \hat{\varepsilon}_j^{1st-3}|^{\tilde{\gamma}}$ versus t_j with $\tilde{\gamma} = 2$ is shown in the left panel of Figure 4, where the $\hat{\varepsilon}_j^{1st-3}$ are obtained by applying the first-order differencing operator three times. We observe from this plot that it has an inverted fan shaped pattern. This indicates that a proper value for γ is between 0 and 2. We then plotted $\eta_j^{\tilde{\gamma}}$ versus t_j with $\tilde{\gamma} = 1$ (shown in the right panel of Figure 4) and found that they appear to be identically distributed. Figure 5 depicts the results obtained using the second-order differencing method. We observe similar patterns. These results demonstrate that our proposed method works well. One may argue that this proposed method is somewhat subjective and that one can use difference-based methods to obtain directly the estimate for the variance of measurement errors using the kernel function approach mentioned in the Introduction. One could then use this estimate to obtain the estimate for γ . However, as we remarked in the Introduction, such an estimate for the variance highly depends on the value

Figure 4: Plot of $\eta_j^{\tilde{\gamma}} = \hat{\varepsilon}_j^{1st-3} / |y_j - \hat{\varepsilon}_j^{1st-3}|^{\tilde{\gamma}}$ versus t_j for the case where the simulated data were generated with $\gamma = 1$: (left panel) $\tilde{\gamma} = 2$; (right panel) $\tilde{\gamma} = 1$.

Figure 5: Plot of $\eta_j^{\tilde{\gamma}} = \hat{\varepsilon}_j^{2nd} / |y_j - \hat{\varepsilon}_j^{2nd}|^{\tilde{\gamma}}$ versus t_j for the case where the simulated data were generated with $\gamma = 1$: (left panel) $\tilde{\gamma} = 2$; (right panel) $\tilde{\gamma} = 1$.

of the bandwidth and that a proper choice of bandwidth is still a subject of research. Hence, our proposed method provides an alternative and practical way to determine a reasonable value for γ .

3.1.2 Example 2: SIR Model

We next consider a simple SIR model described by the following system of ordinary differential equations

$$\begin{aligned}\dot{S} &= -\beta SI, \\ \dot{I} &= \beta SI - \delta I, \\ \dot{R} &= \delta I, \\ (S(t_s), I(t_s), R(t_s)) &= (S_0, I_0, R_0).\end{aligned}\tag{3.3}$$

Here S , I and R respectively denote the ratios of the numbers of susceptible, infected, and recovered individuals to the total number of individuals (so they are dimensionless), β denotes the infection rate, and δ is the recover rate.

For demonstration purpose, we assume that we can observe all these three states; that is, $\mathbf{f} = (S, I, R)^T$. For all the results shown below, the parameter values and initial values are chosen as $\beta = 6, \delta = 3, S_0 = 0.7, I_0 = 0.2, R_0 = 0.1$. To generate the simulated data, we impose a normal distribution on $\mathcal{E}_{1j}$ with zero mean and standard deviation $\sigma_{0,1} = 0.01$, a normal distribution on $\mathcal{E}_{2j}$ with zero mean and $\sigma_{0,2} = 0.005$ and a normal distribution on $\mathcal{E}_{3j}$ with zero mean and $\sigma_{0,3} = 0.02$, where the measurement time points are $t_j = t_s + (j-1)\frac{t_f - t_s}{N-1}$ with $t_s = 0, t_f = 2$ and $N = 201, j = 1, 2, \dots, N$.

Table 2 summarizes the estimates for $\sigma_{0,i}$ by using the three methods introduced in Section 2 for the case where $\gamma_i = 0, i = 1, 2, 3$. From this table, we see that all three methods give reasonable estimates for $\sigma_{0,i}$.

$\sigma_{0,i}$	1.000e-02	5.000e-03	2.000e-02
$\hat{\sigma}_{0,i}^{1st}$	1.074e-02	5.295e-03	2.049e-02
$\hat{\sigma}_{0,i}^{2nd}$	1.039e-02	5.297e-03	2.009e-02
$\hat{\sigma}_{0,i}^{1st-3}$	1.043e-02	5.368e-03	1.996e-02

Table 2: Results for the SIR example in the case where $\gamma_i = 0, i = 1, 2, 3$: the value of $\sigma_{0,i}$ as well as its estimates obtained using the three methods introduced in Section 2.

Figure 6 presents the plots of pseudo measurement errors $\hat{\varepsilon}_{ij}^{1st}$ (obtained using first-order differencing method and denoted as “estimates” in the legend of this figure) versus t_j and the plot of ε_{ij} (simulated measurement errors and denoted as “true value” in the legend of this figure) versus t_j for the cases where $\gamma_1 = \gamma_2 = \gamma_3 = 0$ (left column) and $\gamma_1 = \gamma_2 = \gamma_3 = 1$ (right column). From this figure, we see that for the case $\gamma_1 = \gamma_2 = \gamma_3 = 0$ the time plot for $\hat{\varepsilon}_{ij}^{1st}$ exhibits exactly the same pattern as that for ε_{ij} . However, for the case $\gamma_1 = \gamma_2 = \gamma_3 = 1$, there is a discernible divergence between the plot of $\hat{\varepsilon}_{ij}^{1st}$ versus t_j and the plot of ε_{ij} versus t_j . This again demonstrates that the first-order differencing method does not consistently perform well. Figure 7 illustrates the results for the pseudo measurement errors $\hat{\varepsilon}_{ij}^{2nd}$ obtained using the second-order differencing method while Figure 8 depicts the results for the pseudo measurement errors $\hat{\varepsilon}_{ij}^{1st-3}$ obtained by applying the first-order differencing operator 3 times. From these two figures, we see that the time plots for the pseudo measurement errors exhibit

Figure 6: Comparison of the plot of $\hat{\varepsilon}_{ij}^{1st}$ (denoted as “estimates” in the legend) versus t_j and the plot of ε_{ij} (denoted as “true value”) versus t_j for the case $\gamma_1 = \gamma_2 = \gamma_3 = 0$ (left column) and the case $\gamma_1 = \gamma_2 = \gamma_3 = 1$ (right column).

the same pattern as that for the true ones. Hence, these two methods work well for both the constant variance error case and the relative error case.

Figure 7: Comparison of the plot of $\hat{\varepsilon}_{ij}^{2\text{nd}}$ (denoted as “estimates” in the legend) versus t_j and the plot of ε_{ij} (denoted as “true value”) versus t_j for the case $\gamma_1 = \gamma_2 = \gamma_3 = 0$ (left column) and the case $\gamma_1 = \gamma_2 = \gamma_3 = 1$ (right column).

Figure 8: Comparison of the plot of $\hat{\varepsilon}_{ij}^{1st-3}$ (denoted as “estimates” in the legend) versus t_j and the plot of ε_{ij} (denoted as “true value”) versus t_j for the case $\gamma_1 = \gamma_2 = \gamma_3 = 0$ (left column) and the case $\gamma_1 = \gamma_2 = \gamma_3 = 1$ (right column).

The above numerical results again reveal that if difference-based methods work, then the obtained pseudo measurement errors can be used to play the role as measurement errors and hence can be used to determine appropriate values for γ_i 's. Here we take this SIR model with simulated data generated using $\gamma_1 = \gamma_2 = \gamma_3 = 1$ as another example to further verify the method used in Section 3.1.1 for determining appropriate values for γ_i 's in the case where the pseudo measurement errors do not appear to be identically distributed. Specifically, we try different values for γ_i until one finds a value $\hat{\gamma}_i$ such that the plot of $\hat{\varepsilon}_{ij}/|y_{ij} - \hat{\varepsilon}_{ij}|^{\hat{\gamma}_i}$ versus t_j forms a horizontal band around the horizontal axis. We note that for this data set both the plot of $\hat{\varepsilon}_{1j}$ versus t_j and the plot of $\hat{\varepsilon}_{2j}$ versus t_j (e.g., see the right column of Figure 8) exhibit a fan shaped pattern while the plot of $\hat{\varepsilon}_{3j}$ versus t_j exhibits an inverted fan shaped pattern. For a start, we choose $\tilde{\gamma}_i = 2$, $i = 1, 2, 3$. The resulting plots for $\eta_{ij}^{\tilde{\gamma}_i} = \hat{\varepsilon}_{ij}^{1st-3}/|y_{ij} - \hat{\varepsilon}_{ij}^{1st-3}|^{\tilde{\gamma}_i}$ versus t_j with $\tilde{\gamma}_i = 2$ and $i = 1, 2, 3$, are shown in the left column of Figure 9, where the $\hat{\varepsilon}_{ij}^{1st-3}$ are obtained by applying the first-order differencing operator for three times. We observe from these plots that the patterns for all these three plots are inverted (i.e., the plot previously having a fan shaped pattern now has an inverted fan shaped pattern, and the plot previously having an inverted fan shaped pattern now have a fan shaped pattern). This indicates that a proper value for γ_i is between 0 and 2, $i = 1, 2, 3$. We then plotted $\eta_{ij}^{\tilde{\gamma}_i}$ versus t_j with $\tilde{\gamma}_1 = \tilde{\gamma}_2 = \tilde{\gamma}_3 = 1$ (shown in the right column of Figure 9) and found that they all appear to be identically distributed. Figure 10 depicts the results using the second-order differencing method. We observe similar patterns. These results again demonstrate that our proposed method of using pseudo measurement errors to determine an appropriate value for γ works well.

Figure 9: Plot of $\eta_{ij}^{\tilde{\gamma}_i} = \hat{\varepsilon}_{ij}^{1st-3} / |y_{ij} - \hat{\varepsilon}_{ij}^{1st-3}|^{\tilde{\gamma}_i}$ versus t_j for the case where the simulated data were generated with $\gamma_1 = \gamma_2 = \gamma_3 = 1$: (left panel) $\tilde{\gamma}_1 = \tilde{\gamma}_2 = \tilde{\gamma}_3 = 2$; (right panel) $\tilde{\gamma}_1 = \tilde{\gamma}_2 = \tilde{\gamma}_3 = 1$.

Figure 10: Plot of $\eta_{ij}^{\tilde{\gamma}_i} = \hat{\varepsilon}_{ij}^{2\text{nd}} / |y_{ij} - \hat{\varepsilon}_{ij}^{2\text{nd}}|^{\tilde{\gamma}_i}$ versus t_j for the case where the simulated data were generated with $\gamma_1 = \gamma_2 = \gamma_3 = 1$: (left panel) $\tilde{\gamma}_1 = \tilde{\gamma}_2 = \tilde{\gamma}_3 = 2$; (right panel) $\tilde{\gamma}_1 = \tilde{\gamma}_2 = \tilde{\gamma}_3 = 1$.

3.2 Numerical Results for Experimental Data Sets

In this section, we apply the difference-based methods to some experimental data sets to determine an appropriate value for γ . Since the first-order differencing method does not consistently perform well, we only use the second-order differencing method and the method for applying the first-order differencing operator 3 times for these data sets.

3.2.1 Daphnia magna Data Set

Here we consider the survival data collected for *Daphnia magna* that were presented in [1]. Specifically, ninety daphnids (neonates) were longitudinally observed and survival was recorded daily, and an ordinary least squares method was used in this paper to estimate the mortality rate.

Figure 11 presents the time plot results for the pseudo measurement errors obtained using the second-order differencing method (left) and the method for applying the first-order differencing operator three times (right). We observe from the right plot of this figure that pseudo measurement errors form a horizontal band around the line $\varepsilon = 0$. The similar pattern can be observed from the left plot of Figure 11 except several outliers. This indicates that the absolute error model (i.e., $\gamma = 0$) may be correct for this case. This provides support for using ordinary least squares method for parameter estimation in [1].

Figure 11: Time plots for pseudo measurement errors obtained for the *Daphnia* data set presented in [1]: (left panel) using the second-order differencing method; (right panel) using the method for applying the first-order differencing operator for three times (right).

3.2.2 CFSE Data Set

Here we consider the flow cytometry data presented in [8, Section 3.5.3] for a dividing population of lymphocytes labeled with the intracellular dye CFSE. The observable is the number of cells measured at time s_k with log-fluorescence intensity in the region $[z_i, z_{i+1})$, $k = 1, 2, \dots, 7$, $i = 1, 2, \dots, 513$. To be consistent with the notation introduced earlier,

we re-index the data collection points $\{(s_k, z_i)\}$ by $\{t_j\}$; that is, the elements in the set $\{t_j\}_{j=513(k-1)+1}^{513k}$ correspond to the elements in the set $\{(s_k, z_i)\}_{i=1}^{513}$, $k = 1, 2, \dots, 7$.

Figure 12 depicts results obtained using the second-order differencing method. Specifically, the left panel shows the plot of pseudo measurement errors $\hat{\varepsilon}_j^{2\text{nd}}$ versus j , and the right panel illustrates the plot of $\eta_j^{\tilde{\gamma}} = \hat{\varepsilon}_j^{2\text{nd}} / |y_j - \hat{\varepsilon}_j^{2\text{nd}}|^{\tilde{\gamma}}$ versus j with $\tilde{\gamma} = 0.5$, where the vertical lines delineate the pseudo measurement errors obtained in time intervals $[s_k, s_{k+1})$, $k = 1, 2, \dots, 7$. We observe from the left panel of Figure 12 that pseudo measurement errors

Figure 12: Plots for pseudo measurement errors obtained for the CFSE data set presented in [8, Section 3.5.3] by using the second-order differencing method: (left panel) plot of $\hat{\varepsilon}_j$ versus j ; (right panel) plot of $\eta_j^{\tilde{\gamma}} = \hat{\varepsilon}_j^{2\text{nd}} / |y_j - \hat{\varepsilon}_j^{2\text{nd}}|^{\tilde{\gamma}}$ versus j with $\tilde{\gamma} = 0.5$. The vertical lines delineate the pseudo measurement errors obtained in time intervals $[s_k, s_{k+1})$, $k = 1, 2, \dots, 7$.

are far from identically distributed. This is also true even in each subinterval $[s_k, s_{k+1})$. Hence, $\gamma = 0$ is not a reasonable choice for this data set. This conclusion is consistent with the one made in [8, Section 3.5.3] where residual plots were used to determine an appropriate value for γ . Results in the right panel of Figure 12 imply that in each subinterval $[s_k, s_{k+1})$ the $\eta_j^{\tilde{\gamma}}$'s appear to be identically distributed except some outliers. This indicates that $\gamma = 0.5$ may be appropriate in each of these subintervals. We also observe from this plot that the bandwidth formed by the plot of $\eta_j^{\tilde{\gamma}}$'s in the last two subintervals are larger than the ones for those subintervals located in the middle, which suggests that we may have different constant variance for $\eta_j^{\tilde{\gamma}}$ in these subintervals. This is inconsistent with the conclusion made in [8, Section 3.5.3] where residual plots reveal that the modified residuals $r_j / |y_j - f(t_j; \hat{\theta})|^{\tilde{\gamma}}$ with $\tilde{\gamma} = 0.5$ appear to be identically distributed in the whole interval. Figure 13 demonstrates the results obtained by applying the first-order differencing operator three times. We observe similar patterns as that obtained by the second-order differencing method. As we discussed earlier in the Introduction, this inconsistency suggests that there may be a mathematical model misspecification involved.

Motivated by the recent CFSE label division studies of Bocharov, Luzyanina and colleagues [13, 20] on asymmetric label division and time-lags to division in proliferating mouse cell data, we further investigated the data sets presented in [8, 9, 11, 18] and discussed then

here. The Bocharov, et al., investigations raised the question of whether cell proliferation models which allowed for asymmetric label division might be better suited to describe our human cell proliferation data. In [10] we revisited these data sets for the possibility of mathematical model misspecification. In these investigations we used statistically based model comparison tests and found seemingly contradictory results. In one third of the data sets studied, we found support for the hypothesis that mathematical models permitting asymmetric label division *did not* improve the fits-to-data. However, for two thirds of the data sets, it was found that allowing asymmetric division does appear to lead to statistically significantly better agreement with the data. While there may be other confounding factors, the findings of [10] support the suggestion that there may be mathematical modeling error in the earlier CFSE labeled cell proliferation studies of [8, 9, 11, 18]. Thus the findings in the present analysis are consistent with notion of a mathematical modeling misspecification in the earlier findings reported in [10].

Figure 13: Plots for pseudo measurement errors obtained for the CFSE data set presented in [8, Section 3.5.3] by applying the first-order differencing operators for three times: (left panel) plot of $\hat{\varepsilon}_j$ versus j ; (right panel) plot of $\eta_j^{\tilde{\gamma}} = \hat{\varepsilon}_j^{1st-3} / |y_j - \hat{\varepsilon}_j^{1st-3}|^{\tilde{\gamma}}$ versus j with $\tilde{\gamma} = 0.5$. The vertical lines delineate the pseudo measurement errors obtained in time intervals $[s_k, s_{k+1})$, $k = 1, 2, \dots, 7$.

4 Bootstrapping in the Presence of Mathematical Model Misspecification

In this section, we use an example to demonstrate that one could use pseudo measurement errors to determine whether there is a possible mathematical model error in the case where the statistical model is known to be correct. In addition, we use this example to demonstrate how to use the information provided by the pseudo measurement errors to carry out bootstrapping to quantify the uncertainty of parameter estimators in the presence of mathematical model misspecification.

For demonstration purpose, we use the simulated data set that was generated by the logistic growth model (3.1) using an absolute error model ($\gamma = 0$). Specifically, the data set $\{y_j\}$ is generated as follows: we first simulate (3.1) with model parameters and initial values given by $b = 0.8, \kappa = 200, x_0 = 10$; we then impose a normal distribution on $\mathcal{E}_j$ with zero mean and standard deviation $\sigma_0 = 0.2$ to generate a realization ε_j of measure error $\mathcal{E}_j$ (i.e., constant variance error), where again the measurement time points are $t_j = t_s + (j-1)\frac{t_f - t_s}{N-1}$ with $t_s = 0, t_f = 2$ and $j = 1, 2, \dots, N$, with $N = 201$. We finally add this noise to the simulated model solution (that is, $y_j = x(t_j) + \varepsilon_j$ with $x(t)$ the solution to the logistic model (3.1)). The resulting data are illustrated in Figure 14.

Figure 14: Results for fitting exponential growth model (4.1) to the simulated data generated by the logistic growth model (3.1).

In practice, one has no idea how $\{y_j\}$ is generated. Based on the information provided by the plot of this data set, one may choose to use an exponential growth model to describe the data, i.e.,

$$\dot{\tilde{x}} = b\tilde{x}, \quad \tilde{x}(t_s) = x_0, \quad (4.1)$$

where the parameter b needs to be estimated from the data. Hence, the data is assumed to be generated by

$$Y_j = f(t_j; b_0) + \mathcal{E}_j, \quad j = 1, 2, \dots, N, \quad (4.2)$$

where $f(t; b) = \tilde{x}(t; b) = x_0 \exp(bt)$. In practice, one may have no knowledge of the measurement errors either. Since both second-order differencing method and the method for applying the first-order differencing operator 3 times work well, here we just use the second-order differencing method to determine a proper statistical model. The resulting plot for the pseudo measurement errors $\hat{\varepsilon}_j$ versus t_j is illustrated in the left panel of Figure 15. From this plot, we observe that the obtained pseudo measurement errors form a horizontal band around the horizontal axis. This implies that $\gamma = 0$ is appropriate to use (it conforms with the actual data collection process). Based on this information, we use the following ordinary

least squares method to do parameter estimation.

$$\hat{b} = \arg \min_{b \in [\underline{b}, \bar{b}]} \sum_{j=1}^N (y_j - f(t_j; b))^2, \quad (4.3)$$

where $\underline{b}$ and $\bar{b}$ are some constants. The resulting model fit is illustrated in Figure 14. From this figure, we see that we obtain a very good fit to the data. Hence, one might conclude there is no/negligible mathematical modeling error. However, when one plots the residuals $r_j = y_j - f(t_j; \hat{b})$ versus t_j (shown in the right panel of Figure 15), one clearly sees that residuals are far from identically distributed. This divergence between the residual plot and

Figure 15: (left panel): plot of pseudo measurement errors $\hat{\varepsilon}_j^{2nd}$ versus t_j , where ε_j^{2nd} 's are obtained by applying the second-order differencing method to the simulated data generated by the logistic growth model (3.1); (right panel): residual plot (i.e., a plot of r_j versus t_j) obtained by fitting exponential growth model (4.1) to the simulated data generated by logistic growth model (3.1).

the plot of $\hat{\varepsilon}_j^{2nd}$ versus t_j implies that there is a mathematical model error. In addition, this example clearly demonstrates that residual plots may give incorrect information for the variance of measurement errors in the case where there is a mathematical model error. That is, if one looked solely at the residual plots, the conclusion might be drawn that the statistical model has been incorrectly specified. Hence, one needs to be cautious when one attempts to use only residual plots to determine whether or not the assumed statistical model is appropriate.

We next consider how to quantify uncertainty in parameter estimates in the presence of mathematical model misspecification. Here we demonstrate a proposed method for how to use the information provided by the pseudo measurement errors to quantify uncertainty through bootstrapping. For simplicity, we take the scalar observation case (i.e., $m = 1$) as an example and assume an absolute error model. We use the second-order differencing method to obtain pseudo measurement errors. For a given data set $\{y_j\}_{j=1}^N$ that was generated by some unknown function $\psi(t; \boldsymbol{\vartheta}_0)$, we assume that we have a mathematical model $f(t; \boldsymbol{\theta}_0)$

which approximates $\psi(t; \boldsymbol{\vartheta}_0)$. That is, the data is generated by

$$y_j = \psi(t_j; \boldsymbol{\vartheta}_0) + \varepsilon_j, \quad j = 1, 2, \dots, N,$$

but since ψ is unknown, we assume the data was generated by

$$y_j = f(t_j; \boldsymbol{\theta}_0) + \varepsilon_j, \quad j = 1, 2, \dots, N, \quad (4.4)$$

where $f(t_j; \boldsymbol{\theta}_0)$ denotes the observed part of the solution of the chosen mathematical model with $\boldsymbol{\theta}_0 \in \mathbb{R}^{\kappa_\theta}$ (κ_θ is an integer) at the measurement point t_j . Algorithm 4.1 illustrates how to use the bootstrapping method to quantify uncertainty for parameter estimator in the presence of model misspecification when the pseudo measurement errors appear to be independent and identically distributed.

Algorithm 4.1 Bootstrapping in the presence of model misspecification

1. Apply the second-order differencing method to data set $\{y_j\}_{j=1}^N$ to obtain pseudo measurement errors $\hat{\varepsilon}_j^{2\text{nd}}$, $j = 2, 3, \dots, N-1$, and then use these pseudo measurement errors to obtain estimates for the true regression function at t_j 's by

$$\hat{\psi}_j = y_j - \hat{\varepsilon}_j^{2\text{nd}}, \quad j = 2, 3, \dots, N-1.$$

Set $k = 1$.

2. Create a bootstrapping sample using random sampling with replacement from $\{\hat{\varepsilon}_j^{2\text{nd}}\}_{j=2}^{N-1}$ to form a bootstrapping sample $\{\hat{\varepsilon}_2^{(k)}, \dots, \hat{\varepsilon}_{N-1}^{(k)}\}$.
3. Create bootstrap sample points

$$y_j^{(k)} = \hat{\psi}_j + \hat{\varepsilon}_j^{(k)}, \quad j = 2, \dots, N-1.$$

4. Obtain an estimate $\hat{\boldsymbol{\theta}}^{(k)}$ from the bootstrapping sample $\{y_j^{(k)}\}$ using the ordinary least squares method given by

$$\hat{\boldsymbol{\theta}}^{(k)} = \arg \min_{\boldsymbol{\theta} \in \Omega} \sum_{j=2}^{N-1} (y_j^{(k)} - f(t_j; \boldsymbol{\theta}))^2,$$

where Ω is some compact set in $\mathbb{R}^{\kappa_\theta}$.

5. Set $k = k + 1$ and repeat steps 2–5 until $k > K$ (e.g., typically $K = 1000$ as in our calculations below).
-

We then calculate the mean and covariance matrix for the bootstrapping estimator using

the formulae

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \frac{1}{K} \sum_{k=1}^K \hat{\boldsymbol{\theta}}^{(k)}, \\ \hat{\Sigma} &= \frac{1}{K-1} \sum_{k=1}^K (\hat{\boldsymbol{\theta}}^{(k)} - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{(k)} - \hat{\boldsymbol{\theta}})^T.\end{aligned}\tag{4.5}$$

The standard error for the i th component of the bootstrapping estimator is then given by $\sqrt{\hat{\Sigma}_{ii}}$, where $\hat{\Sigma}_{ii}$ is the (i, i) th component of $\hat{\Sigma}$, $i = 1, 2, \dots, \kappa_\theta$. It is worth emphasizing that Algorithm 4.1 can be used not only for the case where model misspecification is due to modeling error but also for the case where model misspecification is due to some approximations such as approximating infinite-dimensional parameters by finite-dimensional parameters as was considered in [6].

To illustrate the accuracy of this algorithm, we take the logistic example of this section, that is, ψ is the solution of the logistic model used to generate the data, and f is the solution to the exponential model which is used as the mathematical model. We compare the results obtained by Algorithm 4.1 and the results obtained by an algorithm similar to Algorithm 4.1 (referred to as Modified Algorithm 4.1) except that bootstrapping sample points are created by random sampling with replacement from the simulated measurement errors $\{\varepsilon_j\}_{j=2}^{N-1}$ and then adding them to the true regression function (that is, $y_j^{(k)} = \psi(t_j; \boldsymbol{\theta}_0) + \varepsilon_j^{(k)}$, $j = 2, 3, \dots, N-1$, where $\{\varepsilon_j^{(k)}\}_{j=2}^{N-1}$ is the bootstrapping sample set created by random sampling with replacement from $\{\varepsilon_j\}_{j=2}^{N-1}$). Table 3 illustrates the mean and standard error of the bootstrapping estimator for the logistic example of this section obtained using Algorithm 4.1 and Modified Algorithm 4.1 for different values of σ_0 . From this table, we see that the mean and standard error for the bootstrapping estimators obtained by Algorithm 4.1 are quite similar to those obtained by Modified Algorithm 4.1 for both cases considered. This implies that Algorithm 4.1 works well.

Value of σ_0	Algorithm	Mean	Standard error
$\sigma_0 = 0.2$	Algorithm 4.1	7.213e-01	4.095e-04
	Modified Algorithm 4.1	7.213e-01	3.978e-04
$\sigma_0 = 1$	Algorithm 4.1	7.218e-01	2.045e-03
	Modified Algorithm 4.1	7.220e-01	1.986e-03

Table 3: The mean and standard error of the bootstrapping estimator for the logistic example of this section obtained using Algorithm 4.1 and Modified algorithm 4.1.

We also observe from Table 3 that the ratios of the standard errors to the corresponding means are very small for both cases considered. This suggests that we obtain very reliable estimates for b_0 . However, the estimated value for b_0 for each of these cases is smaller than its true value 0.8 (i.e., the value used to generate the data). This clearly demonstrates that obtaining excellent fits to the data and very reliable estimates for parameters is not sufficient to obtain accurate estimates for the true values of the parameters with which to make reliable predictions. Thus, it is important to recognize mathematical modeling error no matter how small this error may be.

5 Concluding Remarks and Future Research Questions

In this paper, we demonstrate with a number of examples how to employ pseudo measurement errors to determine an appropriate statistical model. Numerical results demonstrate that this method works well in practice and is more efficient and accurate than the traditional method of using residual plots to investigate statistical model errors. Once a proper statistical model is determined, we then use the information provided by pseudo measurement errors to determine whether there is a mathematical model error. This is done through comparing the plot for pseudo measurement errors and residual plots. We demonstrate this with an example where the assumed mathematical model has a different form than the model that is used to generate the data. We found that even though the model fit is superb, there is a discernible difference between residual plots and the plot for pseudo measurement errors. Thus, this method provides a much more reliable way to detect a mathematical modeling error.

In the presence of model misspecification, we investigated how to use the information provided by pseudo measurement errors to quantify uncertainty in parameter estimators through bootstrapping methods. Numerical results demonstrate the effectiveness of this method. In addition, these results reveal that even in the case where the modeling error is small, it is important to recognize it in order to avoid obtaining biased estimates and hence making false predictions. One of our future efforts involve developing a method to account for this error so that one can obtain accurate estimates for the true values of the parameters and make reliable prediction. We remark that the effect of model misspecification on parameter estimates and model prediction has been addressed by a number of researchers (e.g., see [14] and the references therein). There have been numerous attempts to incorporate this error to obtain accurate estimates for parameters (e.g., see [2, 14, 23] and the recent review made in [16]). However, as remarked in [16], none of the approaches attempted so far is universally applicable due to the complexity and wide diversity of the underlying problems to be treated. In the future, we hope to use the information provided by pseudo measurement errors and the ideas in [14] to properly calibrate parameters in the presence of model misspecification. It was demonstrated with an example in [14] that with realistic priors on modelling error one might be able to obtain accurate estimates for the true values of parameters.

Acknowledgements

This research was supported in part by the Air Force Office of Scientific Research under grant number AFOSR FA9550-12-1-0188, and in part by the US Department of Education Graduate Assistance in Areas of National Need (GAANN) under grant number P200A120047. The authors wish to gratefully acknowledge the efforts of two referees of an earlier version of this ms. whose questions and suggestions resulted in significant improvements of the current ms.

References

- [1] K. Adoteye, H.T. Banks, K.B. Flores and G.A. LeBlanc, Estimation of time-varying mortality rates using continuous models for *Daphnia Magna*, CRSC-TR14-17, Center

- for Research in Scientific Computation, North Carolina State University, December, 2014; *Applied Mathematics Letters*, **44** (2015), 12–16. DOI:10.1016/j.aml.2014.12.01.
- [2] P.D. Arendt, D.W. Apley and W. Chen, Quantification of model uncertainty: calibration, model discrepancy and identifiability, *Journal of Mechanical Design*, **134** (2012), 100908, 12pp.
 - [3] H.T. Banks, J.E. Banks, Neha Murad, J. A. Rosenheim, and K. Tillman, Modelling pesticide treatment effects on *Lygus hesperus* in cotton fields, CRSC-TR15-09, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, September, 2015.
 - [4] H.T.Banks, J.E. Banks, J. Rosenheim, and K. Tillman, Modeling populations of *Lygus Hesperus* on cotton fields in the San Joaquin Valley of California: The importance of statistical and mathematical model choice, CRSC-TR15-04, Center for Research in Scientific Computation, N. C. State University, Raleigh, NC, May, 2015; *J. Biological Dynamics*, to appear.
 - [5] H.T. Banks, R. Baraldi, K. Cross, K. Flores, C. McChesney, L. Poag, and E. Thorpe, Uncertainty quantification in modeling HIV viral mechanics, CRSC-TR13-16, N. C. State University, Raleigh, NC, December, 2013; *Math. Biosciences and Engr.*, **12** (2015), 937–964.
 - [6] H.T. Banks, J. Catenacci and S. Hu, Asymptotic properties of probability measure estimators in a nonparametric model, *SIAM/ASA Journal on Uncertainty Quantification*, to appear.
 - [7] H.T. Banks, Marie Doumic, Carola Kruse, Stephanie Prigent, and Human Rezaei, Information content in data sets for a nucleated-polymerization model, CRSC-TR14-15, N. C. State University, Raleigh, NC, November, 2014; *J. Biological Dynamics*, (2015), DOI: 10.1080/17513758.2015.1050465
 - [8] H.T. Banks, S. Hu, and W.C. Thompson, *Modeling and Inverse Problems in the Presence of Uncertainty*, Chapman & Hall/CRC Press, Boca Raton, FL, 2014.
 - [9] H.T. Banks, D.F. Kapraun, K.G. Link, W.C. Thompson, C. Peligero, J. Argilaguet, and A. Meyerhans, Analysis of variability in estimates of cell proliferation parameters for cyton-based models using CFSE-based flow cytometry data, CRSC-TR13-14, North Carolina State University, November 2013; *J. Inverse Ill-Posed Probl.*, **23** (2014), 135–171; ISSN (Online) 1569-3945, ISSN (Print) 0928-0219, DOI: 10.1515/jiip-2013-0065, July 2014.
 - [10] H.T. Banks D.F. Kapraun, C. Peligero, J. Argilaguet, and A. Meyerhans, Evaluating the importance of mitotic asymmetry in cyton-based models for CFSE-based flow cytometry data, CRSC-TR15-02, N. C. State University, Raleigh, NC, February, 2015; *Intl. J. Pure and Applied Mathematics*, **100** (2015), No. 1; DOI: 10.12732/ijpam.v100i1.12

- [11] H.T. Banks, D.F. Kapraun, W.C. Thompson, C. Peligero, J. Argilaguet, and A. Meyerhans, A novel statistical analysis and interpretation of flow cytometry data, CRSC-TR12-23, North Carolina State University, March 2013; *J. Biol. Dyn.* **7** (2013), 96–132.
- [12] H.T. Banks and H.T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, CRC Press, Boca Raton, FL, 2009.
- [13] G. Bocharov, T. Luzyanina, J. Cupovic, and B. Ludewig, Asymmetry of cell division in CFSE-based lymphocyte proliferation analysis, *Front. Immunol.* **4** (2013), published online.
- [14] J. Brynjarsdóttir and A. O’Hagan, Learning about physical parameters: the importance of model discrepancy, *Inverse Problems*, **30** (2014), 114007, 24pp.
- [15] H. Dette, A. Munk and T. Wagner, Estimating the variance in nonparametric regression – what is a reasonable choice? *J.R. Statist. Soc. B*, **60** (1998), 751–764.
- [16] J. Doherty and D. Welter, A short exploration of structural noise, *Water Resource Research*, **46** (2010), W05525.
- [17] B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall/CRC Press, Boca Raton, 1998.
- [18] D.F. Kapraun, *Cell Proliferation Models, CFSE-Based Flow Cytometry Data, and Quantification of Uncertainty*, Ph.D. Dissertation, Department of Mathematics, North Carolina State University, Raleigh, NC (24 July 2014).
- [19] M. Levine, Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: A possible approach, *Computational Statistics & Data Analysis*, **50** (2006), 3405–3431.
- [20] T. Luzyanina, J. Cupovic, B. Ludewig, and G. Bocharov, Mathematical models for CFSE labelled lymphocyte dynamics: asymmetry and time-lag in division, *J. Math. Biol.* **69** (2014), 1547–83.
- [21] H.-G. Müller and U. Stadtmüller, Estimation of heteroscedasticity in regression analysis, *Annals of Statistics*, **15** (1987), 610–625.
- [22] T. Tong, A. Liu and Y. Wang, Relative errors of difference-based variance estimators in nonparametric regression, *Communications in Statistics – Theory and Methods*, **37** (2008), 2890–2902.
- [23] J.A. Vrugt, C.G.H. Diks, H.V. Gupta, W. Bouten, and J.M. Verstraten, Improved treatment of uncertainty in hydrologic modeling: combining the strengths of global optimization and data assimilation, *Water Resources Research*, **41** (2005), W01017.
- [24] W. Clayton Thompson, *Partial Differential Equation Modeling of Flow Cytometry Data from CFSE-based Proliferation Assays*, Ph.D. Dissertation, Dept. of Mathematics, North Carolina State University, Raleigh, December, 2011.