International Journal of Pure and Applied Mathematics
Volume 103 No. 4 2015, 819-843
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
doi: http://dx.doi.org/10.12732/ijpam.v103i4.15

ijpam.eu

EXISTENCE AND CONSISTENCY OF A NONPARAMETRIC ESTIMATOR OF PROBABILITY MEASURES IN THE PROHOROV METRIC FRAMEWORK

H.T. Banks[1 §], W. Clayton Thompson[2]

[1]Center for Research in Scientific Computation
Department of Mathematics
North Carolina State University
Raleigh, NC 27695-8212, USA

[2]Quantitative Systems Pharmacology Lab
Cardiovascular and Metabolic Diseases Research Unit, Pfizer Inc.
Cambridge, MA 02139, USA

Abstract: We consider nonparametric estimation of probability measures for parameters in problems where only aggregate (population level) data are available. We summarize an existing computational method for the estimation problem which has been developed over the past several decades [3, 6, 15, 18, 20]. New theoretical results are presented which establish the existence and consistency of very general (ordinary, generalized and other) least squares estimates for the measure estimation problem.

1. Motivation and Problem Formulation

In a standard nonlinear regression problem, a mathematical model is proposed which links one or more states of interest to the independent variables (regressors) of an experiment and to a vector of parameters whose values are

Received: September 6, 2015

[§]Correspondence author

unknown to the experimenter. An experiment is then conducted on the physical or biological system and data is collected for one or more states of interest. The unknown parameters of interest are then estimated in an *inverse* or *parameter estimation* problem, the theory for which is well-established [14, 24, 26]. Yet in many situations physical, biological, or experimental limitations do not permit one to sample indivudal data directly. Rather, one obtains data at the *aggregate* level as multiple individuals are sampled. In this case, it is commonly assumed that while the states of interest for these individuals are described by a single mathematical framework, each individual is described by a unique set of parameters within that framework. For instance, the growth of mosquitofish [7, 15, 16] and shrimp [10, 12] have been shown to be described by a size-structured partial differential equation model in which the rate of individual growth is assumed to vary probabilistically across the population. HIV replication data has been shown to be accurately described by a cellular-level model in which intracellular delays vary from cell to cell [5]. The probabilistic distribution of parameters has also been observed in models of electromagnetic polarization [9, 17]. These examples and others are considered at greater length in the recent book [2].

More precisely, suppose that the quantities of interest for a single individual can be described by the mathematical model

$$\begin{aligned} \frac{dy}{dt} &= g(t, y(t); q), \\ y(t_0) &= y_0. \end{aligned} \tag{1.1}$$

The parameter vector $q \in \mathbb{R}^r$ is specific to each individual within the population. The model solution is

$$y(t;\theta) = Cf(t; q, y_0) \tag{1.2}$$

where $\theta = (q, y_0) \in \mathbb{R}^{r+s} = \mathbb{R}^p$. It is assumed $f(t;\theta) \in \mathbb{R}^s$ and $C \in \mathbb{R}^{l\times s}$ so that $y \in \mathbb{R}^l$. (In the notation that follows, we tacitly assume $l = 1$; this is only for convenience and all theory presented holds for vector observations.) It is assumed that $\theta \in \Theta$ for all individuals in the population, where Θ is a set of admissible parameters.

For the aggregate data problem, one can consider n observations as random variables resulting from the direct sampling of the mean population state, but measured subject to random error. Then it is possible to define the random variables

$$V_j = v(t; P_0) + \mathcal{E}_j \tag{1.3}$$

for $j = 1, \ldots, n$ where

$$v(t; P) = E[Cy(t;\cdot)|P] = \int_\Theta Cy(t;\theta)dP(\theta),$$

and the random variables $\mathcal{E}_j$ represent measurement noise, modeling error, microfluctuations, etc. Let $\vec{\mathcal{E}} = (\mathcal{E}_1, \ldots, \mathcal{E}_n)$. It is assumed that the first two central moments of the random vector $\vec{\mathcal{E}}$ are

$$\begin{aligned} E[\vec{\mathcal{E}}] &= \vec{0} \\ Var[\vec{\mathcal{E}}] &= R. \end{aligned} \tag{1.4}$$

Without loss of generality, it may be assumed (by transforming the data and model) that the random variables $\mathcal{E}_j$ are independent and identically distributed, so that $R = \sigma^2 I_n$, where I_n is the $n \times n$ identity matrix [24].

In the context of (1.3) one does not have information on a fixed, single parameter, but rather on the distribution of parameters which characterizes the behavior of the entire population. Given n realizations v_j of the random variables V_j (which we will sometimes write $\vec{v}$ and $\vec{V}$ for notational convenience), the goal of an inverse or parameter estimation problem is to produce an estimate of the hypothetical true measure P_0. Significantly, the data is sampled from the state space of the mathematical system and not from the parameter space; thus one does not sample directly from the distribution of interest.

The estimated measure should be one that best fits the data in some appropriate sense, so that one must first choose a framework in which to work. Given that choice of framework, one must establish a set of theoretical and computation tools with which to treat the parameter estimation problem. For the results presented here, we focus on a frequentist approach using least squares estimation. Theoretical results for likelihood estimation (also in a frequentist framework) can be established with little difficulty from the results presented here. We do not consider a Bayesian approach in this manuscript, except to remark that a comparison of the frequentist and Bayesian approaches to the estimation of the unknown distribution P_0 is an interesting avenue for future work.

For the least squares problem, define the estimator

$$P_n = \arg \min_{P \in \mathcal{P}(\Theta)} J_n(\vec{V}, P) = \arg \min_{P \in \mathcal{P}(\Theta)} \sum_{j=1}^{n} \left(V_j - v(t_j; P)\right)^2 . \tag{1.5}$$

We remark that P_n is itself a random variable in that it is a function of the random variables V_j (and hence $\mathcal{E}_j$). This dependence is generally suppressed with the exception of the subscripted n, but should be carefully noted, particularly in the consideration of the existence and consistency of the estimator (see below). The inverse problem is then to use realizations v_j of the random variables V_j to compute

$$\hat{P}_n = \arg \min_{P \in \mathcal{P}(\Theta)} J_n(\vec{v}, P) = \arg \min_{P \in \mathcal{P}(\Theta)} \sum_{j=1}^{n} (v_j - v(t_j; P))^2 . \tag{1.6}$$

However, one cannot typically compute $\hat{P}_n$ as defined. In most practical problems, the model $v(t; P, \psi)$ cannot be computed exactly and must be approximated by some numerical scheme (e.g., finite difference methods, Galerkin methods, etc.). Similarly, the space $\mathcal{P}(\Theta)$ has (uncountably) infinitely many elements so that it must also be approximated. Thus, given a set of realizations $\{v_j\}$ of the random variables V_j, what one computes in practice is

$$\hat{P}_{n,M}^{N} = \arg \min_{P \in \mathcal{P}_{\mathcal{M}}(\Theta)} J_n^N(\vec{v}, P) = \arg \min_{P \in \mathcal{P}_{\mathcal{M}}(\Theta)} \sum_{j=1}^{n} \left(v_j - v^N(t_j; P)\right)^2 . \tag{1.7}$$

The immediate question of interest is how these formal definitions relate back to the actual quantity of interest, the unknown 'true' probability measure P_0. In answering this question, it must be shown that the least squares estimator P_n defined by (1.5) is well-defined and subsequently that $\hat{P}_{n,M}^{N}$ converges (in some sense) to $\hat{P}_n$ as M and N grow large. Of course, the answer to this question depends largely upon the approximation schemes used. For instance, one could define $\mathcal{P}_M(\Theta)$ to be the subset of the space of probability measures consisting of those measures with a specific parametric form. While this technique has the advantage of creating a standard nonlinear estimation problem, it may lead to inaccurate and misleading results unless there is strong evidence to suggest a particular parametric form for the unknown measure. In this document, we are concerned with *nonparametric estimation*, so that only a minimal set of restrictions is placed on the class of admissible measures. Finally, it must be shown that the estimator $\hat{P}_n$ converges to P_0 as the number of observations n increases. This is a question of the *consistency* of the least squares estimator P_n.

In developing a framework to address these issues, one encounters a rich body of mathematical theory. In this manuscript, it will be shown that the Prohorov Metric Framework (PMF) provides a natural setting in which to work. While this framework has been used extensively for computational construction of nonparametric estimates [5, 7, 9, 10, 12, 15, 16, 17], several theoretical components have remained unresolved. Here, we provide a proof of the existence (as a measureable function) and consistency of the least squares estimator underlying the computational framework. Particularly in the latter case, this provides a basis for additional work in defining and constructing confidence 'intervals' for the associated estimates [9].

2. The Prohorov Metric Framework

We begin with several general definitions and theorems which are meant to motivate the PMF and provide some background. In the interest of brevity, no proofs are given here for this motivating material. Extensive proofs are provided in the Technical Report version of this manuscript [21].

First, the Riesz Representation Theorem [25, pg. 357-358] on the space of bounded continuous functions is stated. This theorem can be used to characterize the weak* topology on the continuous dual of the space of bounded continuous functions, which provides an intuitive motivation for the weak topology on the space of probability measures. Consider the metric space Θ with its metric d, which we can write together as (Θ, d). Define the space $C_B(\Theta) = \{f : \Theta \to \mathbb{R} | f \text{ bounded, continuous}\}$.

Theorem 2.1 (Riesz). *Assume (Θ, d) is a compact (Hausdorff[1]) space. For every $f^* \in C_B(\Theta)^*$ (the continuous dual of the space $C_B(\Theta)$), there exists a unique finite signed Borel measure μ such that*

$$f^*(f) = \int_\Theta f(\theta) d\mu(\theta)$$

for all $f \in C_B(\Theta)$. Moreover, $||f|| = |\mu|(\Theta)$.

Given this identification, we may write $f^* = f^*_\mu$ when convenient. We see that the set $\mathcal{P}(\Theta)$ of probability measures on (Θ, d) can be identified with those $f^*_\mu \subset C_B(\Theta)^*$ such that $f^*_\mu(f) \geq 0$ for all $f \in C_B(\Theta)$ and $||f^*_\mu|| = \mu(\Theta) = 1$. Thus we have, in a sense, that $\mathcal{P}(\Theta) \subset C_B(\Theta)^*$. In fact, given any $f \in C_B(\Theta)$, the map from $C_B(\theta)$ into $\mathbb{R}$ given by $f^{**}_f(f^*) = f^*(f)$ defines the natural embedding of $C_B(\Theta) \hookrightarrow C_B(\Theta)^{**}$. The image of $f^{**}(C_B(\Theta))$ induces a topology on the space $C_B(\Theta)^*$, known to functional analysts as the weak* topology [2, 49–57]. (That is, $f^*_n \xrightarrow{w^*} f^*$ if and only if $f^*_n(f) \to f^*(f)$ for all $f \in C_B(\Theta)$.) When viewed in the context of $\mathcal{P}(\Theta) \subset C_B(\Theta)^*$, this is the *weak convergence of measures* known from the theory of probability and stochastic processes.

With this motivation, we now turn to the problem of characterizing the weak topology of measures using the Prohorov metric. This metric can be

[1] The assumption that Θ is Hausdorff will be maintained throughout this document.

shown to metrize the weak topology of measures, and can thus be used to establish several desirable properties of the space of probability measures.

Definition 2.2. Let (Θ, d) be any metric space (not necessarily compact) and define the set $C_B(\Theta)$ as above. Given any probability measure $P \in \mathcal{P}(\Theta)$ and some $\epsilon > 0$, an ϵ-neighborhood of P is

$$B_\epsilon(P) = \left\{ Q \middle| \left| \int_\Theta f(\theta) dQ(\theta) - \int_\Theta f(\theta) dP(\theta) \right| < \epsilon, \forall f \in C_B(\Theta) \right\}. \tag{2.1}$$

Comparing the Riesz Represention Theorem (Theorem 2.1) with the definition of $B_\epsilon(P)$, there is a clear connection between the open balls on $\mathcal{P}(\Theta)$ and the weak topology of measures. In fact, we may take the collection of all open balls as the *definition* of the weak topology of measures [23, pg. 236]. Alternatively, we have the following equivalent characterizations of the weak topology.

Theorem 2.3. *Let Θ be a topological space with σ-algebra Σ_Θ. Let $P \in \mathcal{P}(\Theta)$. The following are equivalent:*

1. *$B_\epsilon(P)$;*
2. *$\{Q | Q(C) < Q(C) + \epsilon, C \subset \Theta$ closed$\}$;*
3. *$\{Q | Q(O) < Q(O) + \epsilon, O \subset \Theta$ open$\}$;*
4. *$\{Q | Q(F) < Q(F) + \epsilon, F \in \Sigma_\Theta, P(\partial F) = 0$ (such sets are called P-continuity sets)$\}$.*

Proof. See [23, pgs. 236-237]. □

The weak topology of measures, in turn, gives rise to notions of weak (topological) convergence of measures.

Definition 2.4. Given a sequence of measures $P_M \in \mathcal{P}(\Theta)$ for all $M = 1, \ldots, \infty$, we say P_M converges weakly to P, $P_M \xrightarrow{w^*} P$, if any one (and hence all) of the following equivalent conditions holds:

1. $\left| \int_\Theta f(\theta) dP_M(\theta) - \int_\Theta f(\theta) dP(\theta) \right| \to 0$ for all $f \in C_B(\Theta)$;
2. $\limsup P_M(C) \leq P(C)$ for all C closed in Θ;
3. $\liminf P_M(O) \geq P(O)$ for all O open in Θ;
4. $\lim P_M(F) = P(F)$ for all sets $F \in \Sigma_\Theta$ such that F is a P-continuity set.

The equivalence of the above notions of convergence is often referred to as the portmanteau theorem [23, pgs. 11-12]. We remark that the notation $P_M \xrightarrow{w^*} P$ is slightly abusive as it implies weak* convergence when what is meant is the *weak convergence of measures.* Yet it should be emphasized that the two notions are *equivalent* on the space of *probability measures.*

The above definitions and theorem provide several characterizations of the weak* topology on the set of probability measures. While this characterization is mathematically sufficient, our discussions of approximation and convergence would be facilitated by some metric ρ defined on the space $\mathcal{P}(\Theta)$ which metrizes the above notions of topological convergence. That is, given two probability measures P and Q, we would like ρ to have the property that $Q \in B_\epsilon(P)$ if and only if $\rho(P, Q) < \epsilon$. Such a metric could then be used to establish intuitive measures of convergence, compactness, etc., in the space of probability measures. In fact, such a metric does exist, named for the Russian probabilist Y.V. Prohorov who first defined the metric and derived its properties.

Definition 2.5. Let (Θ, d) be a metric space. For all $F \in \Sigma_\Theta$, $F \neq \emptyset$, define the ϵ-neighborhood of F,

$$F^\epsilon = \{\phi \in \Theta | \inf_{\theta \in \Theta} d(\theta, \phi) < \epsilon\}.$$

If $F = \emptyset$, define $F^\epsilon = \emptyset$.

Definition 2.6. Let (Θ, d) be a metric space and let $\mathcal{P}(\Theta)$ be the set of all probability measures on Θ. For any two measures $P, Q \in \mathcal{P}(\Theta)$, the Prohorov metric ρ is

$$\rho(P, Q) = \inf \left\{\epsilon > 0 | Q(F) \leq P(F^\epsilon) + \epsilon \text{ and } P(F) \leq Q(F^\epsilon) + \epsilon, \text{ for all } F \in \Sigma_\Theta \right\}.$$

While this definition is far from intuitive, it gives rise to a number of desireable properties, namely that (1) the Prohorov metric, as defined, is in fact a metric, and (2) the Prohorov metric metrizes the weak topology of measures:

Theorem 2.7. *Let (Θ, d) be a separable metric space. Then ρ is a metric on $\mathcal{P}(\Theta)$.*

Theorem 2.8. *Assume (Θ, d) is separable. Assume $P_M \in \mathcal{P}(\Theta)$ for all $M = 1, \ldots, \infty$, and $P \in \mathcal{P}(\Theta)$. Then $P_M \xrightarrow{w^*} P$ if and only if $\rho(P_M, P) \to 0$.*

With these results, we have obtained the desired result–the weak topology of measures (weak* topology) is equivalent to the topology induced by the Prohorov metric on the space of probability measures over a separable metric space (Θ, d). It should be noted that in the definition of the Prohorov metric, it is sufficient to consider only sets F which are closed (see [27, Online Supplement] for a proof), so that the definitions and results presented here are in agreement with similar results obtained previously [3, 4, 10, 15, 20]. We now proceed to use the Prohorov metric to characterize the properties of the space $(\mathcal{P}(\Theta), \rho)$ which will prove useful in establishing results for the nonparametric estimation of measures.

Define

$$D = \{\Delta_{\theta_k} | \theta_k \in \Theta\}.$$

That is, D is the space of Dirac measures on Θ, defined for all $F \in \Sigma_\Theta$ as

$$\Delta_{\theta_k}(F) = \begin{cases} 1, & \theta_k \in F \\ 0, & \theta_k \notin F \end{cases}$$

Proposition 2.9. *Let (Θ, d) be a separable metric space and define $D \subset \mathcal{P}(\Theta)$ as above. Then*

$$\rho(\Delta_{\theta_1}, \Delta_{\theta_2}) = \min\{d(\theta_1, \theta_2), 1\}.$$

Corollary 2.10. *The sequence $\{\theta_k\}_{k=1}^{\infty}$ is Cauchy in the separable space (Θ, d) if and only if the sequence $\{\Delta_{\theta_k}\}_{k=1}^{\infty}$ is Cauchy in $(\mathcal{P}(\Theta, \rho)$.*

Corollary 2.11. *Let (Θ, d) be a separable metric space and let the space D be defined as above. Then D is closed in $(\mathcal{P}(\Theta), \rho)$. (That is, D is weak* closed in the space of probability measures.)*

Definition 2.12. $P \in \mathcal{P}(\Theta)$ is tight if for all $\epsilon > 0$ there exists a compact set $K \subset \Theta$ such that $P(K) > 1 - \epsilon$. A family of measures $\Pi \subset \mathcal{P}(\Theta)$ is tight if for all $P \in \Pi$, P is tight.

Theorem 2.13. *Assume (Θ, d) is complete. If for all $\epsilon, \delta > 0$ there exist $\theta_1, \ldots, \theta_M \in \Theta$ such that*

$$P\left(\bigcup_{k=1}^{M} B_\delta(\theta_k)\right) \geq 1 - \epsilon,$$

for all $P \in \Pi$, then Π is tight.

Theorem 2.14 (Prohorov). *Assume (Θ, d) is separable and let $\Pi \subset (\mathcal{P}(\Theta), \rho)$. The following are equivalent:*

- *Π is relatively (sequentially) compact;*
- *Π is tight.*

Corollary 2.15. *Assume (Θ, d) is separable. Then (Θ, d) is complete if and only if $(\mathcal{P}(\Theta), \rho)$ is complete.*

Corollary 2.16. *Assume (Θ, d) is separable. Then (Θ, d) is compact if and only if $(\mathcal{P}(\Theta), \rho)$ is compact.*

The compactness of the space $(\mathcal{P}(\Theta), \rho)$ given the compactness of (Θ, d) is of vital importance for the theoretical framework. In effect, one need only show that the cost functional $J_n(\vec{v}, P)$ in (1.6) is a continuous function of P in order to be guaranteed the existence of a minimizer to the least squares estimation problem. We need one final result which will be useful in establishing computational tools for the parameter estimation problem.

Theorem 2.17. *Assume (Θ, d) is a separable, compact metric space. Let $\{\theta_k\}_{k=1}^{\infty}$ be an enumeration of of the countable dense subset of Θ. Take $\mathbb{Q} \subset \mathbb{R}$ to be the set of all rational numbers. Define*

$$\tilde{\mathcal{P}}(\Theta) = \left\{ P \in \mathcal{P}(\Theta) \middle| P = \sum_{k=1}^{M} p_k \Delta_{\theta_k}, M \in \mathbb{N}, p_k \in [0,1] \cap \mathbb{Q}, \sum_{k=1}^{M} p_k = 1 \right\}.$$

That is, $\tilde{\mathcal{P}}(\Theta)$ is the collection of all convex combinations of Dirac measures on Θ with rational weights. Then $\tilde{\mathcal{P}}(\Theta)$ is dense in $\mathcal{P}(\Theta)$, and thus $\mathcal{P}(\Theta)$ is separable.

Taken together, these results establish that if (Θ, d) is separable and compact, then $(\mathcal{P}(\Theta), \rho)$ is also compact and separable. This, in turn, can be used to establish a number of useful results for the original estimation problem. These results are presented in the next two sections.

3. Existence of the Estimator

In this section we present new results for the existence of estimators in the Prohorov Metric Framework. We begin by proving the existence of P_n and $\hat{P}_n$ as measurable functions mapping a subset of $\mathbb{R}^n$ (that is, the data) into the space of probability measures on Θ. We remark that the statement of Theorem 3.1 concerns the estimate $\hat{P}_n$ obtained from the data realizations $\vec{v} \in \mathbb{R}^n$. This is sufficient to establish the existence of the estimator P_n as a measureable function as well, since the random vector $\vec{V}$ is by definition a measurable function from a probability triple into $\mathbb{R}^n$, and the composition of measurable functions is measurable.

Theorem 3.1. *Define the function $J_n : \mathbb{R}^n \times \mathcal{P}(\Theta) \to \mathbb{R}$ according to Equation* (1.6). *Assume (Θ, d) is separable and compact and take the space* *of probability measures $\mathcal{P}(\Theta)$ with the Prohorov metric ρ. Assume further that $J_n(\cdot, P)$ is a measurable function from $\mathbb{R}^n \to \mathbb{R}$ for each $P \in \mathcal{P}(\Theta)$, and that $J_n(\vec{v}, \cdot) : \mathcal{P}(\Theta) \to \mathbb{R}$ is continuous for each $\vec{v} \in \mathbb{R}^n$. Then there exists a measurable function $\hat{P}_n : \mathbb{R}^n \to \mathcal{P}(\Theta)$ such that*

$$J(\vec{v}, \hat{P}_n(\vec{v})) = \inf_{P \in \mathcal{P}(\Theta)} J(\vec{v}, P).$$

Proof. Let $\{\theta_k\}_{k=1}^{\infty}$ be an enumeration of the countable dense subset of Θ. For each $M \geq 1$, define

$$\mathcal{P}_M = \left\{ P \in \tilde{\mathcal{P}}(\Theta) \middle| P = \sum_{k=1}^{M} p_k \Delta_{\theta_k} \right\} \subset \tilde{\mathcal{P}}(\Theta).$$

(That is, $\mathcal{P}_M$ is the set of all discrete measures consisting of a convex combination of M Dirac measures weighted with rational coefficients.) Thus $\mathcal{P}_M$ is countable. Let $\{P_j^M\}_{j=1}^{\infty}$ be an enumeration of the elements of $\mathcal{P}_M$. (We remark that, because the M nodes θ_k are fixed in advance, the space $\mathcal{P}_M$ can be analogously considered as a subset of $\mathbb{R}^M$, a fact which will be exploited in some of the notation below.) Finally, define $\mathcal{P}_J^M = \{P_j^M\}_{j=1}^{J}$, the first J enumerated elements of $\mathcal{P}_M$.

Fix $J \geq 1$. Define the function $\tilde{P}_J^M(\vec{v})$ implicitly as

$$J(\vec{v}, \tilde{P}_J^M(\vec{v})) = \min_{P \in \mathcal{P}_J^M} J(\vec{v}, P).$$

Such a function must exist because the minimum is begin taken over a finite number of elements from a point set; if the minimum occurs at multiple elements of $\mathcal{P}_J^M$, we may arbitrarily choose the element which comes first in the enumeration so that the function $\tilde{P}_J^M(\vec{v})$ is well-defined. First, we show that $\tilde{P}_J^M(\vec{v})$ is measurable.

Let $F \in \Sigma_{\mathcal{P}_J^M}$. (Thus F is a finite point set.) We must show that the set B defined as

$$B = \left\{ \vec{v} \middle| \tilde{P}_J^M(\vec{v}) \in F \right\}$$

is contained within measurable sets $\Sigma_{\mathbb{R}^n}$ in $\mathbb{R}^n$. Since F is a finite point set, we can define for each $P_j^M \in F$ the sets

$$\begin{aligned} B_j &= \left\{ \vec{v} \middle| \tilde{P}_J^M(\vec{v}) = P_j^M \right\} \\ &= \left\{ \vec{v} \middle| J(\vec{v}, P_j^M(\vec{v})) = \min_{P \in \mathcal{P}_J^M} J(\vec{v}, P) \right\} \end{aligned}$$

$$= \left\{ \vec{v} \middle| J(\vec{v}, P_j^M(\vec{v})) = \min_{1 \leq j \leq J} J(\vec{v}, P_j^M) \right\}.$$

By assumption, the functions $J(\vec{v}, P_j^M)$ are measurable from $\mathbb{R}^n$ into $\mathbb{R}$ for all P_j^M, $j \geq 1$. The minimum over a finite set of functions is also measurable, as is the test for equality between two measureable functions. Hence $B_j \in \Sigma_{\mathbb{R}^n}$. Finally, $B = \cup B_j$, the union being over the finite number of sets B_j, hence $B \in \Sigma_{\mathbb{R}^n}$ and the function $\tilde{P}_J^M(\vec{v})$ is measurable.

As mentioned previously, we can identify the function $\tilde{P}_J^M(\vec{v})$ with $[0,1]^M \cap \mathbb{Q}^M$ via the map $\tilde{P}_J^M(\vec{v}) \mapsto (p_1^M(\vec{v}), \ldots, p_M^M(\vec{v}))$. Let $\tilde{p}_J^M$ be the first component of the vector representation for $\tilde{P}_J^M(\vec{v})$. Now consider the sequence $\{\tilde{p}_J^M\}_{J=1}^\infty$. Define

$$\hat{p}_1^M(\vec{v}) = \liminf_{J \to \infty} \tilde{p}_J^M(\vec{v}).$$

Since each $\tilde{p}_J^M(\vec{v})$ is a measurable function, so is $\hat{p}_1^M(\vec{v})$. Also, since the space $[0,1]^M$ is compact, there must exist a convergent subsequence of (the vector representation of) $\tilde{P}_J^M$ to some vector $(\hat{p}_1^M(\vec{v}), \bar{p}_2^M(\vec{v}), \ldots, \bar{p}_M^M(\vec{v}))$, which can be identified with the measure $\bar{P}_M$. Now

$$\begin{aligned} \inf_{[0,1]^{M-1} \cap \mathbb{Q}^{M-1}} J_n(\vec{v}, (\hat{p}_1^M, p_2, \ldots, p_M)) &\leq J_n(\vec{v}, \bar{P}_M) \\ &= \lim_l J_n(\vec{v}, \tilde{P}_{j_l}^M) \\ &= \lim_l \inf_{P \in \mathcal{P}_{j_l}^M} J_n(\vec{v}, P) \\ &= \inf_{P \in \mathcal{P}_M} J_n(\vec{v}, P). \end{aligned}$$

The first equality comes from the definition of $\tilde{P}_M$ and the continuity of the function J; the second equality comes from the definition of the probability measures $\tilde{P}_{j_l}^M$; the final equality arises from the density of $\{P_j^M\}$ in $\mathcal{P}$.

Now, define (with some abuse of notation)

$$J_n^{(1,M)}(\vec{v}, P) = J_n(\vec{v}, (\hat{p}_1^M, p_2, \ldots, p_M)).$$

Applying the same arguments above inductively on $J_n^{(j,M)}$, we obtain a set of measureable functions $\hat{p}_1^M(\vec{v}), \ldots, \hat{p}_M^M(\vec{v})$ such that

$$J_n(\vec{v}, (\vec{p}_1^M, \ldots, \hat{p}_M^M)) = \inf_{P \in \mathcal{P}_M} J_n(\vec{v}, P)$$

and we have proven the existence of a measurable function $\hat{P}_M \in \mathcal{P}_M$ mapping $\mathbb{R}^n \to \mathcal{P}(\Theta)$ which minimizes the cost functional J_n. We conclude the proof by noting that

$$J_n(\vec{v}, \hat{P}(\vec{v})) = \inf_{P \in \mathcal{P}(\Theta)} J_n(\vec{v}, P) = \lim_{M \to \infty} \inf_{P \in \mathcal{P}_M(\Theta)} J_n(\vec{v}, P) = \lim_{M \to \infty} J_n(\vec{v}, \hat{P}_M).$$

As the final term in the equation above is the composition of measureable functions, it is measurable, and thus $J(\vec{v}, \hat{P}(\vec{v}))$ must be measurable, so that $\hat{P}(\vec{v})$ must be measurable as well. □

4. Consistency of the Estimator

We turn to consistency results for the PMF estimator. Unlike the new results of the last section, these results have appeared (some without proofs) in the recent research monograph [18]. Theorem 3.1 shows that for any fixed n the estimator P_n and the corresponding estimate $\hat{P}_n$ exist as measurable functions mapping the data into the space of probability measures. An obvious question then, is what the resulting measures P_n or $\hat{P}_n$ represent. Since $\hat{P}_n$ is just a realization of P_n (given a specific set of data), we focus on characterization the properties of the estimator P_n. Given the problem formulation (1.5) and the statistical model (1.3), one would certainly hope that the estimator provides some information regarding the underlying 'true' distribution P_0. In particular, we would hope that $P_n \to P_0$ in some appropriate sense. If this is the case, then the estimator is said to be *consistent*. Of course, the estimator itself is a random variable, and thus this convergence must be discussed in terms of probability. With this in mind, we consider the following set of assumptions.

(A1) For any fixed n, the error random variables $\{\mathcal{E}_j\}_{j=1}^n$ are independent and identically distributed, defined on some probability triple $(\Omega, \Sigma_\Omega, P_\Omega)$.

(A2) For $\vec{\mathcal{E}} = (\mathcal{E}_1, \ldots, \mathcal{E}_n)$, $E[\vec{\mathcal{E}}] = 0$ and $Cov[\vec{\mathcal{E}}] = \sigma^2 I_n$, where I_n is the $n \times n$ identity matrix.

(A3) (Θ, d) is a separable, compact metric space; the space $\mathcal{P}(\Theta)$ is taken with the Prohorov metric ρ.

(A4) For all j, $1 \le j \le n$, $t_j \in T$ for some compact space T.

(A5) The model function $v \in C(\mathcal{P}(\Theta), C(T))$.

(A6) There exists a measure μ on T such that

$$\frac{1}{n}\sum_{j=1}^{n} g(t_j) = \int_T g(t)d\mu_n(t) \to \int_T g(t)d\mu(t)$$

for all $g \in C(T)$.

(A7) The functional

$$J_0(P) = \sigma^2 + \int_T \left(v(t;P_0) - v(t;P)\right)^2 d\mu(t)$$

is uniquely minimized at $P_0 \in \mathcal{P}(\Theta)$.

Assumption (A1) establishes the probability triple on which the error random variables $\mathcal{E}_j$ are assumed to be defined. As we will see, this probability triple will permit us to make probabilistic statements regarding the consistency of the estimator P_n. These assumptions as well as the two theorems below follow closely the theoretical results of [14] which establish the consistency of the ordinary least squares estimator for a traditional nonlinear least squares problem. The key idea is to first argue that the functions $J_n(\vec{V};P)$ converge to J_0 as n increases; then the minimizer P_n of J_n should converge to the unique minimizer P_0 of J_0 [1].

Because the functions J_n are functions of the vector $\vec{V}$, which itself depends on the random variables $\mathcal{E}_j$, these functions are themselves random variables, as are the estimators P_n. Though we have generally refrained from doing so up to this point, it will occasionally be convenient to evaluate these functions at points in the underlying probability triple. Thus we may write $J_n(\vec{V};P)(\omega)$, $\mathcal{E}_j(\omega)$, etc., whenever the particular value of ω is of interest.

The following results is stated without proof in [18]. We give a proof here for the sake of completeness.

Theorem 4.1. *Under assumptions (A1)-(A7), there exists a set $A \in \Sigma_\Omega$ with $P_\Omega(A) = 1$ such that for all $\omega \in A$,*

$$\frac{1}{n}J_n(\vec{V};P)(\omega) \to J_0(P)(\omega)$$

as $n \to \infty$ and for each $P \in \mathcal{P}(\Theta)$. Moreover, the convergence is uniform on $\mathcal{P}(\Theta)$.

Proof. As in [14], the proof will proceed in three parts. First, for any fixed element $P \in \mathcal{P}(\Theta)$, a set A_P is constructed with $P_\Omega(A_P) = 1$ such that the convergence statement holds. The sets A_P are then used to construct a set A as described. Finally, the uniform convergence is shown.

Let $P \in \mathcal{P}(\Theta)$ be fixed. We may rewrite

$$\begin{aligned}\frac{1}{n}J_n(\vec{V};P) &= \frac{1}{n}\sum_{j=1}^{n}(V_j - v(t_j;P))^2\\ &= \frac{1}{n}\sum_{j=1}^{n}(\mathcal{E}_j + v(t_j;P_0) - v(t_j;P))^2\\ &= \frac{1}{n}\sum_{j=1}^{n}\mathcal{E}_j^2 + \frac{2}{n}\sum_{j=1}^{n}(v(t_j;P_0) - v(t_j;P))\,\mathcal{E}_j\\ &\qquad + \frac{1}{n}\sum_{j=1}^{n}(v(t_j;P_0) - v(t_j;P))^2\,.\end{aligned}$$

We consider the three terms on the right. For the first term, define

$$B_1 = \left\{\omega \in \Omega \middle| \frac{1}{n}\sum_{j=1}^{n}\mathcal{E}_j(\omega)^2 \to \sigma^2\right\}.$$

By the Strong Law of Large Numbers, $P_\Omega(B_1) = 1$. For the third term, observe that

$$\frac{1}{n}\sum_{j=1}^{n}(v(t_j;P_0) - v(t_j;P))^2 \to \int_T (v(t;P_0) - v(t;P))^2\,d\mu(t) = J_0(P) - \sigma^2$$

by assumption (A6) and the continuity of $v(t;\cdot)$. (Note also that this convergence is independent of $\omega \in \Omega$.) For the second term, define

$$\tilde{\mathcal{E}}_j = (v(t_j;P_0) - v(t_j;P))\,\mathcal{E}_j.$$

Then

$$\begin{aligned}E[\tilde{\mathcal{E}}_j] &= 0\\ Var[\tilde{\mathcal{E}}_j] &= \sigma^2\,(v(t_j;P_0) - v(t_j;P))^2\\ &\le \sigma^2 \sup_{t\in T}(v(t;P_0) - v(t;P))^2\end{aligned}$$

$$\leq M_P$$

where the final inequality follows from the continuity of v and the compactness of T. Hence we have

$$\sum_{j=1}^{\infty} \frac{Var[\tilde{\mathcal{E}}_j]}{j^2} \leq M_P \sum_{j=1}^{\infty} \frac{1}{j^2} < \infty$$

and therefore the set B_P defined by

$$B_P = \left\{ \omega \in \Omega \Big| \frac{2}{n} \sum_{j=1}^{n} (v(t_j; P_0) - v(t_j; P))\, \mathcal{E}_j \to 0 \right\}$$

satisfies $P_\Omega(B_P) = 1$ by Kolmogorov's Law of Large Numbers. Finally, we may define $A_P = B_1 \cap B_P$. Then $P_\Omega(A_P) = 1$ and $\frac{1}{n} J_n(\vec{V}; P)(\omega) \to J_0(P)$ for each $\omega \in A_P$, which completes the first part of the proof.

For the second part of the proof, we must find a set A with $P_\Omega(A) = 1$ such that $\frac{1}{n} J_n(\vec{V}; P)(\omega) \to J_0(P)$ for each $\omega \in A$ and for all $P \in \mathcal{P}(\Theta)$. Naively, we desire $A = \cap A_P$, but this intersection is (in general) uncountable. Rather, we construct the set A using the dense countable subset of $\mathcal{P}(\Theta)$ (Theorem 2.17). Define

$$A_1 = \left\{ \omega \Big| \frac{1}{n} \sum_{j=1}^{n} |\mathcal{E}_j(\omega)| \to E[|\mathcal{E}_1|] \right\}.$$

Again by the Strong Law of Large Numbers, $P_\Omega(A_1) = 1$. Now define the set $\tilde{P}(\Theta)$ as before and set

$$A = A_1 \bigcap \left[\bigcap_{P \in \tilde{\mathcal{P}}} A_P \right].$$

Since the intersection is taken over a countable number of sets, each having probability one (with respect to P_Ω), $P_\Omega(A) = 1$. To complete the second part of the proof, we must show that $A \subset A_P$ for all $P \in \mathcal{P}(\Theta)$ (and not merely for all $P \in \tilde{\mathcal{P}}(\Theta)$, which holds by the definition of A). If this is the case, then $\frac{1}{n} J(\vec{V}; P)(\omega) \to J_0(P)$ for all ω in A and for all $P \in \mathcal{P}(\Theta)$.

Consider any $P \in \mathcal{P}(\Theta)$ and take $\omega \in A$, $\epsilon > 0$. Since $\omega \in A$, $\omega \in A_1$ and we may choose n_1 such that for all $n \geq n_1$,

$$\frac{1}{n} \sum_{j=1}^{n} |\mathcal{E}_j| < 1 + E[|\mathcal{E}_1|].$$

By the continuity of v and the density of $\tilde{\mathcal{P}}(\Theta)$ in $\mathcal{P}(\Theta)$, we may choose $P_M \in \tilde{\mathcal{P}}(\Theta)$ such that

$$\sup_{t\in T} |v(t;P) - v(t;P_M)| < \frac{\epsilon}{4\left(E[|\mathcal{E}_1|]+1\right)}.$$

Finally, $\omega \in A$ imples $\omega \in A_{P_M}$ wich in turn implies $\omega \in B_{P_M}$. Thus we may choose n_2 such that for all $n \geq n_2$,

$$\left| \frac{2}{n}\sum_{j=1}^{n} \left(v(t_j;P_0) - v(t_j;P_M)\right)\mathcal{E}_j \right| < \frac{\epsilon}{2}.$$

Then for $n \geq \max\{n_1, n_2\}$,

$$\begin{aligned} \left|\frac{1}{n}J_n(\vec{V};P) - J_0(P)\right| \leq \left|\sigma^2 - \frac{1}{n}\sum_{j=1}^{n}\mathcal{E}_j^2\right| + \left|\frac{2}{n}\sum_{j=1}^{n}\left(v(t_j;P_0) - v(t_j;P)\right)\mathcal{E}_j\right| \\ + \left|\frac{1}{n}\sum_{j=1}^{n}\left(v(t_j;P_0) - v(t_j;P)\right)^2 - \int_T \left(v(t;P_0) - v(t;P)\right)^2 d\mu(t)\right|. \end{aligned}$$

The first term goes to zero since $\omega \in A$ implies $\omega \in B_1$. The final term goes to zero by assumptions (A5) and (A6). For the second term,

$$\begin{aligned} \left|\frac{2}{n}\sum_{j=1}^{n}\left(v(t_j;P_0) - v(t_j;P)\right)\mathcal{E}_j\right| &\leq \left|\frac{2}{n}\sum_{j=1}^{n}\left(v(t_j;P_0) - v(t_j,P_M)\right)\mathcal{E}_j\right| \\ &\quad + \frac{2}{n}\sum_{j=1}^{n}|v(t_j;P_M) - v(t_j;P)|\cdot|\mathcal{E}_j| \\ &< \frac{\epsilon}{2} + \left(2\sup_{t\in T}|v(t;P_M) - v(t;P)|\right)\left(\frac{1}{n}\sum_{j=1}^{n}|\mathcal{E}_j|\right) \\ &< \frac{\epsilon}{2} + 2\left(\frac{2}{4\left(E[|\mathcal{E}_1|]+1\right)}\right)\left(E[|\mathcal{E}_1|]+1\right) \\ &< \epsilon. \end{aligned}$$

Thus $\frac{1}{n}J_n(\vec{V};P)(\omega) \to J_0(P)$ and thus $\omega \in A_P$. Thus $A \subset A_P$ for all $P \in \mathcal{P}(\Theta)$ and the second part of the proof is complete.

Finally, we must show the convergence is uniform on $\mathcal{P}(\Theta)$ for $\omega \in A$. To do so we will show that the sequence of functions $\frac{1}{n}J_n(\vec{V};P)(\omega)$ is equicontinuous (viewed as functions of P) and then use the Arzela-Ascoli Theorem. For fixed $\omega \in A$, let $\epsilon > 0$. Take $P \in \mathcal{P}(\Theta)$. By the continuity of v (A5) and compactness of T (A4), there exists a $\delta > 0$ such that

$$\sup_{t\in T}\left|v(t;P) - v(t;\tilde{P})\right| < \frac{1}{6}\left\{\frac{\epsilon}{E[|\mathcal{E}_1|]+1}, \frac{\epsilon}{\sup_{t\in T}|v(t;P_0)|}\right\}$$
$$\sup_{t\in T}\left|v(t;P)^2 - v(t;\tilde{P})^2\right| < \frac{\epsilon}{3},$$

for all $\tilde{P} \in B_\delta(P)$. Since $\omega \in A$, $\omega \in A_1$ and we can choose N such that $n \geq N$ implies

$$\frac{1}{n}\sum_{j=1}^{n}|\mathcal{E}_j| < E[|\mathcal{E}_1|] + 1.$$

Then for $n \geq N$ and for all $\tilde{P} \in B_\delta(P)$,

$$\begin{aligned}
\left|\frac{1}{n}J_n(\vec{V};P) - \frac{1}{n}J_n(\tilde{P})\right| \leq & \left|\frac{1}{n}\sum_{j=1}^{n}(\mathcal{E}_j + v(t_j;P_0) - v(t_j;P))^2\right.\\
& \left. -\frac{1}{n}\sum_{j=1}^{n}\left(\mathcal{E}_j + v(t_j;P_0) - v(t_j;\tilde{P})\right)^2\right|\\
= & \left|\frac{1}{n}\sum_{j=1}^{n}\left(2\mathcal{E}_j + v(t_j;P_0) - v(t_j;P) - v(t_j;\tilde{P})\right)\right.\\
& \left. \times\left(v(t_j;\tilde{P}) - v(t_j;P)\right)\right|\\
\leq & \left|\frac{2}{n}\sum_{j=1}^{n}(\mathcal{E}_j + v(t_j;P_0))\left(v(t_j;\tilde{P}) - v(t_j;P)\right)\right|\\
& + \frac{1}{n}\sum_{j=1}^{n}\left|v(t_j;P)^2 - v(t_j;\tilde{P})^2\right|\\
\leq & \frac{2}{n}\sum_{j=1}^{n}|\mathcal{E}_j|\left(\sup_{t\in T}\left|v(t;P) - v(t;\tilde{P})\right|\right)\\
& + \frac{2}{n}\left(\sup_{t\in T}|v(t;P_0)|\right)\left(\sup_{t\in T}|v(t;P) - v(t;\tilde{P})|\right)\\
& + \sup_{t\in T}\left|v(t;P)^2 - v(t;\tilde{P})^2\right|
\end{aligned}$$

$$\leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.$$

Thus the sequence of functions $\frac{1}{n} J_n(\vec{V}; P)(\omega)$ is equicontinuous for each $\omega \in A$ and by the Arzela-Ascoli Theorem, $\frac{1}{n} J_n(\vec{V}; P)(\omega) \to J_0(P)$ uniformly on compact subsets of $\mathcal{P}(\Theta)$, and hence on $\mathcal{P}(\Theta)$ itself. □

Theorem 4.2. *Under assumptions (A1)-(A7), the estimators* $P_n \xrightarrow{w^*} P_0$ *as* $n \to \infty$ *with probability 1. That is,*

$$P_\Omega \left(\left\{ \omega \middle| P_n(\vec{V})(\omega) \to P_0 \right\} \right) = 1.$$

Proof. Take the set A as in the previous theorem and fix $\omega \in A$. Then by the previous theorem, $\frac{1}{n} J_n(\vec{V}; P)(\omega) \to J_0(P)$ for all $P \in \mathcal{P}(\Theta)$. Let $\delta > 0$ be arbitrary and define $O = B_\delta(P_0)$. Then O is open in $\mathcal{P}(\Theta)$ (in the subspace topology) and O^C is compact (again, in the subspace topology). Since P_0 is the unique minimizer of $J_0(P)$ by assumption (A7), there exists $\epsilon > 0$ such that

$$J_0(P) - J_0(P_0) > \epsilon$$

for all $P \in O^C$. By the previous theorem, there exists n_0 such that for $n \geq n_0$,

$$\left| \frac{1}{n} J_n(\vec{V}; P)(\omega) - J_0(P) \right| < \frac{\epsilon}{4}$$

for all $P \in \mathcal{P}(\Theta)$. Then for $n \geq n_0$ and $P \in O^C$,

$$\begin{aligned} \frac{1}{n} \left(J_n(\vec{V}; P)(\omega) - J_n(\vec{V}; P_0)(\omega) \right) =& \frac{1}{n} J_n(\vec{V}; P)(\omega) - J_0(P) + J_0(P) - J_0(P_0) \\ &+ J_0(P_0) - \frac{1}{n} J_n(\vec{V}; P_0)(\omega) \\ \geq& -\frac{\epsilon}{4} + \epsilon - \frac{\epsilon}{4} > 0. \end{aligned}$$

But $J_n(\vec{V}; P_n)(\omega) \leq J_n(\vec{V}; P_0)(\omega)$ by definition of P_n. Hence we must have $P_n \in O = B_\delta(P_0)$ for all $n \geq n_0$, which implies $P_n(\omega) \xrightarrow{w^*} P_0$ since $\delta > 0$ was arbitrary. □

Theorem 4.2 establishes the consistency of the estimator (1.5). Given a set of data $\vec{v}$, it follows that the estimate $\hat{P}_n$ corresponding to the estimator P_n will converge to the true distribution P_0 under the stated assumptions. We remark that these assumptions are not overly restrictive (compare [14, 19, 24])s though some of the assumptions may be difficult to verify in practice. Assumptions (A3)–(A5) are mathematical in nature and may be verified directly for each specific problem. Assumptions (A1) and (A2) describe the error process which is assumed to generate the collected data. While it is unlikely that one will be able to prove a priori that the error process satisfies these assumptions, posterior analysis such as residual plots [22, Ch. 3] can be used to investigate the appropriateness of the assumptions of the statistical model. Assumption (A6) reflects the manner in which data is sampled and, together with Assumption (A7), constitutes an identifiability condition for the model. The limiting sampling distribution function μ may be known if the experimenter has complete control over the values t_j of the independent variables (e.g., if the t_j are measurement times) but this is not always the case.

5. Computational Convergence

The novel results in the previous two sections establish the desirable property of consistency of the estimator P_n as a measureable function mapping the data observation process to the space of probability measures. However, it is generally not possible to directly solve the optimization problems (1.5) or (1.6) for P_n or $\hat{P}_n$ as a function of $\vec{V}$ or $\vec{v}$. As a result, approximate (generally numerical) methods must be used in order to solve (1.7) and obtain an approximate estimate $\hat{P}^N_{n,M}$. We must ascertain, then, how the approximate estimate $\hat{P}^N_{n,M}$ relates to the exact estimate $\hat{P}_n$ (for any fixed value of n.) The following result establishes the *computational* convergence of the Prohorov Metric framework. Together with the results of the previous two sections, these results establish a comprehensive body of theory for the least squares estimation of the measure P_0 that is assumed to have generated the observed data.

Theorem 5.1. *Let (Θ, d) be a compact, separable metric space and consider the space $(\mathcal{P}(\Theta), \rho)$ of probability measures on Θ with the Prohorov metric, as before. Let $\mathcal{P}_M(\Theta)$ be as defined after the proof of Theorem 3.1. Assume*

1. *the map $P \mapsto J^N_n(\vec{v}, P)$ is continuous for all n, N;*

2. *for any sequence of probability measures $P_k \to P$ in $\mathcal{P}(\Theta)$, $v^N(t; P_k) \to v(t; P)$ as $N, k \to \infty$;*

3. *$v(t; P)$ is uniformly bounded for all t, P.*

Then there exists minimizers $\hat{P}^N_{n,M}$ satisfying (1.7). Moreover, for fixed n, there *exists a subsequence (as $M, N \to \infty$) of the approximate estimates $\hat{P}^N_{n,M}$ which converges to some $\hat{P}^*_n$ which satisfies* (1.6).

This theorem provides a set of conditions under which a sequence of approximate estimates $\hat{P}^N_{n,M}$ converges to the estimate $\hat{P}_n$ of interest. This estimate is itself a realization (for a particular data set) of the estimator P_n which has been shown to exist and to be consistent, so that $P_n \to P_0$ with probability one. Thus we are assured that a computed measure $\hat{P}^N_{n,M}$ is an accurate estimate of the true distribution P_0. The assumptions of Theorem 5.1 are not restrictive. In typical problems (and, indeed, in the assumptions of other theorems appearing in this document) it is assumed that the parameter space Θ as well as the independent variable space T are compact. In such a case, Assumptions 1 and 3 above are satisfied in the individual model solutions $y(t;\theta)$ are continuous on $T \times \Theta$. Assumption 2 is then simply a condition on the convergence of the numerical procedure used in obtaining model solutions.

Significantly, the Prohorov Metric Framework is computationally constructive. In practice, one does not construct a sequence of estimates for increasing values of M and N; rather, one fixes the values of M and N to be sufficiently large to attain a desired level of accuracy. By Theorem 2.17, we need only to have some enumeration of the elements of $\mathcal{P}_M(\Theta)$ in order to compute an approximate estimate $\hat{P}^N_{n,M}$. Practically, this is accomplished by selecting M nodes in Θ, $\{\theta^M_k\}^M_{k=1}$. The optimization problem (1.7) is then reduced to a standard constained estimation problem over Euclidean M-space in which one determines the values of the weights p^M_k corresponding to each node. Thus,

$$\begin{aligned}\hat{P}^N_{n,M} &= \arg\min_{\mathcal{P}_M(\Theta)} \sum_{j=1}^{n} (v_j - v(t_j;P))^2 \\ &= \arg\min_{\mathcal{P}_M(\Theta)} \sum_{j=1}^{n} \left(v_j - \int_\Theta Cy(t_j;\theta)dP(\theta)\right)^2 \\ &= \arg\min_{\widetilde{\mathbb{R}^M}} \sum_{j=1}^{n} \left(v_j - \left(\sum_{k=1}^{M} Cy(t_j;\theta^M_k)p^M_k\right)\right)^2,\end{aligned}$$

where in the final line we seek the weights $\bar{p}^M = (p^M_1, \ldots, p^M_M)^T \in \widetilde{\mathbb{R}^M} = \{\bar{p}^M | p^M_k \in \mathbb{R}^+, \sum_{k=1}^M p^M_k = 1\}$. These are sufficient to characterize the approximating discrete estimate $\hat{P}^N_{n,M}$ since the nodes are assumed to be fixed in advance. Moreover, define

$$H_{kl} = 2\sum_j \left(Cy(t_j;\theta_k)\right)\left(Cy(t_j;\theta_l)\right)$$

$$f_k = -2 \sum_j v_j \left(C y(t_j; \theta_k) \right)$$

$$c = \sum_j (v_j)^2 .$$

Then one can equivalently compute [10]

$$\hat{P}^N_{n,M} = \arg \min_{\widetilde{\mathbb{R}^M}} \left(\frac{1}{2} \left(\bar{p}^M \right)^T H \bar{p}^M + f^T \bar{p}^M + c \right) . \tag{5.1}$$

From this reformulation, it is clear that the approximate problem (1.7) has a unique solution if H is positive definite. If the individual mathematical model (1.2) is independent of P (See [2, Sec. 14.1.2] for a complete discussion) then the matrices H and f can be precomputed in advance. Then one can rapidly (and exactly) compute the gradient and Hessian of the objective function in a numerical optimization routine. As M grows large, the quadratic optimization problem (5.1) becomes poorly conditioned [10]. Thus there is a trade-off: M must be chosen sufficiently large so that the computational approximation is accurate, but not so large that ill-conditioning leads to large numerical errors. The efficient choice of M as well as the choice of the nodes $\{\theta_k\}_{k=1}^M$ is an open research problem.

It should be acknowledged that the uniquness of the computational problem (i.e., when H is positive definite) is not sufficient to ensure the uniqueness of the limiting estimate $\hat{P}^*_n$ in Theorem (5.1) (as there could be multiple convergent subsequences). However, if $J_n(\vec{v}; P)$ is uniquely minimized, then every subsequence of $\hat{P}^N_{n,M}$ which converges must converge to that unique minimizer. Moreover, under assumptions (A1)–(A7), it has been shown that $\frac{1}{n} J_n(\vec{v}, P) \to J_0(P)$ (as n grows large) with probability one, and the function $J_0(P)$ is assumed to be uniquely minimized by P_0.

6. Concluding Remarks

In this document we have defined a parameter estimation problem in which one has a mathematical model describing the dynamics of an individual biological or physical process but data which is sampled from a population of individuals. Because each individual is assumed to be described my a unique set of parameters, the data is described not by a single parameter but by the probability distribution (over all individuals) from which these parameters are sampled.

Theoretic results for the nonparametric measure estimation problem are presented which establish the existence and consistency of the nonparametric estimator. Combined with established computational/approximation techniques, these results form a comprehensive theoretical basis for the nonparametric least squares estimation of a probability measure.

Several open problems remain. First, while the computational scheme is simple, it is not clear how one should go about choosing the M nodes θ_k from the dense subset of Θ which are then used to estimate weights p_k. From a theoretical perspective, the nodes need only to be added so that they 'fill up' the parameter space in an appropriate way. In practice, however, rounding error and ill-conditioning can be quite problematic, particularly for a poor choice of nodes. A more complete computational algorithm will include information on how to optimally choose the M nodes θ_k (as well as the value of M). Some results in these directions can be found in [8, 10, 11].

Additionally, given the consistency of the estimator P_n, it would be desirable to place some measure of confidence on the estimated probability distribution. The traditional frequentist approach relies on either asymptotic theory or bootstrapping to construct such measures of confidence. In the former case, it is not clear how one might extend notions of sensitivity to the space of probability measures, which would require a notion of differentiability on the space of probability measures. In the latter case, the results provide some computational estimates but do not enjoy any theoretical rigor. Some work on these topics has been considered [8, 9, 11, 13] and is ongoing.

Acknowledgements

This research was supported by the Air Force Office of Scientific Research under grant number AFOSR FA9550-12-1-0188 and grant number AFOSR FA9550-15-1-0298.

References

[1] Takeshi Amemiya, Nonlinear regression models, Ch. 6 in *Handbook of Econometrics, Volume I*, Z. Griliches and M. D. Intriligator, Eds. North Holland, Amsterdam, 333–389.

[2] H.T. Banks, *A Functional Analysis Framework for Modeling, Estimation, and Control in Science and Engineering*, CRSC Press, Boca Raton, 2012.

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

[4] 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.

[5] H.T. Banks, D.M. Bortz, and S.E. Holte, Incorporation of variability into the mathematical modeling of viral delays in HIV infection dynamics, *Mathematical Biosciences* **183** (2003), 63–91.

[6] H.T. Banks, D.M. Bortz, G.A. Pinter, and L.K. Potter, Modeling and imaging techniques with potential for application in bioterrorism, CRSC-TR03-02, North Carolina State University, January 2003; Chapter 6 in *Bioterrorism: Mathematical Modeling Applications in Homeland Security* (H.T. Banks and C. Castillo-Chavez, eds.), Frontiers in Applied Math, **FR28**, SIAM, Philadelphia, 2003, 129–154.

[7] H.T. Banks, L.W. Botsford, F. Kappel, C. Wang, Modeling and estimation in size structured population models, LCDS-CCS Report 87-13, Brown University; *Proc. 2nd Course on Mathematical Ecology*, (Trieste, December 8–12, 1986) World Press (1988), Singapore, 521–541.

[8] H.T. Banks, J. Catenacci, and A. Criner, Quantifying the degradation in thermally treated ceramic matrix composite, to appear.

[9] H.T. Banks, J. Catenacci, and S. Hu, Asymptotic Properties of Probability Measure Estimators in a Nonparametric Model, CRSC-TR14-05, North Carolina State University, May 2014; *SIAM J. Uncertainty Quantification*, **3** (2015), 417–433.

[10] H.T. Banks and J.L. Davis, A comparison of approximation methods for the estimation of probability distributions on parameters, *Appl. Num. Math.*, **57** (2007), 753–777.

[11] H.T. Banks and J.L. Davis, Quantifying uncertainty in the estimation of probability distributions, *Math. Biosci. Eng.* **5** (2008), 647–667.

[12] H.T. Banks, J.L. Davis, S.L. Ernsterberger, S. Hu, E. Artimovich, and A.K. Dhar, Experimental design and estimation of growth rate distributions in size-structured shrimp populations, CRSC-TR08-20, North Carolina State University, November 2008; *Inverse Problems* **25** (2009), 095003 (28 pages).

[13] H.T. Banks, S. Dediu, H.K. Nguyen, Sensitivity of dynamical systems to parameters in a convex subset of a topological vector space, *Math. Biosci. Eng.* **4** (2007), 403–430.

[14] 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.

[15] H.T. Banks and B.G. Fitzpatrick, Estimation of growth rate distributions in size structured population models, *Quarterly of Applied Mathematics*, **49** (1991), 215–235.

[16] H.T. Banks, B.G. Fitzpatrick, L.K. Potter, and Y/ Zhang, Estimation of probability distributions for individual parameters using aggregate population data, CRSC-TR98-06, North Carolina State University, January 1998; In *Stochastic Analysis, Control, Optimization, and Applications*, (W. McEneaney, G. Yin, and Q. Zhang, eds.), Birkhauser, Boston, 1989.

[17] H.T. Banks and N.L. Gibson, Well-posedness in Maxwell systems with distributions of polarization relaxation parameters, CRSC-TR04-01, North Carolina State University, January 2004; *Appl. Math Letters* **18** (2005), 423–430.

[18] 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.

[19] H.T. Banks, Z.R. Kenz, and W.C. Thompson, An extension of RSS-based model comparison tests for weighted least squares, *Intl. J. Pure and Appl. Math* **79** (2012), 155–183.

[20] H.T. Banks, Z.R. Kenz, and W.C. Thompson, A review of selected techniques in inverse problem nonparametric probability distribution estimation, CRSC-TR12-13, North Carolina State University, May 2012; *J. Inverse and Ill-Posed Problems*, accepted.

[21] H.T. Banks and W.C. Thompson, Least squares estimation of probability measures in the Prohorov Metric Framework, CRSC-TR12-21, North Carolina State University, November 2012.

[22] H.T. Banks and H.T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, CRC Press, Boca Raton London New York, 2009.

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

[24] A. Ronald Gallant, *Nonlinear Statistical Models*, John Wiley and Sons, New York, 1987.

[25] H.L. Royden, *Real Analysis*, 3rd Ed., Prentice Hall, Upper Saddle River, New Jersey, 1988.

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

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