Asymptotic Properties of Probability Measure Estimators in a Nonparametric Model

H.T. Banks, Jared Catenacci and Shuhua Hu

Center for Research in Scientific Computation
North Carolina State University
Raleigh, NC 27695-8212 USA

May 16, 2014

Abstract

We consider probability measure estimation in a nonparametric model using a least-squares approach under the Prohorov metric framework. We summarize the computational methods and their convergence results that were developed by our group over the past two decades. New results are presented on the bias and the variance due to the approximation and the pointwise asymptotic normality of the approximated probability measure estimator. We propose use of model selection criterion to balance the bias and the variance, and compare the pointwise confidence band constructed using the asymptotic normality results with that obtained by Monte Carlo simulations.

Key Words: Least-squares method, Prohorov metric, model selection criterion, consistency, pointwise asymptotic normality, confidence band.

Mathematics Subject Classification: 34A55, 46S50, 62G05, 62G20

1 Introduction

We consider nonparametric estimation of an unknown probability measure in the case where the regression function is dependent on this measure. More precisely, the statistical model, the model describing the observation process, is described by

$$Y_j = f(t_j; P_0) + \mathcal{E}_j, \quad j = 1, 2, 3, \ldots, N. \tag{1.1}$$

In the above equation, $f(t_j; P_0)$ denotes the observed part of the solution of a mathematical model with the true probability measure P_0 (unknown) at the measurement point t_j, $\mathcal{E}_j$ is the measurement error at t_j, and N is the total number of observations, where $t_j \in [t_s, t_f]$, $j = 1, 2, 3, \ldots, N$, with t_s and t_f being some real numbers. Following popular conventions we will not always distinguish between probability measures and their associated cumulative distribution functions in the discussions below.

Equation (1.1) is often referred to as a nonparametric model (a model with all the unknown parameters being in an infinite-dimensional parameter space) in the statistics literature. Such models are motivated by a number of applications arising in biology and physics, for example, in modeling mosquitofish populations [9] and shrimp populations [6], in wave propagation in biotissue [15], in modeling of a complex nonmagnetic dielectric materials [4, 10], and in HIV cellular models [3]. Here we only elaborate one of the motivating examples, a recent project investigated by our group. In this project, the goal is to develop a noninvasive technique to characterize the degradation of a complex nonmagnetic dielectric material by assessing the small physical and chemical changes in the material using reflectance spectroscopy. This involves determining the components of the permittivity of the dielectric medium using the measured spectral responses. The relative permittivity of the dielectric medium is described by

$$\widehat{\varepsilon}_r(k; P_0) = \varepsilon_\infty - \int_{\Omega_\theta} \frac{k_p^2}{k^2 - ik/\tau - k_0^2} dP_0(\boldsymbol{\theta}). \tag{1.2}$$

In the above equation, ε_∞ denotes the relative permittivity of the dielectric medium at infinite frequency, k is the wavenumber ($k = \omega/(2\pi c)$, where ω is the angular frequency and c is the speed of light), k_0 represents the resonance wavenumber, and τ denotes the relaxation time. The composite parameter k_p is given by $k_p = k_0\sqrt{\varepsilon_s - \varepsilon_\infty}$ with ε_s being the relative permittivity of the medium at zero frequency, $i = \sqrt{-1}$ is the imaginary unit, and $\boldsymbol{\theta} = k_0 \in \Omega_\theta \subset \mathbb{R}$. If we assume that a monochromatic uniform wave is incident at an angle zero on a plane interface between free space and a nonmagnetic dielectric medium with the electric field polarized perpendicular to the plane of incidence, then the reflection coefficient is given by

$$r_s(k; P_0) = \frac{1 - \sqrt{\widehat{\varepsilon}_r(k; P_0)}}{1 + \sqrt{\widehat{\varepsilon}_r(k; P_0)}}, \tag{1.3}$$

where $\widehat{\varepsilon}_r$ is defined by (1.2). The observations f_j are the reflectance (the square of the magnitude of the reflection coefficient) at different wave numbers k_j; that is, $f_j = |r_s(k_j; P_0)|^2$. The goal is then to use these observations to estimate the unknown probability measure P_0.

We note here that the problem we outlined above is different from those, for example, in pharmacokinetics studies and HIV studies, where one desires to estimate both individual-specific parameters $\boldsymbol{\theta}$ (such as clearance rate of the virus and infection rate in HIV studies) and their associated probability distribution function P_0 from blood samples taken serially in time from *individuals* in the population (e.g., see [11, 19]). This is because in this case the data f_j is dependent on $\boldsymbol{\theta}$ instead of P_0; that is, one has *individual* longitudinal data instead of *aggregate* longitudinal data (i.e., data collected by sampling from the population at large). Hence, the methods used to solve these two types of problems are *fundamentally different*. We refer the interested reader to [12, 13] for more details on this topic.

Traditional parametric methods, which assume the sought-after probability measure P_0 has a particular distributional form, are not preferred as they are overly restrictive and will produce inaccurate results if the parametric form is misspecified. Here we propose to use a least-squares approach for nonparametric estimation of probability measures. Least-squares methods and maximum likelihood estimation (MLE) methods are two widely used frequentist-based approaches for parameter estimation. It is well-known that in the case that the parameter space is finite-dimensional both least-squares methods and maximum likelihood estimation methods have nice limiting properties for the parameter estimator, i.e., asymptotic normality and consistency (e.g., see [23]). However, MLE methods require knowledge of the probability density function of observations in order to define the likelihood function. Unfortunately, this knowledge is often not available in practice. In contrast, for least-squares methods, one only needs to assume the first two moments, i.e., the mean and variance or covariance matrix, of the measurement errors in order to define the cost function. To this point, we have not discussed the Bayesian approach, which is another widely used methodology for parameter estimation. The difficulties for a Bayesian analysis in a nonparametric model setting (the involved approach is often referred to as Bayesian nonparametric estimation) include prior (a stochastic process in this case) construction, algorithmic development (as it depends on the prior and the problem itself, and the common MCMC techniques cannot be directly applied to the aggregate data and infinite-dimensional parameter space setting) and posterior asymptotics. Thus, compared to frequentist approaches, the Bayesian nonparametric approach is more difficult to implement and hence we do not consider this approach in this presentation. We refer the interested reader to [18] for a review of recent developments on this topic.

The remainder of this paper is organized as follows. In Section 2, we first give an overview of the computational methods developed by our group in the past two decades (see [1]) for probability measure estimation, and provide a consistency result for the probability measure estimator. Then we discuss the bias and the variance due to the approximation and present the asymptotic normality of the approximated probability measure estimator. In Section 3, we give some numerical results to illustrate the efficacy of our approach. Finally, in Section 4 we conclude the paper with some summary remarks and future research questions.

2 Theoretical and Computational Framework for Probability Measure Estimation

For notational convenience, we assume that observations Y_j in (1.1) are scalar (the multi-dimensional case can be treated similarly). We also assume throughout the remainder of this discussion that measurement errors $\mathcal{E}_j$, $j = 1, 2, 3, \ldots, N$, are independent and identically distributed (i.i.d.) with zero mean and constant variance σ_0^2. We note that the existence of a true probability measure, a standard assumption in statistical formulations, along with the assumption that the measurement errors have zero mean implies that $f(t; P_0)$ correctly describes the observed part of the dynamical system (that is, the underlying mathematical model is correct).

With the i.i.d. assumption on the measurement errors, the estimator of P_0 can be obtained using the ordinary least-squares method as defined by

$$P^N = \arg\min_{P \in \mathbb{P}(\Omega_\theta)} \sum_{j=1}^{N} (Y_j - f(t_j; P))^2, \tag{2.1}$$

where $\mathbb{P}(\Omega_\theta)$ denotes the set of probability measures on the space $\Omega_\theta \subset \mathbb{R}^{\kappa_\theta}$ with κ_θ being a positive integer. We remark that P^N itself is random in that it is a function of random variables Y_j (and hence $\mathcal{E}_j$) on a probability space $(\Omega, \mathcal{F}, \text{Prob})$. The corresponding realization $\hat{P}^N$ of P^N can be calculated through

$$\hat{P}^N = \arg\min_{P \in \mathbb{P}(\Omega_\theta)} \sum_{j=1}^{N} (y_j - f(t_j; P))^2, \tag{2.2}$$

where y_j is a realization of Y_j, $j = 1, 2, 3, \ldots, N$. Thus, we can view P^N as a stochastic process (i.e., $P^N(\boldsymbol{\theta}; \cdot)$ as a one parameter ($\boldsymbol{\theta} \in \Omega_\theta$) family of random variables on the probability space $(\Omega, \mathcal{F}, \text{Prob})$) since each of its realizations yields a probability measure $\hat{P}^N \in \mathbb{P}(\Omega_\theta)$.

The existence of a minimizer to the least-squares optimization problem (2.1) or (2.2) can be established under the Prohorov metric framework. The Prohorov metric was introduced by the Russian probabilist Y.V. Prohorov [22] and is defined as follows (e.g., see [16, pp. 237–238]).

Definition 2.1. Let $\mathbb{F} \subset \Omega_\theta$ be any closed set and define $\mathbb{F}^\epsilon$ as follows:

$$\mathbb{F}^\epsilon = \left\{ \boldsymbol{\theta} \in \Omega_\theta : \inf_{\tilde{\boldsymbol{\theta}} \in \mathbb{F}} d(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) < \epsilon \right\},$$

where d denotes the metric on Ω_θ. For $P, Q \in \mathbb{P}(\Omega_\theta)$, the *Prohorov metric* is given by

$$\begin{aligned} &\rho(P, Q) \\ &\quad = \inf \{ \epsilon > 0 | Q(\mathbb{F}) \leq P(\mathbb{F}^\epsilon) + \epsilon \text{ and } P(\mathbb{F}) \leq Q(\mathbb{F}^\epsilon) + \epsilon, \text{ for all } \mathbb{F} \text{ closed in } \Omega_\theta \}. \end{aligned}$$

It is clear from the definition above that the meaning of Prohorov metric is far from intuitive. Yet one can provide several useful characterizations. For example, convergence in the Prohorov metric is equivalent to the weak* convergence if we view $\mathbb{P}(\Omega_\theta) \subset C_B^*(\Omega_\theta)$, where $C_B^*(\Omega_\theta)$ denotes the topological dual of the space $C_B(\Omega_\theta)$ of bounded and continuous functions on Ω_θ. In other words, the statement $\rho(P_j, P) \to 0$ is equivalent to the statement

$$\int_{\Omega_\theta} h(\boldsymbol{\theta}) dP_j(\boldsymbol{\theta}) \to \int_{\Omega_\theta} h(\boldsymbol{\theta}) dP(\boldsymbol{\theta}) \text{ for any } h \in C_B(\Omega_\theta).$$

The Prohorov metric also possesses many useful and important properties. For example, if we assume that Ω_θ is compact, then $\mathbb{P}(\Omega_\theta)$ is a compact metric space when taken with the Prohorov metric ρ. We refer interested readers to [1, 12, 16, 20, 26] for more information on the Prohorov metric and its properties. Based on these discussions, we see that if Ω_θ is compact and f is continuous with respect to P, then there exists a solution to (2.1) or (2.2).

2.1 Consistency of the Probability Measure Estimator

The ideas for establishing the consistency of probability measure estimators follow closely those given in [7] and [12]. Here we only present the necessary assumptions as well as the result, and refer the interested reader to [12, Section 5.5] for details. For a given *sampling set* $\{t_j\}_{j=1}^N$ in the interval $[t_s, t_f]$, one can define the empirical distribution function

$$\mu_N(t) = \frac{1}{N} \sum_{j=1}^{N} \Delta_{t_j}(t), \tag{2.3}$$

where Δ_{t_j} is the Dirac measure with atom at t_j; that is,

$$\Delta_{t_j}(t) = \begin{cases} 0, & t < t_j \\ 1, & \text{otherwise.} \end{cases}$$

Clearly, $\mu_N \in \mathbb{P}([t_s, t_f])$, the space of probability measures (or, equivalently the cumulative distribution functions) on $[t_s, t_f]$. We assume

(A1) For each fixed N, $\mathcal{E}_j$, $j = 1, 2, \ldots, N$, are independent and identically distributed with zero mean and constant variance σ_0^2, and they are defined on some probability space $(\Omega, \mathcal{F}, \text{Prob})$.

(A2) There exists a measure μ on $[t_s, t_f]$ such that

$$\frac{1}{N} \sum_{j=1}^{N} h(t_j) = \int_{t_s}^{t_f} h(t) d\mu_N(t) \to \int_{t_s}^{t_f} h(t) d\mu(t)$$

for all $h \in C([t_s, t_f])$.

(A3) The space $\Omega_\theta \subset \mathbb{R}^{\kappa_\theta}$ is compact; the space $\mathbb{P}(\Omega_\theta)$ is taken with the Prohorov metric ρ.

(A4) The function f satisfies $f \in C([t_s, t_f], C(\mathbb{P}(\Omega_\theta)))$.

(A5) The functional J^0 defined by

$$J^0(P) = \sigma^2 + \int_{t_s}^{t_f} \left(f(t; P_0) - f(t; P)\right)^2 d\mu(t)$$

is uniquely minimized at $P_0 \in \mathbb{P}(\Omega_\theta)$.

Theorem 2.2. *Under assumptions (A1)–(A5),* $\rho(P^N, P_0) \xrightarrow{a.s.} 0$ *as* $N \to \infty$, *where* $\xrightarrow{a.s.}$ *denotes convergence almost surely in* $(\Omega, \mathcal{F}, Prob)$. *That is,*

$$Prob\left\{\omega \in \Omega \;\middle|\; \lim_{N\to\infty} \rho(P^N(\omega), P_0) = 0\right\} = 1.$$

2.2 Approximation Schemes for Probability Measure Estimation

We note that (2.2) is an infinite-dimensional optimization problem. Hence, the infinite-dimensional space $\mathbb{P}(\Omega_\theta)$ must be approximated by some finite dimensional space $\mathbb{P}_M(\Omega_\theta)$ so that one has a computationally tractable finite-dimensional optimization problem given by

$$\hat{P}_M^N \;=\; \arg\min_{P \in \mathbb{P}_M(\Omega_\theta)} \sum_{j=1}^{N} (y_j - f(t_j; P))^2. \tag{2.4}$$

However, one needs to choose $\mathbb{P}_M(\Omega_\theta)$ in a meaningful way so that $\hat{P}_M^N$ approaches the solution to (2.2) as $M \to \infty$.

One such approximation method involves using Dirac measures to approximate the probability measures. The following theorem is useful in establishing the convergence results as well as in constructing approximation schemes. We refer the interested reader to [2] for a proof of this result.

Theorem 2.3. *Assume* $\Omega_\theta \subset \mathbb{R}^{\kappa_\theta}$ *is compact. Let* $\Omega_{\theta D} = \{\boldsymbol{\theta}_j\}_{j=1}^{\infty}$ *be an enumeration of a countable dense subset of* Ω_θ. *Define*

$$\tilde{\mathbb{P}}_D(\Omega_\theta) = \left\{ P \in \mathbb{P}(\Omega_\theta) \middle| P = \sum_{j=1}^{M} a_j \Delta_{\boldsymbol{\theta}_j}, \boldsymbol{\theta}_j \in \Omega_{\theta D}, a_j \in [0,1] \cap \mathbb{Q}, \sum_{j=1}^{M} a_j = 1, M \in \mathbb{N} \right\},$$

where $\Delta_{\boldsymbol{\theta}_j}$ *is the Dirac measure with atom at* $\boldsymbol{\theta}_j$, *and* $\mathbb{Q} \subset \mathbb{R}$ *denotes the set of all rational numbers. (That is,* $\tilde{\mathbb{P}}_D(\Omega_\theta)$ *is the collection of all convex combinations of Dirac measures on* Ω_θ *with atoms* $\boldsymbol{\theta}_j \in \Omega_{\theta D}$ *and rational weights.) Then* $\tilde{\mathbb{P}}_D(\Omega_\theta)$ *is dense in* $(\mathbb{P}(\Omega_\theta), \rho)$, *and thus* $\mathbb{P}(\Omega_\theta)$ *is separable.*

Under this Dirac measure approximation framework, we define $\mathbb{P}_M(\Omega_\theta)$ to be the set of all atomic probability measures with nodes placed at the first M elements in the enumeration of the countable dense subset of Ω_θ; that is,

$$\mathbb{P}_M(\Omega_\theta) = \left\{ P \in \mathbb{P}(\Omega_\theta) \;\middle|\; P = \sum_{j=1}^{M} a_j \Delta_{\boldsymbol{\theta}_j}, \text{ where } a_j \geq 0 \text{ and } \sum_{j=1}^{M} a_j = 1 \right\}. \tag{2.5}$$

By Theorem 2.3 we know that we can approximate any element $P \in \mathbb{P}(\Omega_\theta)$ by a sequence $\{P_{M_j}\}$, $P_{M_j} \in \mathbb{P}_{M_j}(\Omega_\theta)$, such that $\rho(P_{M_j}, P) \to 0$ as $M_j \to \infty$. We also see that this Dirac measure approximation method can be used regardless of the smoothness of probability measures. This is especially useful in the situations where one has no knowledge of the sought-after probability measures.

Another family of approximation methods is based on linear spline approximations, which are designed for the case where the sought after probability measures are absolutely continuous so that their corresponding probability density functions exist. The following theorem is useful in establishing the convergence results as well as in constructing approximation schemes. We refer the interested reader to [15] for a proof of this result.

Theorem 2.4. *Assume $\Omega_\theta \subset \mathbb{R}^{\kappa_\theta}$ is compact. Define*

$$\tilde{\mathbb{P}}_S(\Omega_\theta) = \left\{ P \in \mathbb{P}(\Omega_\theta) \middle| P'(\boldsymbol{\theta}) = \sum_{j=1}^{M} a_j l_j^M(\boldsymbol{\theta}), a_j \in [0,\infty) \cap \mathbb{Q}, \sum_{j=1}^{M} a_j \int_{\Omega_\theta} l_j^M(\boldsymbol{\xi}) d\boldsymbol{\xi} = 1, M \in \mathbb{N} \right\},$$

where P' denotes the derivative of P with respect to $\boldsymbol{\theta}$, the $\{l_j^M\}$ denote the usual piecewise linear splines, and $\mathbb{Q} \subset \mathbb{R}$ denotes the set of all rational numbers. Then $\tilde{\mathbb{P}}_S(\Omega_\theta)$ is dense in $\mathbb{P}(\Omega_\theta)$.

Under this linear spline approximation framework, we define $\mathbb{P}_M(\Omega_\theta)$ to be

$$\mathbb{P}_M(\Omega_\theta) = \left\{ P \in \mathbb{P}(\Omega_\theta) \;\middle|\; P'(\boldsymbol{\theta}) = \sum_{j=1}^{M} a_j l_j^M(\boldsymbol{\theta}), \text{ where } a_j \geq 0 \text{ and } \sum_{j=1}^{M} a_j \int_{\Omega_\theta} l_j^M(\boldsymbol{\xi}) d\boldsymbol{\xi} = 1 \right\}. \tag{2.6}$$

By Theorem 2.3 we know that we can approximate any element $P \in \mathbb{P}(\Omega_\theta)$ by a sequence $\{P_{M_j}\}$, $P_{M_j} \in \mathbb{P}_{M_j}(\Omega_\theta)$, such that $\rho(P_{M_j}, P) \to 0$ as $M_j \to \infty$.

The following theorem provides the desired convergence result, and it follows immediately from the Prohorov metric framework and convergence theorems of [14] as well as the results above (Theorems 2.3 and 2.4).

Theorem 2.5. *Assume Ω_θ is compact and $\mathbb{P}(\Omega_\theta)$ is taken with the Prohorov metric. If f is continuous with respect to P, then there exists a minimizer $\hat{P}_M^N$ to (2.4), where $\mathbb{P}_M(\Omega_\theta)$ is chosen as either (2.5) or (2.6). Moreover, the sequence $\{\hat{P}_M^N\}$ has at least one convergent subsequence, and the limit $\hat{P}^{N*}$ of such a subsequence is a minimizer to the least-squares problem (2.2).*

Remark 2.6. Both the Dirac measure approximation methods and the spline-based approximation methods have been successfully used to estimate probability measures in a number of applications (e.g., see [4, 8, 9, 10, 15]). However, it was demonstrated in [5] that if the sought-after probability measure is absolutely continuous, then the spline-based approximation methods converge much faster than do the Dirac measure approximation methods (in terms of the value of M). In addition, it was observed in [5] that the spline-based approximation methods also provide convergence for the associated probability density functions while the Dirac measure approximation methods do not do this. This is not surprising since in the spline-based approximation methods one directly approximates the associated probability density functions instead of the cumulative distribution functions.

2.3 Bias and Variance in Probability Measure Estimation

As we discussed in the above section, what one actually does in practice is to minimize the cost functional in a finite-dimensional space; that is, one solves the optimization problem

$$P_M^N \;=\; \arg\min_{P\in\mathbb{P}_M(\Omega_\theta)} \sum_{j=1}^{N} (Y_j - f(t_j;P))^2. \tag{2.7}$$

For example, if one uses the Dirac measure approximation methods, then $P_M^N = \boldsymbol{\Delta}^T \mathbf{A}_M^N$, where $\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{\theta}) = (\Delta_{\boldsymbol{\theta}_1}, \Delta_{\boldsymbol{\theta}_2}, \ldots, \Delta_{\boldsymbol{\theta}_M})^T$, and

$$\mathbf{A}_M^N = \operatorname*{argmin}_{\mathbf{a}_M^N \in \widetilde{\mathbb{R}}^M} \sum_{j=1}^{N} \left[Y_j - f(t_j; \sum_{l=1}^{M} a_{M,l}^N \Delta_{\boldsymbol{\theta}_l}) \right]^2. \tag{2.8}$$

Here $\widetilde{\mathbb{R}}^M = \left\{ \mathbf{a}^M = (a_1^M, a_2^M, \ldots, a_M^M)^T \;\middle|\; a_j^M \geq 0, j = 1, 2, \ldots, M, \; \sum_{j=1}^{M} a_j^M = 1 \right\}$. The corresponding realization of (2.7) is given by

$$\hat{P}_M^N \;=\; \arg\min_{P\in\mathbb{P}_M(\Omega_\theta)} \sum_{j=1}^{N} (y_j - f(t_j;P))^2; \tag{2.9}$$

that is, $\hat{P}_M^N = \boldsymbol{\Delta}^T \hat{\mathbf{a}}_M^N$, where $\boldsymbol{\Delta} = (\Delta_{\boldsymbol{\theta}_1}, \Delta_{\boldsymbol{\theta}_2}, \ldots, \Delta_{\boldsymbol{\theta}_M})^T$, and

$$\hat{\mathbf{a}}_M^N = \operatorname*{argmin}_{\mathbf{a}_M^N \in \widetilde{\mathbb{R}}^M} \sum_{j=1}^{N} \left[y_j - f(t_j; \sum_{l=1}^{M} a_{M,l}^N \Delta_{\boldsymbol{\theta}_l}) \right]^2. \tag{2.10}$$

In essence, one presumes that the data was generated using the following statistical model

$$Y_j = \tilde{f}(t_j; \mathbf{a}_{0,M}) + \mathcal{E}_j, \quad j = 1, 2, 3, \ldots, N. \tag{2.11}$$

In the above equation, $\tilde{f}(t_j; \mathbf{a}_{0,M}) = f(t_j; P_{0,M})$, where $P_{0,M} = \mathbf{\Delta}^T \mathbf{a}_{0,M} \in \mathbb{P}_M(\Omega_\theta)$, and $\mathbf{a}_{0,M} \in \widetilde{\mathbb{R}}^M$ is the one that minimizes

$$\tilde{J}^0(\mathbf{a}_M) = \sigma_0^2 + \int_{t_s}^{t_f} (f(t; P_0) - \tilde{f}(t; \mathbf{a}_M))^2 d\mu(t) \tag{2.12}$$

over $\widetilde{\mathbb{R}}^M$. In other words, the functional J^0 defined by

$$J^0(P) = \sigma_0^2 + \int_{t_s}^{t_f} (f(t; P_0) - f(t; P))^2 d\mu(t) \tag{2.13}$$

has a minimizer $P_{0,M}$ in $\mathbb{P}_M(\Omega_\theta)$ for each fixed M.

Thus, we have a model "misspecification", which is due to the approximation of the infinite-dimensional space $\mathbb{P}(\Omega_\theta)$ by the finite-dimensional space $\mathbb{P}_M(\Omega_\theta)$. Under this framework, the total error between the true model (1.1) and the approximating model (2.11) can be characterized by (illustrated in Figure 1)

$$\rho(P_0, P_{0,M}) + \rho(P_{0,M}, \hat{P}_M^N),$$

where the first term $\rho(P_0, P_{0,M})$ is a measure of the accuracy of the approximating model and is often called *bias* in the statistics literature, and the second term $\rho(P_{0,M}, \hat{P}_M^N)$ is a measure of estimation accuracy and is often called *variance*. Using properties of the Prohorov metric

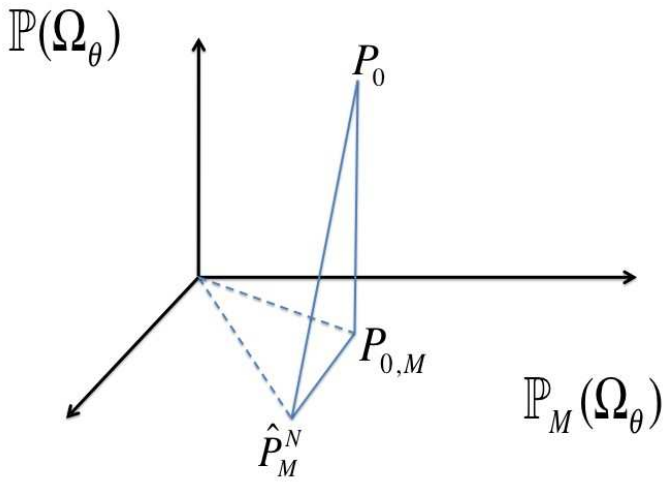

Figure 1: Illustration of the bias and the variance in the probability measure approximation.

and assumptions (A3)–(A5) as well as Theorems 2.3 and 2.4, we readily see that the bias $\rho(P_0, P_{0,M})$ approaches zero as $M \to \infty$. However, for fixed N the variance in general increases as the value of M increases (e.g., see [17]); that is, we have less confidence in the parameter estimates as the number of approximating parameters increases. Hence, there is a trade-off between the bias and the variance (illustrated in Figure 2). Model selection

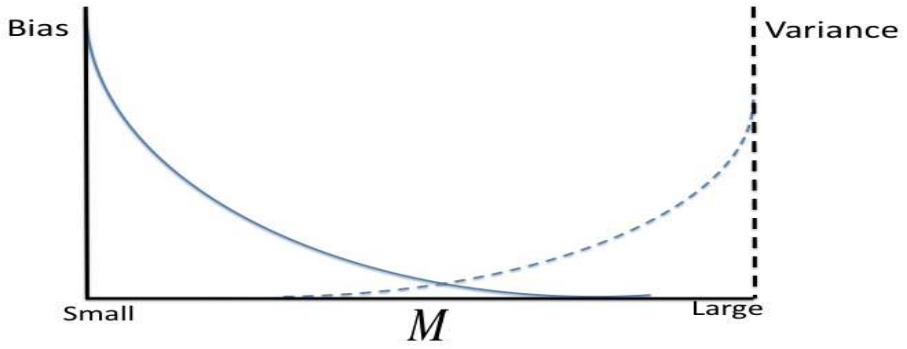

Figure 2: Illustration of the trade-off between the bias and the variance.

criteria such as the Akaike Information Criterion and the Bayesian Information Criterion have been widely used in the literature to select a best approximating model from a prior set of candidate models, and they all are based to some extent on the *principle of parsimony* (again see [12, 17]). The goal in model selection is to simultaneously minimize both bias (modeling error) and variance (estimation error). Thus one can use a model selection criterion to select a best M value (i.e., a best approximating model).

2.4 Pointwise Asymptotic Normality of the Approximatd Probability Measure Estimator

In this section, we consider the pointwise asymptotic normality of the least-squares estimator P_M^N. Since for any given $\boldsymbol{\theta}$, $P_M^N(\boldsymbol{\theta})$ is linearly dependent on $\mathbf{A}_M^N$ (for example, in the case where the Dirac measure approximation is used, $P_M^N(\boldsymbol{\theta}) = (\boldsymbol{\Delta}(\boldsymbol{\theta}))^T \mathbf{A}_M^N$), we first consider the asymptotic normality of $\mathbf{A}_M^N$. As discussed in the above section, (2.11) is misspecified. Hence, the asymptotic normality results established in [25] for the parameter estimator in a misspecified nonlinear regression can be applied to our case. To ensure the results in [25] hold, we make the following additional assumptions.

(A6) For each fixed M, $\tilde{J}^0$ has a unique minimizer $\mathbf{a}_{0,M}$ in $\widetilde{\mathbb{R}}^M$ (that is, J^0 has a unique minimizer $P_{0,M}$ in $\mathbb{P}_M(\Omega_\theta)$), and the minimizer $\mathbf{a}_{0,M}$ is interior to $\widetilde{\mathbb{R}}^M$.

(A7) For each fixed M, $\tilde{f}$ is twice continuously differentiable with respect to $\mathbf{a}_M$, and the

matrices $\mathcal{H}(\mathbf{a}_{0,M})$ and $\mathcal{F}(\mathbf{a}_{0,M})$ defined by

$$\begin{aligned}\mathcal{H}(\mathbf{a}_M) &= \frac{\partial^2 \tilde{J}^0(\mathbf{a}_M)}{\partial \mathbf{a}_M^2} \\ &= 2\int_{t_s}^{t_f} \left[\frac{\partial \tilde{f}(t;\mathbf{a}_M)}{\partial \mathbf{a}_M} \left(\frac{\partial \tilde{f}(t;\mathbf{a}_M)}{\partial \mathbf{a}_M} \right)^T - (f(t;P_0) - \tilde{f}(t;\mathbf{a}_M)) \frac{\partial^2 \tilde{f}(t;\mathbf{a}_M)}{\partial \mathbf{a}_M^2} \right] d\mu(t)\end{aligned} \tag{2.14}$$

and

$$\mathcal{F}(\mathbf{a}_M) = 4\int_{t_s}^{t_f} \left[\sigma_0^2 + (f(t;P_0) - \tilde{f}(t;\mathbf{a}_M))^2 \right] \frac{\partial \tilde{f}(t;\mathbf{a}_M)}{\partial \mathbf{a}_M} \left(\frac{\partial \tilde{f}(t;\mathbf{a}_M)}{\partial \mathbf{a}_M} \right)^T d\mu(t). \tag{2.15}$$

are nonsingular.

Theorem 2.7. *Under assumptions (A1)–(A7), for each fixed M we have*

$$\sqrt{N}\left(\mathbf{A}_M^N - \mathbf{a}_{0,M}\right) \xrightarrow{d} \mathbf{Z} \sim \mathcal{N}\left(0, \Sigma_{0,M}\right), \text{ as } N \to \infty, \tag{2.16}$$

where $\xrightarrow{d}$ denotes convergence in distribution, $\Sigma_{0,M} = \left(\mathcal{H}(\mathbf{a}_{0,M})\right)^{-1} \mathcal{F}(\mathbf{a}_{0,M}) \left(\mathcal{H}(\mathbf{a}_{0,M})\right)^{-1}$, and $\mathcal{N}\left(0, \Sigma_{0,M}\right)$ represents a multivariate normal distribution with zero mean and covariance matrix $\Sigma_{0,M}$. A strongly consistent estimator of $\mathbf{a}_{0,M}$ is $\mathbf{A}_M^N$, and a strongly consistent estimator of $\Sigma_{0,M}$ is given by

$$\Sigma_M^N = (H^N(\mathbf{A}_M^N))^{-1} F^N(\mathbf{A}_M^N) (H^N(\mathbf{A}_M^N))^{-1}, \tag{2.17}$$

where "strongly consistent" means convergence almost surely, and

$$\begin{aligned}&H^N(\mathbf{a}_M) \\ &= \frac{2}{N}\sum_{j=1}^{N} \left[\frac{\partial \tilde{f}(t_j;\mathbf{a}_M)}{\partial \mathbf{a}_M} \left(\frac{\partial \tilde{f}(t_j;\mathbf{a}_M)}{\partial \mathbf{a}_M} \right)^T - (Y_j - \tilde{f}(t_j;\mathbf{a}_M)) \frac{\partial^2 \tilde{f}(t_j;\mathbf{a}_M)}{\partial \mathbf{a}_M^2} \right],\end{aligned} \tag{2.18}$$

and

$$F^N(\mathbf{a}_M) = \frac{4}{N}\sum_{j=1}^{N} \left[(Y_j - \tilde{f}(t_j;\mathbf{a}_M))^2 \left(\frac{\partial \tilde{f}(t_j;\mathbf{a}_M)}{\partial \mathbf{a}_M} \right) \left(\frac{\partial \tilde{f}(t_j;\mathbf{a}_M)}{\partial \mathbf{a}_M} \right)^T \right]. \tag{2.19}$$

Theorem 2.7 implies that for any sufficiently large N, we have

$$\mathbf{A}_M^N \sim \mathcal{N}(\hat{\mathbf{a}}_M^N, \tfrac{1}{N}\hat{\Sigma}_M^N). \tag{2.20}$$

Here $\hat{\Sigma}_M^N$ is given by

$$\hat{\Sigma}_M^N = (\hat{H}^N(\hat{\mathbf{a}}_M^N))^{-1} \hat{F}^N(\hat{\mathbf{a}}_M^N) (\hat{H}^N(\hat{\mathbf{a}}_M^N))^{-1}, \tag{2.21}$$

with

$$\hat{H}^N(\mathbf{a}_M) = \frac{2}{N}\sum_{j=1}^{N}\left[\frac{\partial\tilde{f}(t_j;\mathbf{a}_M)}{\partial\mathbf{a}_M}\left(\frac{\partial\tilde{f}(t_j;\mathbf{a}_M)}{\partial\mathbf{a}_M}\right)^T - (y_j-\tilde{f}(t_j;\mathbf{a}_M))\frac{\partial^2\tilde{f}(t_j;\mathbf{a}_M)}{\partial\mathbf{a}_M^2}\right], \tag{2.22}$$

and

$$\hat{F}^N(\mathbf{a}_M) = \frac{4}{N}\sum_{j=1}^{N}\left[(y_j-\tilde{f}(t_j;\mathbf{a}_M))^2\left(\frac{\partial\tilde{f}(t_j;\mathbf{a}_M)}{\partial\mathbf{a}_M}\right)\left(\frac{\partial\tilde{f}(t_j;\mathbf{a}_M)}{\partial\mathbf{a}_M}\right)^T\right]. \tag{2.23}$$

If one uses the Dirac measure approximation method, then by (2.20) we know that for any sufficiently large N

$$P_M^N(\boldsymbol{\theta}) \sim \mathcal{N}((\boldsymbol{\Delta}(\boldsymbol{\theta}))^T\hat{\mathbf{a}}_M^N, \tfrac{1}{N}(\boldsymbol{\Delta}(\boldsymbol{\theta}))^T\hat{\Sigma}_M^N\boldsymbol{\Delta}(\boldsymbol{\theta})) \tag{2.24}$$

holds for any fixed $\boldsymbol{\theta}\in\Omega_\theta$. Similarly, one can use (2.20) to obtain the pointwise asymptotic result for $P_M^N(\boldsymbol{\theta})$ in the case where the linear spline approximation method is employed.

3 Numerical Results

In this section, we use the motivating example in the Introduction to demonstrate our theoretic results through simulated data. Specifically, we consider the following nonparametric model

$$Y_j = |r_s(k_j;P_0)|^2 + \mathcal{E}_j, \quad j=1,2,3,\ldots,N, \tag{3.1}$$

with r_s given by (1.3). The simulated data is then generated by simulating

$$y_j = |r_s(k_j;P_0)|^2 + \epsilon_j, \quad j=1,2,3,\ldots,N. \tag{3.2}$$

In the above equation, P_0 is chosen as the cumulative distribution function of a truncated normal distribution with its corresponding probability density function p_0 given by

$$p_0(k_0) = \frac{\beta}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(k_0-\mu)^2}{2\sigma^2}\right), \quad k_0\in[\underline{k}_0,\bar{k}_0],$$

where $\mu=700$, $\sigma=50$, $\underline{k}_0=400$, $\bar{k}_0=1090$, and β is the normalizing constant

$$\beta^{-1} = \int_{\underline{k}_0}^{\bar{k}_0}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(k_0-\mu)^2}{2\sigma^2}\right)dk_0.$$

The ϵ_j are realizations of $\mathcal{E}_j$, which are assumed to be normally distributed with zero mean and standard deviation $\sigma_0=0.002$. For all the simulations below, N is chosen as 70, the

measurement wavenumber points are $k_j = 400 + 10(j-1)$, $j = 1, 2, \ldots, 70$, and the values for the rest of model parameters are chosen as

$$\tau = 0.03, \quad \varepsilon_s = 2.7, \quad \varepsilon_\infty = 2.5.$$

Since P_0 is chosen as an absolutely continuous function, we will use the linear spline approximation method in the simulations demonstrated below.

In the presentation below, we first use the Akaike Information Criterion to determine the optimal value of M, where the probability measure is obtained by the linear spline approximation method. We then compare the confidence band obtained using the asymptotic normality results with the one obtained with the Monte Carlo simulations.

3.1 Optimal Value of M

As we stated earlier in this section, we use the Akaike Information Criterion (AIC), one of the most widely used model selection criteria, to determine the optimal value for M. The AIC was developed by Akaike (in 1973), and it is based on Kullback-Leibler information (a well-known measure of "distance" between two probability density functions) and maximum likelihood estimation. There are several advantages in using the AIC. For example, it can be used to compare both nested models and non-nested models, and it can also be used to compare multiple models at a time. For the least squares case, it can be found (e.g., see [17, Section 2.2], [12, Section 4.3.1]) that if the measurement errors are i.i.d. normally distributed, then the AIC is given by

$$\mathrm{AIC} = N \log\left(\frac{\mathrm{RSS}}{N}\right) + 2(M+1). \tag{3.3}$$

Here $M+1$ is the total number of estimated parameters including the coefficients for the splines and the variance of measurement errors, and RSS denotes the residuals of sum squares given by

$$\mathrm{RSS} = \sum_{j=1}^{N} (y_j - |r_s(k_j; \hat{P}_M^N)|^2)^2.$$

Given a prior set of candidate models, one calculates the AIC value for each model, and the best approximating model is the one with minimum AIC value. As might be expected, the AIC value depends on the data set used. Thus, when one tries to select a best model from a set of candidate models, one must use the same data set to calculate AIC values for each of the models. It should be noted that the AIC may perform poorly if the sample size N is small relative to the total number of estimated parameters (it is suggested in [17] that the AIC should be used only if the sample size is at least 40 times the total number of estimated parameters). Otherwise, one needs to use the small sample AIC, the so-called AIC_c, which is given by

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2(M+1)(M+2)}{N-M-2}. \tag{3.4}$$

For more information on the AIC and its variations, we refer the interested reader to [17] and [12, Chapter 4].

The set of our candidate models is chosen as model (2.11) with $M = 5, 10, 15, 20, 25$ and 30, and $\mathcal{E}_j$, $j = 1, 2, 3, \ldots, N$, being i.i.d. normally distributed with zero mean and constant variance. Note that for all our models the sample size is less than 40 times total number of estimated parameters. Hence, we will use the AIC_c to select the best model. Figure 3 depicts the AIC_c values for each of these models. From this figure we see that the model with $M = 15$ is the one with the minimum AIC_c value and thus it is the best approximating model as measured by AIC_c.

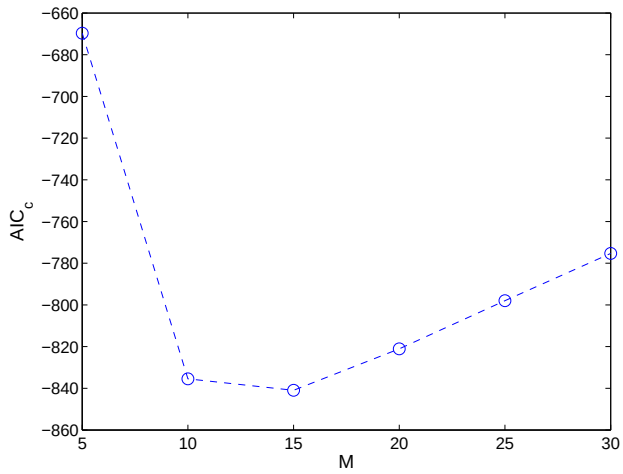

Figure 3: The AIC_c values for model (2.11) with $M = 5, 10, 15, 20, 25$ and 30.

3.2 Pointwise Confidence Band

In this section, we construct the pointwise confidence band for P_M^N by using both the asymptotic normality results presented in Section 2.4 and Monte Carlo simulations, where M is chosen as the optimal value of $M = 15$ obtained in the above analysis.

By (2.20) we know that for any sufficiently large N

$$P_M^N(k_0) \sim \mathcal{N}(\hat{P}_M^N(k_0), \tfrac{1}{N}(\boldsymbol{L}(k_0))^T \hat{\Sigma}_M^N \boldsymbol{L}(k_0)) \tag{3.5}$$

holds for any fixed $k_0 \in [\underline{k}_0, \bar{k}_0]$. In the above equation, $\hat{P}_M^N(k_0) = (\boldsymbol{L}(k_0))^T \hat{\mathbf{a}}_M^N$, where

$$\boldsymbol{L}(k_0) = \left(\int_{\underline{k}_0}^{k_0} l_1(\xi) d\xi, \int_{\underline{k}_0}^{k_0} l_2(\xi) d\xi, \ldots, \int_{\underline{k}_0}^{k_0} l_M(\xi) d\xi \right)^T.$$

One can then use (3.5) to construct the pointwise $100(1-\alpha)\%$ level confidence band, which is given by

$$\left[\hat{P}_M^N(k_0) - t_{1-\alpha/2} \text{SEP}_{\text{AN}}(k_0), \quad \hat{P}_M^N(k_0) + t_{1-\alpha/2} \text{SEP}_{\text{AN}}(k_0) \right], \quad k_0 \in [\underline{k}_0, \bar{k}_0].$$

Here $\mathrm{SEP}_{\mathrm{AN}}(k_0) = \sqrt{\frac{1}{N}(\boldsymbol{L}(k_0))^T \hat{\Sigma}_M^N \boldsymbol{L}(k_0)}$, and the critical value $t_{1-\alpha/2}$ is determined by $\mathrm{Prob}\{T \geq t_{1-\alpha/2}\} = \alpha/2$, where T has a student's t distribution t^{N-M} with $N-M$ degrees of freedom. For the simulations illustrated below, l_j is the j^{th} piecewise linear spline element using equally spaced nodes, and central difference schemes are used to approximate the first and second order derivatives involved in the covariance matrix $\hat{\Sigma}_M^N$.

To construct a pointwise confidence band using the Monte Carlo simulations, we first generate K simulated data sets and then estimate the probability measure for each data set. We denote the estimated weights for the mth simulation as $\hat{\mathbf{a}}_M^{N,(m)}$ and the corresponding estimated probability measure as $\hat{P}_M^{N,(m)}$; that is, $\hat{P}_M^{N,(m)}(k_0) = (\boldsymbol{L}(k_0))^T \hat{\mathbf{a}}_M^{N,(m)}$. Then the mean vector and covariance matrix for the estimator $\mathbf{A}_M^N$ obtained using Monte Carlo simulations are computed as

$$\hat{\mathbf{a}}_{\mathrm{MC}} = \frac{1}{K}\sum_{m=1}^{K} \hat{\mathbf{a}}_M^{N,(m)}, \quad \hat{\Sigma}_{\mathrm{MC}} = \frac{1}{K-1}\sum_{m=1}^{K} (\hat{\mathbf{a}}_M^{N,(m)} - \hat{\mathbf{a}}_{\mathrm{MC}})(\hat{\mathbf{a}}_M^{N,(m)} - \hat{\mathbf{a}}_{\mathrm{MC}})^T,$$

and the mean and covariance for the corresponding probability measure estimator P_M^N are respectively given by

$$\hat{P}_{\mathrm{MC}}(k_0) = \frac{1}{K}\sum_{m=1}^{K} \hat{P}_M^{N,(m)} = \frac{1}{K}\sum_{m=1}^{K} (\boldsymbol{L}(k_0))^T \hat{\mathbf{a}}_M^{N,(m)} = (\boldsymbol{L}(k_0))^T \hat{\mathbf{a}}_{\mathrm{MC}}$$

and

$$\begin{aligned}
(\mathrm{SEP}_{\mathrm{MC}}(k_0))^2 &= \frac{1}{K-1}\sum_{m=1}^{K} (\hat{P}_M^{N,(m)}(k_0) - \hat{P}_{\mathrm{MC}}(k_0))^2 \\
&= \frac{1}{K-1}\sum_{m=1}^{K} \left[(\boldsymbol{L}(k_0))^T \left(\hat{\mathbf{a}}_M^{N,(m)} - \hat{\mathbf{a}}_{\mathrm{MC}} \right) \right]^2 \\
&= (\boldsymbol{L}(k_0))^T \hat{\Sigma}_{\mathrm{MC}} \boldsymbol{L}(k_0).
\end{aligned}$$

The pointwise $100(1-\alpha)\%$ level confidence band is then given by

$$[\hat{P}_{\mathrm{MC}}(k_0) - t_{1-\alpha/2}\mathrm{SEP}_{\mathrm{MC}}(k_0), \hat{P}_{\mathrm{MC}}(k_0) + t_{1-\alpha/2}\mathrm{SEP}_{\mathrm{MC}}(k_0)].$$

Figure 4 depicts the pointwise confidence bands for P_M^N obtained using both the pointwise asymptotic normality results and the Monte Carlo simulations, where α is chosen to be 0.1, and $K = 1000$. We observe from this figure that the confidence bands obtained by these two approaches are similar except at the plateau regions where the confidence band obtained using the asymptotic results is wider than that obtained using the Monte Carlo simulations. To have some idea of this discrepancy, we calculate the confidence intervals of the coefficients for each spline obtained by using both the asymptotic normality results and the Monte Carlo simulations. By (2.20) we know that the $100(1-\alpha)\%$ level confidence intervals for the coefficients obtained from the asymptotic normality results are given by

$$[\hat{a}_{M,j}^N - t_{1-\alpha/2}\mathrm{SEA}_{\mathrm{AN},j},\ \hat{a}_{M,j}^N - t_{1-\alpha/2}\mathrm{SEA}_{\mathrm{AN},j}], \quad j = 1, 2, 3, \ldots, M,$$

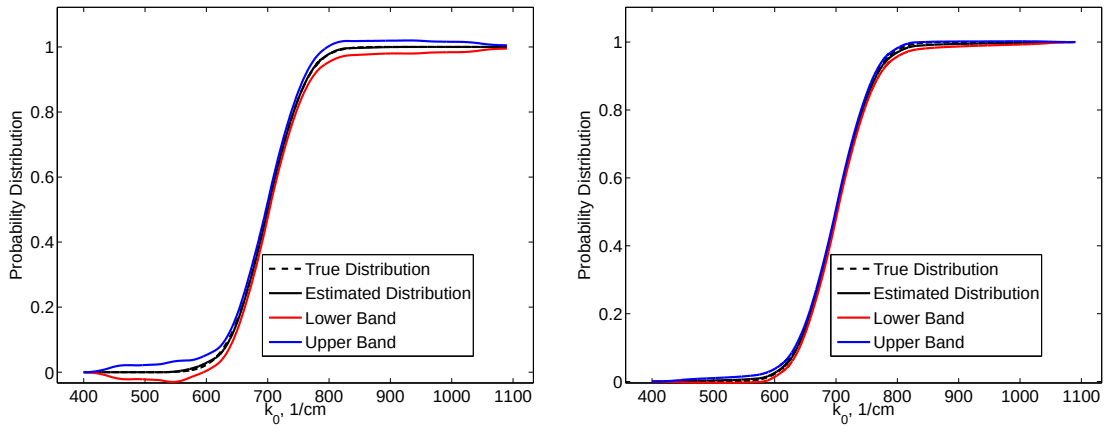

Figure 4: The pointwise confidence bands for the cumulative distribution function obtained using the pointwise asymptotic normality results (left) and the ones obtained using the Monte Carlo simulations (right).

where $\hat{a}^N_{M,j}$ is the j^{th} element of $\hat{\mathbf{a}}^N_M$, and $\text{SEA}_{\text{AN},j} = \sqrt{\frac{1}{N}\hat{\Sigma}^N_{M,jj}}$ with $\hat{\Sigma}^N_{M,jj}$ being the $(j,j)^{th}$ element of $\hat{\Sigma}^N_M$. For the Monte Carlo method, the $100(1-\alpha)\%$ level confidence intervals for the coefficients are calculated as

$$[\hat{a}_{\text{MC},j} - t_{1-\alpha/2}\text{SEA}_{\text{MC},j},\ \hat{a}_{\text{MC},j} + t_{1-\alpha/2}\text{SEA}_{\text{MC},j}], \quad j = 1, 2, 3, \ldots, M.$$

Here $\hat{a}_{\text{MC},j}$ is the j^{th} element of $\hat{\mathbf{a}}_{\text{MC}}$, and $\text{SEA}_{\text{MC},j} = \sqrt{\hat{\Sigma}_{\text{MC},jj}}$ with $\hat{\Sigma}_{\text{MC},jj}$ being the $(j,j)^{th}$ element of $\hat{\Sigma}_{\text{MC}}$. In Table 1, we give the confidence intervals obtained by these two methods. From this table, we see that the confidence intervals obtained by the asymptotic normality results are wider than those obtained by the Monte Carlo simulations except for those with splines located in the middle region where we have a good match. This is in agreement with the plots in Figure 4. It is clear from (1.2) that the relative permittivity is less sensitive to the coefficients for those splines located in the far left and far right (i.e., the plateau regions of the cumulative distribution function) where the values of the corresponding probability density function are negligible. Hence, the model output, the reflectance $|r_s|^2$, is less sensitive to the coefficients for those splines located in these two regions. Thus, one would have wider confidence intervals (i.e., have less confidence in estimates) for these coefficients as we see from (2.21)-(2.23) that the covariance matrix for the coefficients obtained by the asymptotic normality results is dependent on the sensitivity of the model output with respect to the coefficients.

j	CI using AN method	CI using MC method
1	$[-0.0006,\ 0.0006]$	$[-0.0001,\ 0.0002]$
2	$[-0.0007,\ 0.0007]$	$[-0.0001,\ 0.0001]$
3	$[-0.0010,\ 0.0010]$	$[-0.0001,\ 0.0002]$
4	$[-0.0005,\ 0.0014]$	$[-0.0001,\ 0.0004]$
5	$[0.0006,\ 0.0018]$	$[0.0010,\ 0.0021]$
6	$[0.0055,\ 0.0066]$	$[0.0052,\ 0.0063]$
7	$[0.0080,\ 0.0091]$	$[0.0080,\ 0.0090]$
8	$[0.0051,\ 0.0056]$	$[0.0049,\ 0.0058]$
9	$[0.0012,\ 0.0018]$	$[0.0010,\ 0.0018]$
10	$[-0.0004,\ 0.0005]$	$[-0.0001,\ 0.0004]$
11	$[-0.0002,\ 0.0003]$	$[-0.0001,\ 0.0002]$
12	$[-0.0003,\ 0.0003]$	$[-0.0001,\ 0.0001]$
13	$[-0.0004,\ 0.0004]$	$[-0.0001,\ 0.0001]$
14	$[-0.0002,\ 0.0002]$	$[-0.0000,\ 0.0001]$
15	$[-0.0003,\ 0.0003]$	$[-0.0001,\ 0.0001]$

Table 1: The confidence intervals (CI) computed using the pointwise asymptotic normality (AN) results and the ones obtained using the Monte Carlo (MC) simulations.

4 Concluding Remarks and Future Research Questions

In this paper we presented a computational and theoretical framework for nonparametric estimation of a probability measure P_0 in cases where the regression function is dependent on the sought-after probability measure. We also provided a consistency result for the probability measure estimator P^N. Moreover, we discussed the bias and the variance in the parameter estimation process where the infinite-dimensional parameter space $\mathbb{P}(\Omega_\theta)$ is approximated by a finite-dimensional parameter space $\mathbb{P}_M(\Omega_\theta)$, and we established the pointwise asymptotic normality for the approximated probability measure estimator P_M^N. Numerical results verify that we have a good match for the pointwise confidence band obtained by the pointwise asymptotic normality results and the Monte Carlo simulations in the region to which the model output is most sensitive.

Future efforts include investigation of convergence in distribution of the stochastic process $\sqrt{N}(P^N - P_0)$ to a certain Gaussian process. It is worth noting that the asymptotic normality of an infinite-dimensional parameter estimator in a statistic model with smooth regression function has been studied by a number of researchers (e.g., see [21] and the references therein). However, in those efforts the space for the infinite-dimensional parameter was required to be a compact set in a Hilbert space. Thus, the results established in those research efforts cannot be applied to our case as our parameter space $\mathbb{P}(\Omega_\theta)$ is a compact set in the space of all finite regular measures with weak norm $\left(\mathrm{frm}(\Omega_\theta), \|\cdot\|_{\mathrm{frm}(\Omega_\theta),w}\right)$, which is a separable normal linear space (e.g., see [24, Theorem IV.1.4]), but not a Hilbert space.

Acknowledgements

This research was supported in part by Grant Number NIAID R01AI071915-10 from the National Institute of Allergy and Infectious Diseases, in part by the Air Force Office of Scientific Research under grant number AFOSR FA9550-12-1-0188, in part by the Army Research Office under contract number W911NF-13-P-0017, in part by the National Science Foundation under Research Training Grant (RTG) DMS-0636590 and in part by the US Department of Education Graduate Assistance in Areas of National Need (GAANN) under grant number P200A120047. The authors are grateful to Bill Browning, Amanda Criner and Katie Leonard for helpful discussions during the course of parts of the research reported here.

References

[1] H.T. Banks, *A Functional Analysis Framework for Modeling, Estimation and Control in Science and Engineering*, Chapman and Hall/CRC Press, Boca Raton, FL, 2012.

[2] H.T. Banks and K. Bihari, Modelling and estimating uncertainty in parameter estimation, *Inverse Problems*, **17** (2001), 95–111.

[3] H.T. Banks and D.M. Bortz, Inverse problems for a class of measure dependent dynamical systems, *J. Inverse and Ill-posed Problems*, **13** (2005), 103–121.

[4] H.T. Banks, J. Catenacci, S. Hu and Z.R. Kenz, Decomposition of permittivity contributions from reflectance using mechanism models, CRSC-TR13-11, Center for Research in Scientific Computation, North Carolina State University, June, 2013; *Proceedings 2014 American Control Conference*, Portland, Oregon, June 4-6, 2014, to appear.

[5] H.T. Banks and J.L. Davis, A comparison of approximation methods for the estimation of probability distributions on parameters, *Applied Numerical Mathematics*, **57** (2007), 753–777.

[6] H.T. Banks, J.L. Davis, S.L. Ernstberger, S. Hu, E. Artimovich and A.K. Dhar, Experimental design and estimation of growth rate distributions in size-structured shrimp populations, *Inverse Problems,* **25** (2009), 095003(28pp).

[7] H.T. Banks and B.G. Fitzpatrick, Statistical methods for model comparison in parameter estimation problems for distributed systems, *J. Math. Biol.,* **28** (1990), 501–527.

[8] H.T. Banks and B.G. Fitzpatrick, Estimation of growth rate distributions in size-structured population models, *Quarterly of Applied Mathematics*, **49** (1998), 215–235.

[9] H.T. Banks, B.G. Fitzpatrick, L.K. Potter and Y. Zhang, Estimation of probability distributions for individual parameters using aggregate population data, CRSC-TR98-6,

Center for Research in Scientific Computation, North Carolina State University, January, 1998; In *Stochastic Analysis, Control, Optimization and Applications*, (Edited by W. McEneaney, G. Yin and Q. Zhang), Birkhäuser Verlag, Basel, 1998, 353–371.

[10] H.T. Banks and N.L. Gibson, Electromagnetic inverse problems involving distributions of dielectric mechanisms and parameters, *Quarterly of Applied Mathematics*, **64** (2006), 749–795.

[11] H.T. Banks, S.L. Grove, S. Hu and Y. Ma, A hierarchical Bayesian approach for parameter estimation in HIV models, *Inverse Problems*, **21** (2005), 1803–1822.

[12] H.T. Banks, S. Hu and W.C. Thompson, *Modeling and Inverse Problems in the Presence of Uncertainty*, Taylor/Francis-Chapman/Hall-CRC Press, Boca Raton, FL, 2014.

[13] H.T. Banks, Z.R. Kenz and W.C. Thompson, A review of selected techniques in inverse problem nonparametric probability distribution estimation, *J. Inverse and Ill-Posed Problems*, **20** (2012), 429–460.

[14] H.T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems.* Birkhausen, Boston, 1989.

[15] H.T. Banks and G.A. Pinter, A probabilistic multiscale approach to hysteresis in shear wave propagation in biotissue, *SIAM J. Multiscale Modeling and Simulation*, **3** (2005), 395–412.

[16] P. Billingsley, *Convergence of Probability Measures*, Wiley & Sons, New York, 1968.

[17] K.P. Burnham and D.R. Anderson, *Model Selection and Inference: A Practical Information-Theoretical Approach*, 2nd edition, Springer-Verlag, New York, 2002.

[18] N. Choudhuri, S. Ghosal and A. Roy, Bayesian methods for function estimation, *Handbook of Statistics*, **25** (2005), 377–418.

[19] M. Davidian and D. Giltinan, Nonlinear models for repeated measurement data: An overview and update, *J. Agricultural, Biological, and Environmental Statistics*, **8** (2003), 387–419.

[20] R.M. Dudley, *Real Analysis and Probability*, Cambridge University Press, Cambridge, UK, 2002.

[21] A.G. Kukush, Asymptotic normality of the estimator of an infinite-dimensional parameter in the model with a smooth regression function, *Mathematical Methods of Statistics*, **5** (1996), 343–356.

[22] Y.V. Prohorov, Convergence of random processes and limit theorems in probability theory, *Theor. Prob. Appl.,* **1** (1956), 157–214.

[23] G.A. Seber and C.J. Wild, *Nonlinear Regression*, Wiley, Hoboken, 2003.

[24] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[25] H. White, Consequence and detection of misspecified nonlinear regression models, *Journal of the American Statistical Association*, **76** (1981), 419–433.

[26] W. Whitt, *Stochastic-Process Limits*, Springer-Verlag, New York, 2002.